Editors

Radim Briš
Václav Snášel
Chu Duc Khanh
Phan Dao

Applied Mathematics in Engineering and Reliability

CRC Press
Taylor & Francis Group
A BALKEMA BOOK

APPLIED MATHEMATICS IN ENGINEERING AND RELIABILITY

This page intentionally left blank

PROCEEDINGS OF THE 1ST INTERNATIONAL CONFERENCE ON APPLIED MATHEMATICS IN ENGINEERING AND RELIABILITY (ICAMER 2016), HO CHI MINH CITY, VIETNAM, 4–6 MAY 2016

Applied Mathematics in Engineering and Reliability

Editors

Radim Briš & Václav Snášel
Faculty of Electrical Engineering and Computer Science
VŠB—Technical University of Ostrava, Czech Republic

Chu Duc Khanh
Faculty of Mathematics—Statistics, Ton Duc Thang University, Vietnam

Phan Dao
European Cooperation Center, Ton Duc Thang University, Vietnam

CRC Press
Taylor & Francis Group
Boca Raton London New York Leiden

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A BALKEMA BOOK

CRC Press/Balkema is an imprint of the Taylor & Francis Group, an informa business

Typeset by V Publishing Solutions Pvt Ltd., Chennai, India
Printed and bound in Great Britain by CPI Group (UK) Ltd, Croydon, CR0 4YY

Although all care is taken to ensure integrity and the quality of this publication and the information herein, no responsibility is assumed by the publishers nor the author for any damage to the property or persons as a result of operation or use of this publication and/or the information contained herein.

Published by: CRC Press/Balkema
P.O. Box 11320, 2301 EH Leiden, The Netherlands
e-mail: Pub.NL@taylorandfrancis.com
www.crcpress.com – www.taylorandfrancis.com

ISBN: 978-1-138-02928-6 (Hbk)
ISBN: 978-1-315-64165-2 (eBook PDF)

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Table of contents

Applied mathematics in reliability engineering

Bayesian methods, Bayesian reliability

Efficient methods to solve optimization problems

Maintenance modelling and optimization

Monte Carlo methods for parallel computing of reliability and risk

Network and wireless network reliability

Risk and hazard analysis

Stochastic reliability modelling, applications of stochastic processes

System reliability analysis

Advanced mathematical methods for engineering

Advanced methods to solve partial differential equations

Inverse problems

Advanced numerical methods

Statistics and applied statistics

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Preface

ICAMER (International Conference on Applied Mathematics in Engineering and Reliability) is first conference on this topic in Vietnam promoted by the following institutions: Ton Duc Thang University from Ho Chi Minh City, VŠB—Technical University of Ostrava, and the European Safety and Reliability Association (ESRA). The Conference attracts broad international community, which is a good mix of academics and industry participants that present and discuss subjects of interest and application across various industries.

Main theme of the Conference is "Applied Mathematics in Reliability and Engineering". The Conference covers a number of topics within mathematics applied in reliability, risk and engineering, including risk and reliability analysis methods, maintenance optimization, Bayesian methods, methods to solve nonlinear differential equations, etc. The application areas range from nuclear engineering, mechanical engineering, electrical engineering to information technology and communication, safety engineering, finance or health. The Conference provides a forum for presentation and discussion of scientific papers covering theory, methods and applications to a wide range of sectors and problem areas. Integral demonstrations of the use of reliability and engineering mathematics were provided in many practical applications concerning major technological systems and structures.

The ICAMER Conference is organized for the first time in Vietnam. Ho Chi Minh City has been selected as the venue for the Conference. Ho Chi Minh City, one of biggest cities in Vietnam, as well as in the world, lies in the southern part of Vietnam and ranks amongst the most impressive and modern cities in the world. The city has always played an important part in the history of the country.

The host of the conference is the Ton Duc Thang University in close cooperation with VŠB—Technical University of Ostrava and ESRA. The Ton Duc Thang University, as well as VŠB—Technical University of Ostrava rank among top technical universities in both countries. They develop traditional branches of industry as metallurgy, material engineering, mechanical, electrical, civil and safety engineering, economics, computer science, automation, environmental engineering and transportation. Research and development activities of those Universities are crucial for the restructuring process in both countries and corresponding regions.

The program of the Conference includes around 50 papers from prestigious authors coming from all over the world. Originally, about 72 abstracts were submitted. After the review by the Technical Programme Committee of full papers, 40 have been selected to be included in these Proceedings. The work and effort of the peers involved in the Technical Program Committee in helping the authors to improve their papers are greatly appreciated.

Thanks to authors as well as reviewers for their contributions in this process. The review process has been conducted electronically through the Conference webpage.

Finally we would like to acknowledge the local organizing committee for their careful planning of the practical arrangements.

Radim Briš, Václav Snášel,
Chu Duc Khanh and Phan Dao
Editors

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Organization

HONORARY CHAIRS

Prof. Le Vinh Danh
President of Ton Duc Thang University, Vietnam
Prof. Ivo Vondrák
President of VŠB—Technical University of Ostrava, Czech Republic
Prof. Terje Aven
President of ESRA, University of Stavanger, Norway

CONFERENCE CHAIRMAN

Radim Briš, *VŠB—Technical University of Ostrava, Czech Republic*

CONFERENCE CO-CHAIRMEN

Václav Snášel, *VŠB—Technical University of Ostrava, Czech Republic*
Chu Duc Khanh, *Ton Duc Thang University, Vietnam*
Phan Dao, *Ton Duc Thang University, Vietnam*

ORGANIZING INSTITUTIONS

Faculty Mathematics—Statistics, Ton Duc Thang University, Vietnam
Faculty of Electrical Engineering and Computer Science, VŠB—Technical University of Ostrava, Czech Republic
European Safety and Reliability Association
European Cooperation Center, Ton Duc Thang University, Vietnam

INTERNATIONAL CONFERENCE COMMITTEE

John Andrews, *ESRA TC Chair, The University of Nottingham, UK*
Christophe Berenguer, *ESRA TC Chair, Grenoble Institute of Technology, France*
Radim Briš, *Vice-Chairman of ESRA, VŠB—Technical University of Ostrava, Czech Republic*
Marko Čepin, *ESRA TC Chair, University of Ljubljana, Slovenia*
Eric Châtelet, *Troyes University of Technology, France*
Frank Coolen, *Durham University, UK*
Phan Dao, *Ton Duc Thang University, Vietnam*
Tran Trong Dao, *Ton Duc Thang University, Vietnam*
Jesus Ildefonso Diaz, *Complutense University of Madrid, Spain*
Vo Hoang Duy, *Ton Duc Thang University, Vietnam*
Antoine Grall, *Chairman of ESRA Committee for Conferences, Troyes University of Technology, France*
Dang Dinh Hai, *University of Mississippi, USA*
Nguyen Thanh Hien, *Ton Duc Thang University, Vietnam*
Chu Duc Khanh, *Ton Duc Thang University, Vietnam*

Krzysztof Kołowrocki, *Past Chairman of ESRA Committee for Conferences, Gdynia Maritime University, Poland*
Jan Kracík, *VŠB—Technical University of Ostrava, Czech Republic*
Miroslav Vozňák, *VŠB—Technical University of Ostrava, Czech Republic*
Vitaly Levashenko, *University of Zilina, Slovakia*
Gregory Levitin, *ESRA TC Chair, The Israel Electric Corporation, Israel*
Phan Tran Hong Long, *Water Resources University, Vietnam*
Le Van Nghi, *Key Laboratory of River and Coastal Engineering, Vietnam*
Nabendu Pal, *Ton Duc Thang University, Vietnam and University Louisiana, Lafayette, USA*
Do Phuc, *Lorraine University, France*
Pavel Praks, *VŠB—Technical University of Ostrava, Czech Republic*
Joanna Soszyńska-Budny, *Gdynia Maritime University, Poland*
Do Dinh Thuan, *HCMC University of Technology and Education, Vietnam*
Nguyen Thoi Trung, *Ton Duc Thang University, Vietnam*
David Vališ, *University of Defence, Brno, Czech Republic*
Elena Zaitseva, *ESRA TC Chair, University of Zilina, Slovakia*

LOCAL ORGANIZING COMMITTEE

Tran Trong Dao, *Ton Duc Thang University, Vietnam*
Trinh Minh Huyen, *Ton Duc Thang University, Vietnam*
Phan Dao, *Ton Duc Thang University, Vietnam*
Chu Duc Khanh, *Ton Duc Thang University, Vietnam*
Vo Hoang Duy, *Ton Duc Thang University, Vietnam*

SECRETARY OF THE CONFERENCE

Dao Nguyen Anh, *Ton Duc Thang University, Vietnam*
E-mail: icamer@tdt.edu.vn

SPONSORED BY

Ton Duc Thang University, Vietnam
VŠB—Technical University of Ostrava, Czech Republic
European Safety and Reliability Association

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Message from Professor Vinh Danh Le

Welcome to the 1st International Conference on Applied Mathematics in Engineering and Reliability (**ICAMER 2016**), held at Ton Duc Thang University, Vietnam. This Conference aims to offer a forum for scientists, researchers, and managers from universities and companies to share their research findings and experiences in the field. In recognition of its special meaning and broad influence, we consider the organization of this Conference as one of our strategic activities in the development of three decades applied research university.

Ton Duc Thang University (TDTU) has always described itself as a young, inspiring and dynamically growing higher education institution in vibrant Ho Chi Minh City.

TDTU is steadily growing to meet the expanding demand for higher education as well as high-quality human resources in Vietnam. With fifteen faculties and around 25,000 students, the University is now ranked among the largest and fastest growing universities in Vietnam in all aspects.

On behalf of TDTU, the host institution of ICAMER 2016, I would like to express my sincere appreciation to our great partners—European Safety and Reliability Association (ESRA) and VŠB-Technical University of Ostrava (Czech Republic)—for their great efforts in organizing this Conference. I would also like to send my special thanks to conference committees, track chairs, reviewers, speakers and authors around the world for their contributions to and interest in our event.

I believe that you will have an interesting and fruitful conference in Vietnam. I really look forward to welcoming all of you at our campus and hope that this Conference will start a long-term partnership between you and our university.

February 2016

Prof. Vinh Danh Le, Ph.D.
President
Ton Duc Thang University, Vietnam

This page intentionally left blank

Introduction

The Conference covers a number of topics within engineering and mathematics. The Conference is especially focused on advanced engineering mathematics which is frequently used in reliability, risk and safety technologies.

I APPLIED MATHEMATICS IN RELIABILITY ENGINEERING

- Bayesian Methods, Bayesian Reliability
- Efficient Methods to Solve Optimization Problems
- Maintenance Modelling and Optimization
- Monte Carlo Methods for Parallel Computing of Reliability and Risk
- Network and Wireless Network Reliability
- Risk and Hazard Analysis
- Stochastic Reliability Modelling, Applications of Stochastic Processes
- System Reliability Analysis

II ADVANCED MATHEMATICAL METHODS FOR ENGINEERING

- Advanced Methods to Solve Partial Differential Equations
- Inverse Problems
- Advanced Numerical Methods
- Statistics and Applied Statistics

This page intentionally left blank

Applied mathematics in reliability engineering

Bayesian methods, Bayesian reliability

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

A conjugate prior distribution for Bayesian analysis of the Power-Law Process

V.C. Do & E. Gouno
Laboratoire de Mathématiques de Bretagne Atlantique Université de Bretagne Sud, France

ABSTRACT: This paper focuses on the Bayesian analysis of the Power-Law Process. We investigate the possibility of a natural conjugate prior. Relying on the work of Huang and Bier (1998), we introduce and study the H-B distribution. This distribution is a natural conjugate prior since the posterior distribution is a HB-distribution. We describe a strategy to draw out the prior distribution parameters. Results on simulated and real data are displayed.

1 INTRODUCTION

The Power-Law Process (PLP) is a non homogeneous Poisson process $\{N(t), t \geq 0\}$ with a power law intensity $m(t) = \beta t^{\beta-1} / \alpha^{\beta}, \alpha > 0, \beta > 0$. The literature on this process is abundant. It has been widely used to describe reparaible systems, in software reliability, in reliability growth, etc. Inference was carried out by many authors from a frequentist and a Bayesian perspective. Choosing the prior distribution is an important matter. Guida et al. (1989) propose different choice: a joint non informative prior of the form $(\alpha\beta)^{-1}$, a uniform distribution for β and $1/\alpha$ for α. Then considering a gamma prior distribution on $m(t)$, the number of expected failures, they express a distribution for α given β. Bar-Lev et al. (1992) consider a joint prior for $(\alpha\beta)$ of the form $(\alpha\beta^{\nu})^{-1}$. They obtain a chi-square distribution for β posterior distribution but a cumbersome expression for α posterior distribution. Sen and Khattree (1998) study specifically the Bayesian estimator of $m(t)$ considering different lost functions. Our purpose here is to investigate conjugate prior for the Bayesian analysis of PLP. This problem has already been addressed by Huang and Bier (1998) and (Oliveira & Gilardoni 2012). In section 2, we define a 4-parameter distribution that we name the H-B distribution (for Huang-Bier distribution). Properties of this distribution are given and it is shown that this distribution is a natural conjugate prior for Bayesian analysis of the PLP. The Bayes estimates are then obtained and we suggest a technique to elicit the parameters of the prior distribution. This technique is very attractive and simple since the practionner has only to give a prior guess on β and a standard deviation associated with his guess. To end with and before concluding, we apply the method on simulated data and on data from aircraft generator.

2 THE HUANG-BIER DISTRIBUTION

In this section, we introduce a new bi-variate distribution. This distribution has four parameters. One of its component has a gamma distribution while the marginal distribution of the other one do not have an explicit expression. However, the expectation and the variance of each component can be obtained. First of all, we give the definition of the H-B distribution.

Definition 1 – *A bivariate r.v.* $(X,Y) \in \mathbb{R}^{+} \times \mathbb{R}^{+}$ *has a Huang-Bier distribution with parameters* (a,b,c,d) *where* $a,b,c,d > 0$ *and such that* $c < d^{a}$, *if it has a p.d.f. of the form:*

$$f_{X,Y}(x,y) = K\,(xy)^{a-1} c^{y} \exp\{-bd^{y}x\} \tag{1}$$

where $K = [b\log(d^{a}/c)]^{a} / \Gamma(a)^{2}$.

We denote: $(X,Y) \sim HB(a,b,c,d)$. Figure 1 displays a H-B distribution with parameters $(1.5, 5, 0.5, 1)$. As mentioned before, the marginal distribution of X cannot be obtained explicitly. The following theorem provides the conditionnal distribution of X.

Theorem 1 – *Let* $(X,Y) \sim HB(a,b,c,d)$.
Then

i. X *given* $Y = y$ *has a gamma distribution with parameters* (a, bd^{y}),
ii. Y *has a gamma distribution with parameters* $(a, log(d^{a}/c))$.

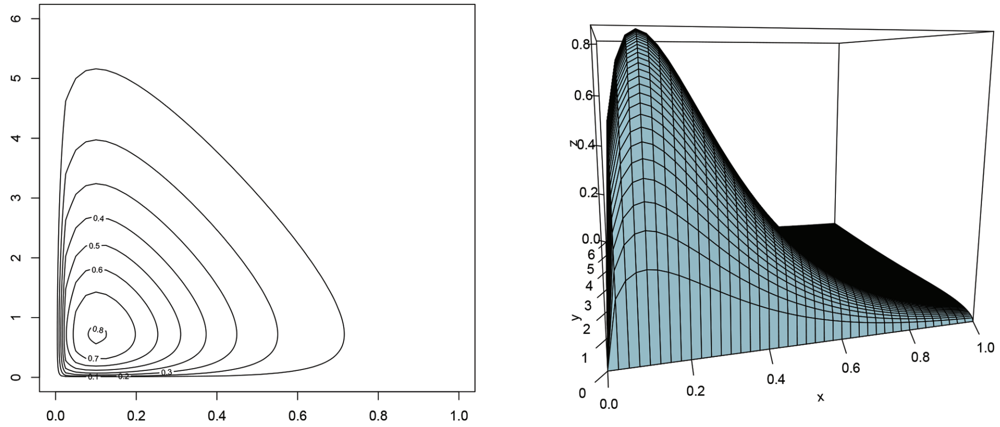


Figure 1. Probability density function for the HB distribution with $a = 1.5$, $b = 5$, $c = 0.5$ and $d = 1$.

Proof:

$$\begin{aligned} f_Y(y) &= \int_0^{+\infty} K\,(xy)^{a-1} c^y \exp\{-bd^y x\} dx \\ &= K\,y^{a-1} c^y \int_0^{+\infty} x^{a-1} \exp\{-bd^y x\} dx \\ &= K\,y^{a-1} c^y \frac{\Gamma(a)}{(bd^y)^a} \\ &= \frac{[\log(d^a/c)]^a}{\Gamma(a)} \left(\frac{c}{d^a}\right)^y y^{a-1} \\ &= \frac{[\log(d^a/c)]^a}{\Gamma(a)}\, y^{a-1} \exp\{-log(d^a/c)\, y \end{aligned}$$

Therefore Y has a gamma distribution with parameters $(a, \log(d^a/c))$.

To prove (ii) we write:

$$\begin{aligned} f_{X|Y=y}(x) &= f_{X,Y}(x,y)/f_Y(y) \\ &= \frac{K\,(xy)^{a-1} c^y \exp\{-bd^y x\}}{K\,y^{a-1} c^y \frac{\Gamma(a)}{(bd^y)^a}} \\ &= \frac{(bd^y)^a}{\Gamma(a)} x^{a-1} \exp\{-bd^y x\} \end{aligned}$$

Thus $X\,|\,Y = y$ has a gamma distribution with parameters (a, bd^y).

The previous theorem allows us to compute the expectation and the variance of X and Y.

Let $k = \log(d^a/c)$.

We have

$$E(Y) = a/k \tag{2}$$

and

$$Var(Y) = a/k^2 \tag{3}$$

To compute $E(X)$ we consider the conditional expectation and compute $E[E(X\,|\,Y)]$ to obtain:

$$E(X) = \frac{a}{b}\left[\frac{k}{k+\log d}\right]^a \tag{4}$$

A similar reasonning provides:

$$Var(X) = \frac{a}{b^2}\left[\frac{k}{k+2\log d}\right]^a. \tag{5}$$

It is interesting to remark that when $d = 1$, X and Y are independent and have gamma distributions. We have the following theorem:

Theorem 2 – *Let* $(X,Y) \sim H - B(a,b,c,1)$. *Then* X *and* Y *are independent,* X *has a gamma distribution with parameters* (a,b) *and* Y *has a gamma distribution with parameters* $(a, \log(1/c))$.

Proof: When $d = 1$,

$$f_{X,Y}(x,y) = K\,(xy)^{a-1} c^y \exp\{-bx\}$$

where $K = [b \log(d^a/c)]^a / \Gamma(a)^2$.

Clearly,

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{b^a}{\Gamma(a)} x^{a-1} \exp\{-bx\} \\ &\quad \times \frac{[\log(1/c)]^a}{\Gamma(a)}\, y^{a-1} \exp\{-\log(1/c)\,y \\ &= f_X(x) f_Y(y). \end{aligned}$$

Note that in this case, the expectations and the variances are easily obtained. We have:

$E(X)=a/b, Var(X)=a/b^2, E(Y)=a/\log(1/c)$ and $Var(Y)=a/[\log(1/c)]^2$.

3 CONJUGATE PRIOR

We now consider the Bayesian inference for the PLP. We reparametrize the intensity setting $1/\alpha^{\beta}=\lambda$. Then $m(t)=\lambda\beta t^{\beta-1}$. Then a prior distribution for (λ,β) is needed. The HB-distribution is a natural conjugate prior.

Theorem 3 – *Let* $\underline{t}=(t_1,\ldots,t_n)$ *be the jump dates of a Power Law Process with intensity* $\lambda\beta t^{\beta-1}$

Then a Huang-Bier distribution with parameters (a,b,c,t_n) *is a natural conjugate prior and the posterior distribution is a Huang-Bier distribution with parameters* $(n+a,1+b,c\prod_{i=1}^{n}t_i,t_n)$.

Proof: We have

$$f(\underline{t}\,|\,\lambda,\beta)=\lambda^n\beta^n\left(\prod_{i=1}^{n}t_i\right)^{\beta-1}\exp\{-\lambda t_n^{\beta}\}$$

Let us consider a H-B (a,b,c,t_n) as the joint prior distribution:

$$\pi(\alpha,\beta)\propto\lambda^{a-1}\beta^{a-1}c^{\beta}\exp\{-bt_n^{\beta}\lambda\}.$$

The posterior distribution is:

$$\begin{aligned}\pi(\lambda,\beta|\underline{t})&\propto f(\underline{t}\,|\,\lambda,\beta)\pi(\lambda,\beta)\\&\propto\lambda^{n+a-1}\beta^{n+a-1}\left(c\prod_{i=1}^{n}t_i\right)^{\beta-1}\\&\quad\times\exp\{-(1+b)t_n^{\beta}\lambda\}\end{aligned}$$

That is to say a HB distribution with parameters $(n+a,1+b,c\prod_{i=1}^{n}t_i,t_n)$.

Assuming a quadratic loss, the Bayes estimators are the expectation of the posterior distributions. Therefore by (4) and (2), we have:

$$\hat{\beta}_{\text{Bayes}}=\frac{n+a}{k'+S_n}\tag{6}$$

and

$$\hat{\lambda}_{\text{Bayes}}=\frac{n+a}{1+b}\left[\frac{k'+S_n}{k'+S_n+\log t_n}\right]^{n+a}\tag{7}$$

where $S_n=\Sigma_{i=1}^{n}\log(t_n/t_i)$ and $k'=log(t_n^a/c)$.

Let us recall the expression of the MLE:

$$\hat{\beta}_{\text{MLE}}=\frac{n}{\sum_{i=1}^{n}\log(t_n/t_i)}\quad\text{and}\quad\hat{\lambda}_{\text{MLE}}=n/t_n^{\hat{\beta}_{\text{MLE}}}.$$

One can see that $\hat{\beta}_{\text{Bayes}}$ can be expressed as a convex combination of the MLE and the expectation of the prior distribution:

$$\hat{\beta}_{\text{Bayes}}=q_n(a,c)\,\hat{\beta}_{\text{MLE}}+[1-q_n(a,c)]\frac{a}{k'},$$

where

$$q_n(a,c)=\frac{S_n}{k'+S_n}.\tag{8}$$

This remark will be usefull to choose the parameters (a,b,c,d) in the sequel.

A relationship between $\hat{\lambda}_{\text{Bayes}}$ and $\hat{\lambda}_{\text{MLE}}$ can be proposed.

From (6), $k'+S_n=\frac{n+a}{\hat{\beta}_{\text{Bayes}}}$.

Substituting in 7,

$$\hat{\lambda}_{\text{Bayes}}=\frac{n+a}{1+b}\left[\frac{1}{1+\log t_n^{\hat{\beta}_{\text{Bayes}}}/(n+a)}\right]^{n+a}$$

which can be approximated by:

$$\hat{\lambda}_{\text{Bayes}}\approx\frac{n+a}{1+b}\exp\{-\log t_n^{\hat{\beta}_{\text{Bayes}}}\}=\frac{n+a}{1+b}\frac{1}{t_n^{\hat{\beta}_{\text{Bayes}}}}.\tag{9}$$

Therefore $\hat{\lambda}_{\text{Bayes}}$ can be expressed as a convex combination of the MLE and the prior expectation of λ given β:

$$\hat{\lambda}_{\text{Bayes}}=(1-\xi)\hat{\lambda}_{\text{MLE}}+\xi\frac{a}{bt_n^{\hat{\beta}_{\text{Bayes}}}},$$

where $\xi=\frac{b}{1+b}$. This approximation will be used in the next section to elicit prior parameters.

4 PRIOR ELICITATION

We suggest a first strategy (elicitation strategy 1) to choose the values for the prior parameters. Suppose that the practioner has a guess $g_{\beta,1}$ at the value of β and a guess $g_{\beta,2}$ at the standard deviation associated with $g_{\beta,1}$. Then a value for a can be obtained by solving the system:

$$\begin{cases} a / k' = g_{\beta,1} & (10) \\ \sqrt{a} / k' = g_{\beta,2} & (11) \end{cases}$$

We have:

$$a = \left[g_{\beta,1} / g_{\beta,2} \right]^2 \text{ and } k' = a / g_{\beta,1}.$$

Then (6) can be computed.

According to (9), $\frac{n+a}{n(1+b)}$ can be interpreted as a confidence or corrective factor α associated with the MLE. A value for b can be obtained solving the equation:

$$\frac{n+a}{n(1+b)} = \alpha \text{ to obtain } b = \frac{n+a}{n\alpha} - 1,$$

with $\alpha = q_n(a,c)$ for example.

A second strategy (elicitation strategy 2) consists in considering a guess at and a guess at β, g_β. From the guess at ξ, a value for b can be deduced. Setting $n = a / b$, a value for a is obtained. The guess at β provides a value for k' since $k' = a / g_\beta$. The results using this strategy are displayed in Table 2.

5 APPLICATIONS

5.1 *Simulated data*

In order to investigate the behaviour of the HB natural conjugate prior, we make a comparison between Bayesian estimation and maximum likelihood estimation relying on simulated data from PLP. For the elicitation strategy 1, three different values of priormean for β are investigated: case [1] prior mean underestimates the true value of parameters, case [2] prior mean overestimates the true value of parameters, and case [3] prior mean is relatively close to the true value of parameter used in generating the data sets. For a each given prior guess $g_{\beta,1}$, computations are carried out using three incertitude values of variability $g_{\beta,2}$ according to the scheme: $g_{\beta,2} = \rho g_{\beta,1}$, where $\rho = 0.3, 0.6, 0.9$ are the coefficient of variation. The sample sizes vary from small size $n = 20$ to medium size $n = 150$ and then to very large size $n = 2000$. The small size case is in favour of showing the advantage of Bayesian approach.

Table 1 and Table 2 describe the results of estimation based on the data sets generated by a PLP with true parameters $\beta = 1.38$, $\lambda = 0.008$. In table 1, we use the elicitation strategy 1 for choosing the values of the prior parameters. For underestimated prior guess, accurate prior guess and overestimated prior guess, we choose respectively $g_{\beta,1} = 0.9, g_{\beta,1} = 1.4, g_{\beta,1} = 2.1$.

Table 1. Mean of the Bayes estimates with elicitation strategy 1 for simulated data from a PLP with input parameter values $\beta = 1.38$ and $\lambda = 0.0008$.

Sample-size	Prior guess		Bayes estimates	
n	$g_{\beta,1}$	$g_{\beta,2}$	$\hat{\beta}_{\text{Bayes}}$	$\hat{\lambda}_{\text{Bayes}}$
10	0,90	0,27	1,1338	0,006949
		0,54	1,4009	0,012574
		0,81	1,5412	0,014838
	1,40	0,42	1,4973	0,002126
		0,84	1,6157	0,007978
		1,26	1,6750	0,011449
	2,10	0,63	1,8461	0,000892
		1,26	1,7543	0,006538
		1,89	1,7318	0,010739
	MLE		1.4343	0.001604
150	0,90	0,27	1,3545	0,001994
		0,54	1,3936	0,001714
		0,81	1,4016	0,001662
	1,40	0,42	1,4114	0,001363
		0,84	1,4124	0,001506
		1,26	1,4127	0,001535
	2,10	0,63	1,4464	0,001085
		1,26	1,4226	0,001411
		1,89	1,4180	0,001484
	MLE		1.3995	0.001082
2000	0,90	0,27	1,3848	0,000904
		0,54	1,3879	0,000881
		0,81	1,3885	0,000877
	1,40	0,42	1,3855	0,000899
		0,84	1,3855	0,000904
		1,26	1,3854	0,000905
	2,10	0,63	1,3906	0,000854
		1,26	1,3887	0,000875
		1,89	1,3883	0,000879
	MLE		1.3803	0.000834

In case of large sample size, it is not surprising that Bayesian estimates are relatively close to MLEs and tend to the true values of the parameters what ever the parameter β is underestimated or overestimated. For small and medium size, one can see that the underestimating scenario is more accurate than the two other scenarios. In more detail, the Bayesian estimators seems to increase when we let the incertitude values become larger. In case of medium sample size, the underestimated prior guess with moderate variability $g_{\beta,2} = 0.6\, g_{\beta,1}$ gives results on Bayesian estimators which are more accurate than MLEs but in small sample size case, the MLEs seems to perform better.

On the other hand, results in table 2 illustrate the elicitation strategy 2. We choose different values for ξ depending on the confidence we might have in the data. We set $\xi = 0.3, 0.6, 0.8, 0.95$. We consider as in Tables 1, 3 values for g_{β}. In small and medium sample size, it turns out that if we choose the accurate prior for β then the Bayesian estimator of λ performs better but the Bayesian estimator of β performs worse compare to strategy 1. We remark that globally the results with strategy 2 are worse than with strategy 1. But for some schemes of prior, the estimation of β for e.g. is closer to the input value. With strategy 2, we observe more dispersion on the estimates.

5.2 *Real data*

The Table 5.2 gives data that have been discussed many times in the literature (Bar-Lev, Lavi, &

Table 2. Mean of the Bayes estimates with elicitation strategy 2 for simulated data from a PLP with input parameter values $\beta = 1.38$ and $\lambda = 0.0008$.

Sample-size	Prior guess		Bayes estimates	
n	g_{β}	ξ	$\hat{\beta}_{\text{Bayes}}$	$\hat{\lambda}_{\text{Bayes}}$
10	0.90	0.30	1.3084	0.0172384
		0.60	1.0856	0.0187112
		0.80	0.9821	0.0205621
		0.95	0.9189	0.0224425
	1.40	0.30	1.5894	0.0080820
		0.60	1.4833	0.0033231
		0.80	1.4350	0.0016563
		0.95	1.4076	0.0009654
	2.10	0.30	1.7735	0.0057058
		0.60	1.8691	0.0008951
		0.80	1.9654	0.0001251
		0.95	2.0617	0.0000164
	MLE		1.4343	0.001604
150	0.90	0.30	1.1988	0.0060447
		0.60	1.0488	0.0177631
		0.80	0.9686	0.0329844
		0.95	0.9162	0.0500529
	1.40	0.30	1.4018	0.0012170
		0.60	1.4000	0.0009408
		0.80	1.3996	0.0008128
		0.95	1.3998	0.0007406
	2.10	0.30	1.5625	0.0003637
		0.60	1.7534	0.0000530
		0.80	1.9103	0.0000106
		0.95	2.0489	0.0000026
	MLE		1.3995	0.001082
2000	0.90	0.30	1.1944	0.0064560
		0.60	1.0475	0.0298960
		0.80	0.9681	0.0687878
		0.95	0.9161	0.1189907
	1.40	0.30	1.3912	0.0008157
		0.60	1.3949	0.0007593
		0.80	1.3974	0.0007268
		0.95	1.3993	0.0007047
	2.10	0.30	1.5447	0.0001643
		0.60	1.7420	0.0000196
		0.80	1.9042	0.0000034
		0.95	2.0474	0.0000007
	MLE		1.3803	0.000834

Table 3. Failure times in hours for aircraft generator.

Failure	Time	Failure	Time
1	55	8	1308
2	166	9	2050
3	205	10	2453
4	341	11	3115
5	488	12	4017
6	567	13	4596
7	731		

Table 4. Bayes estimates with strategy 1 for aircraft generator data.

Prior guess		Bayes estimates	
$g_{\beta,1}$	$g_{\beta,2}$	$\hat{\beta}_{\text{Bayes}}$	$\hat{\lambda}_{\text{Bayes}}$
	0.075	0.3583	0.2561
0.25	0.15	0.4646	0.2642
	0.225	0.5123	0.2457
	0.15	0.5350	0.1054
0.5	0.30	0.5555	0.1730
	0.45	0.5623	0.1959
	0.225	0.6402	0.0604
0.75	0.45	0.5943	0.1441
	0.675	0.5812	0.1797
	MLE	0.5690	0.1076

Table 5. Bayes estimate with strategy 2 for aircraft generator data.

Prior guess		Bayes estimates	
$g_{\beta,1}$	ξ	$\hat{\beta}_{\text{Bayes}}$	$\hat{\lambda}_{\text{Bayes}}$
0.25	0.30	0.4115	0.5399
	0.60	0.3223	0.9559
	0.80	0.2816	1.2621
	0.95	0.2572	1.4992
0.5	0.30	0.5464	0.2120
	0.60	0.5255	0.2041
	0.80	0.5124	0.1981
	0.95	0.5031	0.1934
0.75	0.30	0.6134	0.1355
	0.60	0.6653	0.0735
	0.80	0.7051	0.0439
	0.95	0.7383	0.0277
	MLE	0.5690	0.1076

Reiser 1992). Those are failure times in hours for a complex type of aircraft generator.

The MLE for β and λ are easily obtained: $\beta_{\mathrm{MLE}} = 0.5690$ and $\lambda_{\mathrm{MLE}} = 0.10756$. We compare the MLE with the Bayes estimates in Table 3, for strategy 1 and in Table 4 for strategy 2. Strategy 1 leads to an estimate close to the MLE when the guess on β is 0.5, with a small standard deviation. Strategy is unable what ever be the guess to provide estimate close to the MLE. Again the only case where β is close to the MLE is when $g_\beta = 0.5$. The comments are very similar to those for simulated data.

6 CONCLUDING REMARKS

We introduce in this work a new distribution: the H-B distribution. This distribution is a natural conjugate prior to make Bayesian inference on the PLP. More investigations concerning the properties of this distribution need to be carried out. In particular a better understanding of the properties will be helpful to elicit prior parameters. We suggest two strategies that are easy to implement, relying on expert guessing. The results show that the choice of the elicitation strategy is very sensitive. More need to be done in order to improve the accuracy of the estimates. Other strategies should be investigated. We are working in this direction in the present time.

REFERENCES

Bar-Lev, S., I. Lavi, & B. Reiser (1992). Bayesian infernce for the power law process. *Ann; Inst. Statist. Math. 44*(4), 623–639.

Guida, M., R. Calabria, & G. Pulcini (1989). Bayes inference for non-homogenuous Poisson process with power intensity law. *IEEE Transactions on Reliability 38*, 603–609.

Huang, Y.S. & V.B. Bier (1998). A natural conjugate prior for non-homogeneous Poisson process with power law intensity function. *Communications in Statistics-Simulation and Computation 27*, 525–551.

Oliveira, M.D. and Colosimo, E.A. & G.L. Gilardoni (2012). Bayesian inference for power law process with applications in repairable systems. *Journal of Statistical Planning and Inference 142*, 1151–1160w.

Sen, A. & R. Khattree (1998). On estimating the current intensity of failure for the power law process. *Journal of Statistical Planning and Inference 74*, 252–272.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Bayesian approach to estimate the mixture of failure rate model

R. Briš

Department of Applied Mathematics, Faculty of Electrical Engineering and Computer Science, VSB—Technical University Ostrava, Ostrava, The Czech Republic

T.T. Thach

Ton Duc Thang University, Faculty of Mathematics and Statistics, Ho Chi Minh City, Vietnam

ABSTRACT: Engineering systems are subject to continuous stresses and shocks which may (or may not) cause a change in the failure pattern of the system with unknown probability. A mixture of failure rate models can be used as representation of frequent realistic situations, the failure time distribution is given in the corresponding case. Classical and Bayesian estimation of the parameters and reliability characteristics of this failure time distribution is the subject matter of the paper, where particular emphasis is put on Weibull wear-out failure model.

1 INTRODUCTION

Engineering systems, while in operation, are always subject to environmental stresses and shocks which may or may not alter the failure rate function of the system. Suppose p is the unknown probability that the system is able to bear these stresses and its failure model remains unaffected, and q is the probability of the complementary event. In such situations, a failure distribution is generally used to describe mathematically the failure rate on the system. To some extent, the solution to the proposed problem is attempted through the mixture of distributions (Mann et al. 1974, Sinha 1986, Lawless 1982). However, in this regard we are faced with two problems. Firstly, there are many physical causes that individually or collectively cause the failure of the system or device.

At present, it is not possible to differentiate between these physical causes and mathematically account for all of them, and, therefore, the choice of a failure distribution becomes difficult. Secondly, even if a goodness of fit technique is applied to actual observations of time to failure, we face a problem arising due to the non-symmetric nature of the life-time distributions whose behaviour is quite different at the tails where actual observations are sparse in view of the limited sample size (Mann et al. 1974). Obviously, the best one can do is to look out for a concept which is useful for differentiating between different life-time distributions. Failure rate is one such concept in the literature on reliability. After analyzing such physical considerations of the system, we can formulate a mixture of failure rate functions which, in turn, provide the failure time distributions. In view of the above, and due to continuous stresses and shocks on the system, let us suppose that the failure rate function of a system remain unaltered with probability p, and it undergoes a change with probability q. Let the failure rate function of the system in these two situations be in either of the following two states (Sharma et al. 1997):

1.1 *State 1*

Initially it experiences a constant failure rate model and this model may (or may not) change with probability $q(p = 1 - q)$.

1.2 *State 2*

If the stresses and shocks alter the failure rate model of the system with probability q, then it experiences a wear-out failure model. In comparison with Sharma et al. (1997), this study brings distinctive generalization of the state by implementation of a new parameter, which enables to take into account also more general Weibull model.

In probability theory and statistics, the Weibull distribution is a continuous probability distribution, which is named after the Waloddi Weibull. As a result of flexibility in time-to-failure of a very widespread diversity to versatile mechanisms, the two-parameter Weibull distribution has been recently used quite extensively in reliability and survival analysis particularly when the data are not censored. Much of the attractiveness of the Weibull distribution is due to the wide variety of shapes which can assume by altering its parameters.

Using such a failure rate pattern, the characterization of life-time distribution in the corresponding situation is given. Various inferential properties of this life-time distribution along with the estimation of parameters and reliability characteristics is the subject matter of the present study. Since the estimates based on the operational data can be updated by incorporating past environmental experiences on the random variations in the life-time parameters (Martz & Waller 1982), therefore, the Bayesian analysis of the parameters and other reliability characteristics is also given.

2 BACKGROUND

Let

T: the random variable denoting life-time of the system.

$h(t)$: the failure rate function.

$f(\mathrm{t})$: the probability density function (p.d.f.) of T.

$F(t)$: the cumulative distribution function of T.

$R(t) = \mathbb{P}(T > t)$: the reliability/survival function.

$\mathbb{E}(t) = \int_0^\infty R(t)dt$: Mean Time To Failure (MTTF).

3 ASSUMPTIONS

Let

p: the probability of the event A, that the system is able to bear the stresses and shocks and its failure pattern remains unaltered.

$q = 1 - p$: the probability of the complementary event A^c.

Further, let, the mixture of the failure rate function be

$$h(t) = p\lambda + (1-p)\lambda t^k, \quad \lambda, t > 0, 0 < p < 1 \tag{1}$$

for

1. $p = 1$; represents the failure rate of an exponential distribution.
2. $k = 1$ and $p = 0$; represents the failure rate of the Rayleigh distribution or Weibull distribution with shape parameter 2.
3. $k = 1$; represents the linear ageing process.
4. $0 < k < 1$; represents the concave ageing process.
5. $k > 1$; represents the convex ageing process.

At the beginning of the research our study distinguishes between 3 different practical situations: Case 1 – when p is known, Case 2 – when λ is known, and Case 3 – when both λ and p are unknown.

In Weibull reliability analysis it is frequently the case that the value of the shape parameter is known (Martz & Waller 1982). For example, the Raleigh distribution is obtained when $k = 1$. The earliest references to Bayesian estimation of the unknown scale parameter are in Harris & Singpurwalla (1968). Since that time this case has been considered by numerous authors, see Sharma et al. (1997), Canavos (1974), Moore & Bilikam (1978), Tummala & Sathe (1978), Alexander Aron et al. (2009) & Aslam et al. (2014). This study is free continuation and generalization of the research from Sharma et al. (1997).

4 CHARACTERISTICS OF THE LIFE-TIME DISTRIBUTION

Using the well-known relationship

$$f(t) = h(t)\exp\left\{-\int_0^t h(x)dx\right\} \tag{2}$$

and in view of equation (1), the p.d.f. of the life-time T is

$$f(t) = \begin{cases} (p\lambda + (1-p)\lambda t^k)\exp\left\{-\left(p\lambda t + \dfrac{\lambda(1-p)}{k+1}t^{k+1}\right)\right\}, \\ 0, \qquad \text{otherwise.} \end{cases} \quad t > 0 \tag{3}$$

The reliability function is

$$R(t) = \exp\left\{-\left(p\lambda t + \frac{\lambda(1-p)}{k+1}t^{k+1}\right)\right\}, \quad t > 0. \tag{4}$$

The MTTF is given by

$$\mathbf{MTTF} = \mathbb{E}(T) = \int_0^\infty R(t)dt$$
$$= \int_0^\infty \exp\left\{-\left(p\lambda t + \frac{\lambda(1-p)}{k+1}t^{k+1}\right)\right\}dt \qquad (5)$$

This integral can be obtained by using some suitable numerical methods.

5 ESTIMATION OF PARAMETERS AND RELIABILITY CHARACTERISTICS

Let $t_1, t_2, \ldots, t_n$ be the random failure times of n items under test whose failure time distribution is as given in equation (3). Then the likelihood function is

$$L(t_1, t_2, \ldots, t_n \mid \lambda, p) = \lambda^n \left[\prod_{i=1}^{n}(p + (1-p)t_i^k)\right]$$
$$\times \exp\left\{-\lambda \sum_{i=1}^{n}\left(pt_i + \frac{1-p}{k+1}t_i^{k+1}\right)\right\}. \qquad (6)$$

5.1 *MLE's*

5.1.1 *Case 1: When p is known*

To find the MLE of λ, say $\hat{\lambda}$, we consider

$$\log L(t_1, t_2, \ldots, t_n \mid \lambda, p)$$
$$= n\log\lambda + \sum_{i=1}^{n}\log\left(p + (1-p)t_i^k\right)$$
$$-\lambda\sum_{i=1}^{n}\left(pt_i + \frac{1-p}{k+1}t_i^{k+1}\right). \qquad (7)$$

Now,

$$\frac{\partial \log L(t_1, t_2, \ldots, t_n \mid \lambda, p)}{\partial\lambda} = 0 \qquad (8)$$

gives

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n}\left(pt_i + \frac{1-p}{k+1}t_i^{k+1}\right)}. \qquad (9)$$

By using the invariance property of MLE's,

1. The MLE for $R(t)$, say $\hat{R}_1(t)$, will be

$$\hat{R}_1(t) = \exp\left\{-\hat{\lambda}\left(pt + \frac{1-p}{k+1}t^{k+1}\right)\right\}. \qquad (10)$$

2. The MLE for $h(t)$, say $\hat{h}_1(t)$, will be

$$\hat{h}_1(t) = \hat{\lambda}(p + (1-p)t^k). \qquad (11)$$

3. The MLE for MTTF will be

$$\mathbf{M\hat{T}TF}_1 = \mathbf{MTTF}(p, \hat{\lambda}), \qquad (12)$$

which can be obtained by installing into formula (5) and integrating.

5.1.2 *Case 2: When λ is known*

To find the MLE of p, say $\hat{p}$, we consider

$$\frac{\partial \log L(t_1, t_2, \ldots, t_n \mid \lambda, p}{\partial p} = 0. \qquad (13)$$

or

$$\sum_{i=1}^{n}\left(\frac{1}{p + \frac{t_i^k}{1-t_i^k}}\right) - \frac{\lambda}{k+1}\sum_{i=1}^{n}t_i(1 + k - t_i^k) = 0. \qquad (14)$$

An estimate of p, i.e. $\hat{p}$, can be obtained from equation (14), by using some suitable numerical iteration method. By using the invariance property of MLE's,

1. The MLE for $R(t)$, say $\hat{R}_2(t)$, will be

$$\hat{R}_2(t) = \exp\left\{-\lambda\left(\hat{p}t + \frac{1-\hat{p}}{k+1}t^{k+1}\right)\right\}. \qquad (15)$$

2. The MLE for $h(t)$, say $\hat{h}_2(t)$, will be

$$\hat{h}_2(t) = \lambda(\hat{p} + (1-\hat{p})t^k). \qquad (16)$$

3. The MLE for MTTF will be

$$\mathbf{M\hat{T}TF}_2 = \mathbf{MTTF}(\hat{p}, \lambda), \qquad (17)$$

which can be obtained by installing into formula (5) and integrating.

5.1.3 *Case 3: When both λ and p are unknown*

To find the MLE of λ and p, we consider

$$\frac{\partial \log L(t_1, t_2, \ldots, t_n \mid \lambda, p)}{\partial\lambda} = 0 \qquad (18)$$

and

$$\frac{\partial \log L(t_1, t_2, \ldots, t_n \mid \lambda, p)}{\partial p} = 0. \tag{19}$$

We get

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n}\left(pt_i + \frac{1-p}{k+1}t_i^{k+1}\right)} \tag{20}$$

and

$$\sum_{i=1}^{n}\left(\frac{1}{p + \frac{t_i^k}{1-t_i^k}}\right) - \frac{n\sum_{i=1}^{n} t_i(1+k-t_i^k)}{(k+1)\sum_{i=1}^{n}\left(pt_i + \frac{1-p}{k+1}t_i^{k+1}\right)} = 0. \tag{21}$$

Equation (21) may be solve for $\hat{p}$ by Newton-Raphson or other suitable iterative methods and this value is substituted in equation (20) to obtain $\hat{\lambda}$. By using the invariance property of MLE's,

1. The MLE for $R(t)$, say $\hat{R}_3(t)$, will be

   $$\hat{R}_3(t) = \exp\left\{-\hat{\lambda}\left(\hat{p}t + \frac{1-\hat{p}}{k+1}t^{k+1}\right)\right\}. \tag{22}$$

2. The MLE for $h(t)$, say $\hat{h}_3(t)$, will be

   $$\hat{h}_3(t) = \hat{\lambda}(\hat{p} + (1-\hat{p})t^k). \tag{23}$$

3. The MLE for MTTF will be

   $$\mathbf{M\hat{T}TF}_2 = \mathbf{MTTF}(\hat{p}, \lambda), \tag{24}$$

   which can be obtained by installing into formula (5) and integrating.

6 BAYESIAN ESTIMATION

6.1 *Case 1: when p is known*

6.1.1 *Non-informative prior*

We are going to use the non-informative prior

$$\pi(\lambda) = \frac{1}{\lambda}. \tag{25}$$

The likelihood function in equation (6) may be rewritten as

$$L(t_1, t_2, \ldots, t_n \mid \lambda, p) = \lambda^n T_1 \times e^{-\lambda T_2} \tag{26}$$

$$T_1 = \prod_{i=1}^{n}(p + (1-p)t_i^k) \tag{27}$$

and

$$T_2 = \sum_{i=1}^{n}\left(pt_i + \frac{1-p}{k+1}t_i^{k+1}\right). \tag{28}$$

In view of the prior in equation (25), the posterior distribution of λ given t_1, t_2, ..., t_n is given by

$$\begin{aligned}\pi(\lambda \mid t_1, t_2, \ldots, t_n, p) &= \frac{L(t_1, t_2, \ldots, t_n|\lambda,p)\pi(\lambda)}{\int_0^\infty L(t_1, t_2, \ldots, t_n|\lambda,p)\pi(\lambda)d\lambda} \\ &= \frac{T_2^n}{\Gamma(n)}\lambda^{n-1}e^{-\lambda T_2}, \quad \lambda > 0.\end{aligned} \tag{29}$$

Therefore, the Bayes estimate of λ, say λ^*, under the square-error loss function, becomes

$$\lambda^* = \mathbb{E}(\lambda \mid t_1, t_2, \ldots, t_n, p) = \frac{n}{T_2} \quad (= \hat{\lambda}) \tag{30}$$

i.e. it reduces to the usual ML estimator, what is in agreement with Martz & Waller (1982).

Also, the Bayes estimate of $R(t)$, say $R_1^*(t)$, is

$$\begin{aligned}R_1^*(t) &= \mathbb{E}(R(t) \mid t_1, t_2, \ldots, t_n, p) \\ &= \int_0^\infty e^{-\lambda T_3}\pi(\lambda \mid t_1, t_2, \ldots, t_n, p)d\lambda \\ &= \frac{1}{\left(1 + \frac{T_3}{T_2}\right)^n}\end{aligned} \tag{31}$$

where

$$T_3 = pt + \frac{1-p}{k+1}t^{k+1}.$$

Similarly, the Bayes estimation of $h(t)$, say $h_1^*(t)$, is

$$\begin{aligned}h_1^*(t) &= \mathbb{E}(h(t) \mid t_1, t_2, \ldots, t_n, p) \\ &= \int_0^\infty \lambda(p + (1-p)t^k)\pi(\lambda \mid t_1, t_2, \ldots, t_n, p)d\lambda \\ &= \frac{(n+1)(p + (1-p)t^k)}{T_2}.\end{aligned} \tag{32}$$

6.1.2 *Informative prior*

Let the conjugate prior of λ be gamma with p.d.f.

$$\pi(\lambda) = \frac{\alpha^{\beta}}{\Gamma(\beta)} \lambda^{\beta-1} e^{-\alpha\lambda} \quad \lambda > 0, \alpha > 0, \beta > 0. \tag{33}$$

In view of the prior in equation (33), the posterior distribution of A given t_1, t_2, ..., t_n is given by

$$\begin{aligned}\pi(\lambda \mid t_1, t_2, \ldots, t_n, p) &= \frac{L(t_1, t_2, \ldots, t_n|\lambda, p)\pi(\lambda)}{\int_0^{\infty} L(t_1, t_2, \ldots, t_n|\lambda, p)\pi(\lambda)d\lambda} \\ &= \frac{(\alpha + T_2)^{n+\beta}}{\Gamma(n+\beta)} \lambda^{n+\beta-1} e^{-\lambda(\alpha+T_2)}, \\ &\quad \alpha, \beta, \lambda > 0.\end{aligned} \tag{34}$$

Therefore, the Bayes estimate of λ, say λ*, under the square-error loss function, becomes

$$\lambda^* = \mathbb{E}(\lambda \mid t_1, t_2, \ldots, t_n, p) = \frac{n+\beta}{\alpha + T_2}. \tag{35}$$

Also, the Bayes estimate of $R(t)$, say $R_1^*(t)$, is

$$\begin{aligned}R_1^*(t) &= \mathbb{E}(R(t) \mid t_1, t_2, \ldots, t_n, p) \\ &= \int_0^{\infty} e^{-\lambda T_3} \pi(\lambda \mid t_1, t_2, \ldots, t_n, p) d\lambda \\ &= \frac{1}{\left(1 + \frac{T_3}{\alpha + T_2}\right)^{n+\beta}},\end{aligned} \tag{36}$$

where

$$T_3 = pt + \frac{1-p}{k+1} t^{k+1}.$$

Similarly, the Bayes estimation of $h(t)$, say $h_1^*(t)$, is

$$\begin{aligned}h_1^*(t) &= \mathbb{E}(h(t) \mid t_1, t_2, \ldots, t_n, p) \\ &= \int_0^{\infty} \lambda(p + (1-p)t^k)\pi(\lambda \mid t_1, t_2, \ldots, t_n, p) d\lambda \\ &= \frac{(n+\beta+1)(p + (1-p)t^k)}{\alpha + T_2}.\end{aligned} \tag{37}$$

6.1.3 *Super parameter estimation*

Prior parameters α and β can be obtained by Empirical Bayes approach, what is now under study of authors of the paper.

6.2 *Case 2: when λ is known*

6.2.1 *Non-informative prior*

We are going to use the non-informative prior

$$\pi(p) = p. \tag{38}$$

The likelihood function in equation (6) may be rewritten as

$$\begin{aligned}L(t_1, t_2, \ldots, t_n \mid \lambda, p) &= \\ \lambda^n \sum_{j=0}^{n} p^{n-j} (1-p)^j k_j \times e^{-\frac{\lambda}{k+1}(pT_4 + T_5)}\end{aligned} \tag{39}$$

where

$$k_0 = 1 \tag{40}$$

$$k_j = \sum_{1 \le i_1 < i_2 < \ldots < i_j \le n} t_{i_1}^k t_{i_2}^k \ldots t_{i_j}^k, \quad j = \overline{1, n} \tag{41}$$

$$T_4 = (k+1) \sum_{i=1}^{n} t_i - \sum_{i=1}^{n} t_i^{k+1}, \tag{42}$$

and

$$T_5 = \sum_{i=1}^{n} t_i^{k+1} \tag{43}$$

or

$$\begin{aligned}L(t_1, t_2, \ldots, t_n \mid \lambda, p) &= \\ e^{-\frac{\lambda T_5}{k+1}} \sum_{r=0}^{\infty} \sum_{j=0}^{n} \frac{\lambda^{n+r}(-T_4)^r}{(k+1)^r r!} p^{n+r-j} (1-p)^j k_j.\end{aligned} \tag{44}$$

In view of the prior in equation (38), the posterior distribution of p given t_1, t_2, ..., t_n is given by

$$\begin{aligned}\pi(p \mid t_1, t_2, \ldots, t_n, \lambda) &= \frac{L(t_1, t_2, \ldots, t_n \mid \lambda, p)\pi(p)}{\int_0^1 L(t_1, t_2, \ldots, t_n \mid \lambda, p)\pi(p) dp} \\ &= \frac{\sum_{r=0}^{\infty} \sum_{j=0}^{n} \left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!} p^{n+r-j+1} (1-p)^j k_j}{\int_0^1 \sum_{r=0}^{\infty} \sum_{j=0}^{n} \left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!} p^{n+r-j+1} (1-p)^j k_j dp} \\ &= \frac{\sum_{r=0}^{\infty} \sum_{j=0}^{n} \left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!} p^{n+r-j+1} (1-p)^j k_j}{\sum_{r=0}^{\infty} \sum_{j=0}^{n} \left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!} B(n+r-j+2, j+1) k_j}\end{aligned} \tag{45}$$

Therefore, the Bayes estimate of p, say p^*, under the square-error loss function, becomes

$$p^* = E(p|t_1, t_2, \ldots, t_n, \lambda) = \int_0^1 p\pi(p \mid t_1, t_2, \ldots, t_n, \lambda)dp$$
$$= \frac{\sum_{r=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!} B(n+r-j+3, j+1)k_j}{\sum_{r=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!} B(n+r-j+2, j+1)k_j}. \tag{46}$$

The Bayes estimate of $R(t)$, say $R_2^*(t)$, is

$$R_2^*(t) = \mathbb{E}(R(t) \mid t_1, t_2, \ldots, t_n, \lambda)$$
$$= \int_0^1 \exp\left\{-\lambda\left(pt + \frac{1-p}{k+1}t^{k+1}\right)\right\}\pi(p \mid t_1, t_2, \ldots, t_n, \lambda)dp$$
$$= \frac{e-\frac{\lambda t^{k+1}}{k+1}\int_0^1 \sum_{r=0}^{\infty}\sum_{m=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!}\left(-\frac{\lambda t\left((k+1)-t^k\right)}{k+1}\right)^m \frac{1}{m!} p^{n+r+m-j+1}(1-p)^j k_j dp}{\sum_{r=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!} B(n+r-j+2, j+1)k_j} \tag{47}$$
$$= \frac{e-\frac{\lambda t^{k+1}}{k+1}\sum_{r=0}^{\infty}\sum_{m=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!}\left(-\frac{\lambda t\left((k+1)-t^k\right)}{k+1}\right)^m \frac{1}{m!} B(n+r+m-j+2, j+1)k_j}{\sum_{r=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!} B(n+r-j+2, j+1)k_j}$$

Similarly, the Bayes estimate of h(t), say $h_2^*(t)$, becomes

$$h_2^*(t) = \mathbb{E}(h(t) \mid t_1, t_2, \ldots, t_n, \lambda)$$
$$= \int_0^1 \left(p\lambda + (1-p)\lambda t^k\right)\pi\left(p \mid t_1, t_2, \ldots, t_n, \lambda\right)dp$$
$$= \lambda(1-t^k)p^* + \lambda t^k$$
$$= \lambda(p^* + (1-p^*)t^k), \tag{48}$$

where p^* is given in equation (46).

6.2.2 *Informative prior*
Let the prior distribution of p be a Beta distribution with p.d.f.

$$\pi(p) = \frac{1}{B(a,b)}p^{a-1}(1-p)^{b-1}, \quad a, b > 0, 0 < p < 1. \tag{49}$$

In view of the prior in equation (49), the posterior distribution of p given t_1, t_2,..., t_n is given by

$$\pi(p \mid t_1, t_2, \ldots, t_n, \lambda)$$
$$= \frac{L(t_1, t_2, \ldots, t_n \mid \lambda, p)\pi(p)}{\int_0^1 L(t_1, t_2, \ldots, t_n \mid \lambda, p)\pi(p)dp}$$
$$= \frac{\sum_{r=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!} p^{n+r+a-j-1}(1-p)^{b+j-1}k_j}{\int_0^1 \sum_{r=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!} p^{n+r+a-j-1}(1-p)^{b+j-1}k_j dp}$$
$$= \frac{\sum_{r=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!} p^{n+r+a-j-1}(1-p)^{b+j-1}k_j}{\sum_{r=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!} B(n+r+a-j, b+j)k_j} \tag{50}$$

Therefore, the Bayes estimate of p, say p^*, under the square-error loss function, becomes

$$p^* = E(p \mid t_1, t_2, \ldots, t_n, \lambda)$$
$$= \int_0^1 p\pi(p \mid t_1, t_2, \ldots, t_n, \lambda)dp$$
$$= \frac{\sum_{r=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!} B(n+r+a-j+1, b+j)k_j}{\sum_{r=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!} B(n+r+a-j, b+j)k_j} \tag{51}$$

The Bayes estimate of $R(t)$, say $R_2^*(t)$, is

$$
\begin{aligned}
R_2^*(t) &= \mathbb{E}(R(t)\,|\,t_1,t_2,\ldots,t_n,\lambda) \\
&= \int_0^1 \exp\left\{-\lambda\left(pt+\frac{1-p}{k+1}t^{k+1}\right)\right\}\pi(p\,|\,t_1,t_2,\ldots,t_n,\lambda)dp \\
&= \frac{e-\frac{\lambda t^{k+1}}{k+1}\int_0^1 \sum_{r=0}^{\infty}\sum_{m=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!}\left(-\frac{\lambda t\left((k+1)-t^k\right)}{k+1}\right)^m \frac{1}{m!}p^{n+r+a+m-j-1}(1-p)^{b+j-1}k_j dp}{\sum_{r=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!}B(n+r+a-j,b+j)k_j} \\
&= \frac{e-\frac{\lambda t^{k+1}}{k+1}\sum_{r=0}^{\infty}\sum_{m=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!}\left(-\frac{\lambda t\left((k+1)-t^k\right)}{k+1}\right)^m \frac{1}{m!}B(n+r+a+m-j,b+j)k_j}{\sum_{r=0}^{\infty}\sum_{j=0}^{n}\left(\frac{-\lambda T_4}{k+1}\right)^r \frac{1}{r!}B(n+r+a-j,b+j)k_j}
\end{aligned}
\tag{52}
$$

Similarly, the Bayes estimate of h(t), say $h_2^*(t)$, becomes

$$
\begin{aligned}
h_2^*(t) &= \mathbb{E}(h(t)\,|\,t_1,t_2,\ldots,t_n,\lambda) \\
&= \int_0^1 \left(p\lambda+(1-p)\lambda t^k\right)\pi\left(p\,|\,t_1,t_2,\ldots,t_n,\lambda\right)dp \\
&= \lambda(1-t^k)p^* + \lambda t^k \\
&= \lambda(p^* + (1-p^*)t^k),
\end{aligned}
\tag{53}
$$

where p^* is given in equation (51).

6.2.3 *Super parameter estimation*

Prior parameters a and b can be obtained by Empirical Bayes approach, what is now under study of authors of the paper.

7 AN EXAMPLE

In case $k = 1$, $p = 1/3$, λ = 1/5, the failure rate function is

$$h(t) = \frac{1}{15} + \frac{2}{15}t, \qquad t > 0, \tag{54}$$

and the p.d.f. is

$$f(t) = \begin{cases} \frac{1}{15}(1+2t)\exp\left\{-\frac{1}{15}(t+t^2)\right\}, & > 0 \\ 0, & \text{otherwise} \end{cases} \tag{55}$$

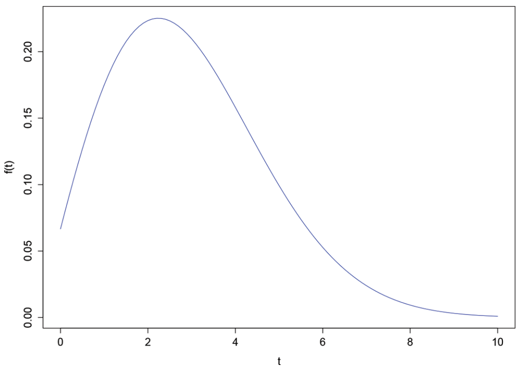


Figure 1. Density curve of T when $k = 1$, $p = 1/3$, $\lambda = 1/5$.

Table 1. Generated data values.

3.6145404	1.2608646	1.9644814	4.5339338
2.1762887	1.7987543	6.7039167	1.1689263
4.5634580	1.3711159	2.7837231	4.7787410
2.3456293	2.1053338	5.0590295	3.6569535
1.8818134	5.2699757	5.9550330	2.8941610
2.4522999	0.8210216	0.8628159	0.4675159
3.7480263	4.4549035	3.3256436	0.8107626
2.4162468	2.4680479		

Figure 1 shows the curve of p.d.f. in (55). Table 1 shows data generated from this p.d.f. with sample size $n = 30$ and the corresponding histogram brings Fig. 2.

Table 2 shows the estimated values for p; λ and MTTF using data in Table 1. Figs. 3–5 demonstrate estimated reliability functions for different cases in comparison with true reliability function computed by formula (4). We can see that Bayes

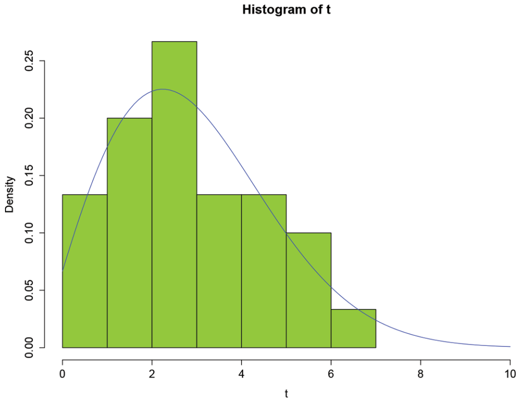


Figure 2. Histogram for generated data.

Table 2. Estimated values for parameters using data in Table 1.

True	$p=\frac{1}{3}=0.3333,\ \lambda=\frac{1}{5}=0.2$, **MTTF** = 2.9844			
Estimate	MLE		Bayes	
Case 1	$p=\frac{1}{3}$	$\hat{\lambda}=0.2124$ $\hat{\mathbf{MTTF}}=2.8841$	$p=\frac{1}{3}$	$\lambda^*=0.2124$ $\hat{\mathbf{MTTF}}=2.8841$
Case 2	$\lambda=\frac{1}{5}$	$\hat{p}=0.1332$ $\hat{\mathbf{MTTF}}=2.8624$	$\lambda=\frac{1}{5}$	$p^*=0.34818$ $\hat{\mathbf{MTTF}}=2.9955$
Case 3	$\hat{p}=0.0080$	$\hat{\lambda}=0.1793$ $\hat{\mathbf{MTTF}}=2.9637$		

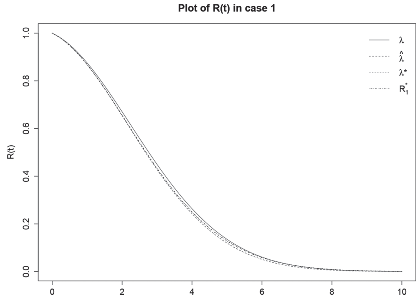


Figure 3. Plot of reliability functions in case 1.

approximation of reliability function is most closely to the true function. Figs. 6–8 show the estimated failure rate functions in comparison with the true function computed by formula (1). Bayes approximation looks promisingly as well, in comparison with MLE method.

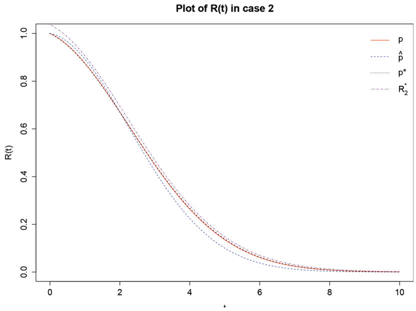


Figure 4. Plot of reliability functions in case 2.

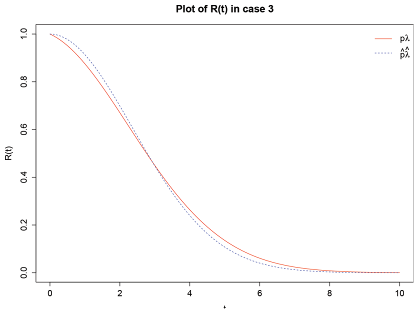


Figure 5. Plot of reliability functions in case 3.

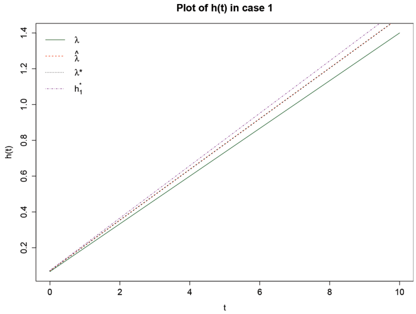


Figure 6. Plot of failure rate functions in case 1.

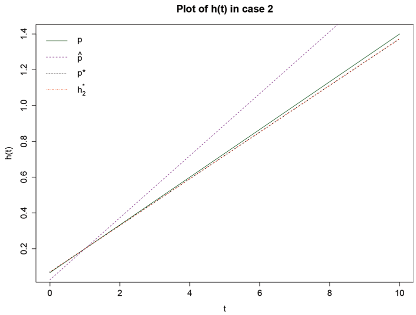


Figure 7. Plot of failure rate functions in case 2.

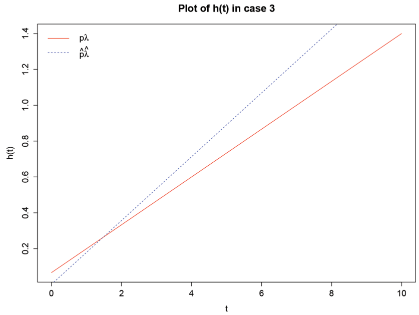


Figure 8. Plot of failure rate functions in case 3.

8 CONCLUSION

Our study shows that Bayesian approach can show better results than MLEs method to estimate the mixture of failure rate model. The study reveals an interesting fact that non-informative prior seems to be working well, especially in case 2. This result would be good platform for selection of non-informative prior, in case of no information about prior. This work gives good evidence that Bayes approach is viable method to model real situations which can be approximated by mixture of failure rate functions.

ACKNOWLEDGEMENT

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4 Innovations excellence in science–LQ1602".

REFERENCES

Alexander Aron, A., H. Guo, A. Mettas, & D. Ogden (2009). Improving the 1-parameter weibull: A bayesian approach. *IEEE*.

Aslam, M., S.M.A. Kazmi, I. Ahmad, & S.H. Shah (2014). Bayesian estimation for parameters of the weibull distribution. *Sci.Int.(Lahore) 26(5)*, 1915–1920.

Canavos, G. (1974). On the robustness of a bayes estimate. *Annual Reliability and Maintainability Symposium*, 432–435.

Harris, C. & N. Singpurwalla (1968). Life distributions derived from stochastic hazard functions. *IEEE Trans, Reliab. 17*, 70–79.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: John Wiley and Sons.

Mann, N., E. Schaffer, & N. Singpurwalla (1974). *Methods for Statistical Analysis ofReliability and Life Data*. NY: Wiley.

Martz, H.F. & R.A. Waller (1982). *Bayesian Reliability Analysis*. New York: John Wiley and Sons.

Moore, A.H. & J.E. Bilikam (1978). Bayesian estimation of parameters of life distributions and reliability from type ii censored samples. *IEEE Trans, Reliab*. 27, 64–67.

Pandey, A., A. Singh, & W.J. Zimmer (1993). Bayes estimation of the linear hazard-rate model. *IEEE Trans, Reliab. 42*.

Sharma, K.K., H. Krishna, & B. Singh (1997). Bayes estimation of the mixture of hazard rate model. *Reliability Engineering and System Safety 55*, 9–13.

Sinha, S.K. (1986). *Reliability and Life Testing*. USA: Wiley Eastern Ltd.

Tummala, V.M.R. & P.T. Sathe (1978). Minimum expected loss estimators of reliability and parameters of certain lifetime distributions. *IEEE Trans, Reliab. 27*, 283–285.

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Cost oriented statistical decision problem in acceptance sampling and quality control

R. Briš
Department of Applied Mathematics, Faculty of Electrical Engineering and Computer Science,
VŠB—Technical University Ostrava, Ostrava, The Czech Republicc

ABSTRACT: Statistical decision problem on the basis of Bayes approach is presented in the paper. Bayes inference for hypothesis testing is introduced in general and its extension in acceptance sampling and quality control as well. Common decision problem for a quality characteristic is described taking into account the cost dependent on the decision. Asymmetric loss function is proposed to quantify cost oriented decision and mainly, to set the optimal Bayes decision rule for acceptance sampling plan. Application of the method on real example from industry is clearly demonstrated.

1 INTRODUCTION

1.1 *Acceptance sampling in modern industry*

In the mass production environments of the twentieth century, quality control rested on the triple sorting principle of inspecting product, detecting inappropriate product, and sorting out inappropriate product. Accordingly, Statistical Product Inspection (SPI) based on sampling was the first strong branch of modern statistical quality control. The sorting principle is clearly expressed by the definition of the purpose of sampling inspection given in Dodge & Roming (1959): "The purpose of the inspection is to determine the acceptability of individual lots submitted for inspection, i.e., sorting good lots from bad." SPI is a tool of product quality control, where the term product is used in a broad sense to include any completed identifiable entity like materials, parts, products, services, or data. Product control is located at the interface of two parties: a supplier or vendor or producer as the party providing product and a customer or consumer as the party receiving product. Traditional product control relied on the sorting principle at the interface of supplier and consumer. Modern product control primarily relies on quality provision and prevention at the interface. The particular task of the interface is to integrate supplier and consumer into joint and concerted efforts to create and maintain favorable conditions for quality in design, manufacture, handling, transport, and service. In this scheme, inspection may be useful, however not as the primary tool of product control, but in an adjunct role, subordinate to the guiding principles of provision and prevention. The adjunct role of SPI is expressed in (Godfrey & Mundel 1984): "The shipper must understand the acceptance sampling plan in order to be sure that the quality control of the production process provides lots of a quality sufficient to meet his objective. This is the basic principle of acceptance sampling—to assure the production of lots of acceptable quality". The important role of SPI is:

- it serves as an supplementary source of information on the product quality level
- it is used to learn about the product quality level if the supplier-consumer interface is new and sufficient quality records do not exist
- it serves as a precaution for the consumer, if the supplier-consumer interface is new or if the quality efforts of the supplier are not completely reliable.

SPI is implemented according to a sampling plan. A single sample plan is defined by a pair (n, r) consisting of the sample size n and the acceptance number r and a sample statistic t. The decision rule of a single sampling plan prescribes the following steps: a) n items are sampled at random; b) the value of the sample statistic t is calculated from the elements of the sample; and c) acceptance is warranted if $t \leq r$, otherwise the decision is rejection.

In acceptance sampling schemes, it is the usual practice for acceptance criteria to be specified in terms of some parameter or parameters of the distribution of the quality characteristic or characteristic which are the basis for acceptance. For example, a minimum value for the percentage of non-functioning items or a minimum value for the mean operating life of some component may be specified. In a classical acceptance sampling scheme, sample test data is used to construct a hypothesis test for the specified parameter. If the hypothesis is not rejected and the sample is accepted, the distribution of the quality characteristic can be calculated from the null distribution. Therefore the null distribution can be used to calculate the number of individual items which will not function or will fail to meet some minimum standard. In practice, the acceptable null standard for the test parameter will have been determined by the performance requirements of individual items. For example, to insure that only a small number of individual items have an operating life less than some minimum standard, the minimum acceptable mean failure time must be considerably higher than this minimum standard for individual items. It is interesting to note that if the failure time has an exponential distribution then fully 63.2% or almost two thirds of individual items will fail before the mean failure time. However, in the classical acceptance sampling scheme, the minimum acceptable parameter values will have been chosen to insure that if the sample is accepted, only a small number of individual items will fail to meet minimum performance standards.

However when prior information is available for the relevant parameter values and a Bayesian procedure is employed, acceptance criteria are not so easy to establish. It is possible to compute a posterior parameter distribution which gives the complete current probability information about the parameter. However, the relation between a posterior distribution for the parameter and the likely distribution of the quality characteristic of individual items is much more complex than the relation of the individual quality characteristic to a single null parameter value. For example, merely requiring the mean of the posterior parameter distribution to meet some specified standard is not sufficient to determine the proportion of individual items which will perform satisfactorily. Other characteristics of the posterior distribution, notably the variance, are also of importance. Even specifying the posterior probability of the parameter lying in some critical range is not sufficient to guarantee the proportion of individual items which will perform satisfactorily. Such naive criteria can produce inconsistent acceptance decisions.

In most cases as well as in this paper, decision in product control is directed toward a single assembled lot of product. A common decision situation is determining whether a product meets a pre-determined quality specification. The decision is to either accept or reject the assembled lot of product, so superficially it would appear to be a simple hypothesis testing problem. However, both a null and alternate hypotheses are composite, embodying many parameter values each. The true loss is likely to depend on the actual parameter value rather than whether parameter value exceeds some precisely defined specification.

In the "Bayesian approach" the best sampling plan is defined as the one which minimizes the average cost. This is essentially the paradigm of Bayesian decision theory, see (Berger 1993). However, it is not restricted to costs in a monetary sense. It only requires a numerical goal function that measures the consequences of possible decision. Bayesian sampling plans in context with optimal Bayes decision rule will be introduced in the paper, applied on real example from industry.

1.2 *Statistical decision problem*

Consider a problem in which a Decision Maker (DM) must choose a decision from some class of available decisions, and suppose that the consequences of this decision depend on the unknown value θ of some parameter Θ. We use the term «parameter» here in a very general sense, to represent any variable or quantity whose value is unknown to DM but is relevant to his or her decision: some authors refer to Θ as the «unknown state of nature ». The set Ω of all possible values of Θ is called the *parameter space*. The set D of all possible decisions *d* that DM might make in the given problem is called the *decision space.*

Conceptually, every combination of state of nature (θ) and decision (*d*) will be associated with some loss. In practical problems of inference it is usually not possible to specify the loss precisely. However, in general we can say that the loss increases with the distance between decision and state of nature. This conceptual loss function is often sufficient to assess known inference procedures and eliminate those which are obviously flawed or unacceptable according to decision theoretic criteria. Thus when applying decision theory to problems of estimation, it is common to assume a quadratic loss function.

$$\ell(\theta, d) = k(\theta - d)^2 \tag{1}$$

and for hypothesis testing problems to assign relative losses to Type I and Type II errors and no loss to the action of choosing the correct hypothesis.

		State of Nature	
		H_0	H_1
Decision	H_0	0	l_2
	H_1	l_1	0

For the purposes of applying decision theory to the problem of hypothesis testing, it is often sufficient to know only whether

$l_1 < l_2$ or $l_1 > l_2$

Suppose that Θ has a specified probability distribution g(θ). An optimal or *Bayes decision* with respect to the distribution g(θ) will be a decision *d** for which the expected loss E(ℓ| g(θ), *d*) is a minimum. For any given distribution g(θ) of Θ the expected loss E(ℓ| g(θ), *d*) is called the *risk* R(g, *d*), of the decision *d*. The risk of the Bayes decision, i.e. the minimum $R_0(g)$ of R(g,*d*) over all decisions: $d \in D$, is called the *Bayes risk*.

In many decisions problems, the DM has the opportunity, before choosing a decision in D, of observing the value of a random variable or random vector X that is related to the parameter Θ. The observation of X provides information that may be helpful to the DM in choosing a good decision. We shall assume that the conditional distribution of X given Θ = θ can be specified for every value of θ, for which $\Theta \in \Omega$. A problem of this type is called a *statistical decision problem*.

Thus the components of a statistical decision problem are a parameter space Ω, a decision space D, a loss function ℓ, and a family of conditional densities f(x|θ) of an observation X whose value will be available to the DM when he or she chooses a decision. In a statistical decision problem, the probability density distribution of Θ is called its prior distribution because it is the distribution of Θ before X has been observed. The conditional distribution of Θ given the observed value X = x is then called the posterior distribution of Θ.

1.3 *The posterior distribution h(θ | x)*

By applying Bayes theorem to the conditional and prior distributions, a posterior distribution for the state of nature can be derived.

$$h(\theta|x) = \frac{f(x|\theta)g(\theta)}{p(x)}, \qquad p(x) = \int_{\Omega} f(x|\theta')g(\theta')d\theta' \tag{2}$$

Note: If g(θ) is actually a discrete probability function, the integral should be replaced by a sum.

Since the posterior distribution of the state of nature depends on the information, X, it is reasonable to expect that the information or data will affect the decision which is chosen as optimal. The relation between information and decision is called the *decision function d(x)*. The decision function is really a statistic. A decision function is a rule that specifies a decision *d*(x), $d(x) \in D$, that will be chosen for each possible observed value x of X.

In inferential problems of estimation, it is the estimator. In problems of hypothesis testing, it is the rule for rejecting the null hypothesis, or equivalently, the critical region. In the presence of information, the decision problem becomes choosing a decision function rather than choosing a single decision.

The risk function R(*θ,d*) is the equivalent of the loss function in the presence of information. The risk function for any decision function *d*, is the expected loss of using that decision function when the state of nature is θ.

$$R(\theta, d) = \int_{X} \ell[\theta, d(x)] f(x|\theta)dx \tag{3}$$

The decision theoretic problem is to find a decision function that is optimal.

The risk R(*g,d*) of any decision function *d*(x) with respect to the prior distribution g(θ) is the expected loss (or expected value of risk function R(θ, *d*) with respect of g(θ):

$$R(g, d) = E\{\ell[\theta, d(X)]\} = \int_{\Omega} R(\theta, d)g(\theta)d\theta = \iint_{\Omega X} \ell[\theta, d(x)] f(x|\theta)g(\theta)dxd\theta \tag{4}$$

We recall that decision function *d**(x) for which the risk R(g, *d*) is minimum, is called a *Bayes decision function* with respect to g and its risk R(g, *d**) is called the *Bayes risk*. We see that the Bayes decision function is equivalent to the Bayes decision.

After the value x has been observed, the DM simply chooses a Bayes decision *d**(x) with respect to posterior distribution h (θ | x) rather than the prior distribution g(θ).

When observed value x is known, the DM can choose a Bayes decision *d**(x) with respect to the posterior distribution. At this stage of the decision—making process, the DM is not interested in the risk R(g,*d*), which is an average over all possible observations, but in the *posterior risk*:

$$\int_{\Omega} \ell[\theta, d^*(x)] h(\theta|x)d\theta, \tag{5}$$

Which is the risk from the decision he or she is actually making.

1.4 *Application to point estimation*

For problems of estimation, it is common to assume a quadratic loss function. In this case, the estimator becomes the mean of the posterior distribution of θ. It can be easy shown (O'Hagan 1994, Weerahandi 1995) that under quadratic loss the optimal Bayes estimate is the posterior mean. When a uniform prior is used, the posterior distribution is proportional to the likelihood function of θ.

$$h(\theta|x)=\frac{f(x|\theta)}{p(x)}\propto f(x|\theta)=l(x|\theta) \tag{6}$$

Thus, the optimal Bayes estimator is the first moment of the likelihood function. If the likelihood is symmetric about its maximum as the case for a normal likelihood, the optimal Bayes estimator is then the Maximum Likelihood Estimator (MLE). Important results concerning an alternative asymmetric precautionary loss functions, presenting a general class of precautionary loss functions with the quadratic loss functions as a special case, introduces Norstrom (1996).

2 BAYES DECISION MAKING IN HYPOTHESIS TESTING

2.1 *Application to hypothesis testing*

Now consider a hypothesis H, that $\theta \in \Omega_0 \subset \Omega$. Inference consists of accepting or rejecting H. A correct inference will result in zero loss. The loss associated with an incorrect inference may depend on the kind of error. Let d_0 be the inference to accept H and d_1 be the inference to reject H.

D = $\{d_0, d_1\}$. Let $\Omega_1 = (\Omega_0)^C$. Then we let

$$\begin{aligned}\ell(\theta,d_i) &= 0 \quad \text{if } \theta\in\Omega_i \\ &= a_i \quad \text{if } \theta\notin\Omega_i\end{aligned} \tag{7}$$

for $i = 0,1$
Therefore

$$\mathrm{E}(\ell(\theta,d_i)) = \mathrm{a}_i \cdot \mathrm{P}(\theta\notin\Omega_i) \tag{8}$$

and the optimal inference (O'Hagan 1994) is to reject H if

$$\mathrm{E}(\ell(\theta,d_1)) < \mathrm{E}(\ell(\theta,d_0)), \tag{9}$$

i.e. if $\mathrm{a}_1 \cdot \mathrm{P}(\theta \in \Omega_0) < \mathrm{a}_0 \cdot \{1 - \mathrm{P}(\theta \in \Omega_0)\}$, or (10)

$$\mathrm{P}(\theta\in\Omega_0) \prec \frac{\mathrm{a}_0}{\mathrm{a}_0+\mathrm{a}_1} \tag{11}$$

Therefore, we reject H if its probability is sufficiently low. The critical value on right side of the equation (11) depends on the relative seriousness of the kinds of error, as measured by $\mathrm{a}_0/\mathrm{a}_1$.

We also in following adopt a conservative strategy of rejecting H when its probability is small. In terms of Bayesian decision theory, this strategy implies that the loss for Type I error is much greater than for Type II error.

This formulation of the hypothesis testing problem is very simple because we have expressed the loss function in very stark terms. If we choose d_0 and it turns out that $\theta \in \Omega_1$, then the inference is wrong and incurs a loss of a_0 regardless of where θ lies in Ω_1. In a practical decision problem, as mentioned below (cost oriented acceptance sampling), it may be more appropriate to make a_0 and a_1 functions of θ. Then $\mathrm{E}(\lambda(\theta, d_i))$ will not have a simple formula, but may be computed for $i = 0,1$ and an optimal inference thereby selected.

3 BAYES DECISION MAKING IN ACCEPTANCE SAMPLING AND QUALITY CONTROL

3.1 *Acceptance sampling*

Acceptance sampling relates to the acceptance or rejection of a product or process on the basis of SPI. SPI is the process of evaluating the quality of a product by inspecting some, but not all of them. Its methods constitute decision rules for the disposition or sentencing of the product sampled. In this sense it may be contrasted with survey sampling, the purpose of which is largely estimation.

Sampling plans, which specify sample size and acceptance criteria, are fundamental to acceptance sampling.

In a classical acceptance sampling scheme, sample test data is used to construct a hypothesis test for the specified parameter. If the hypothesis is not rejected and the sample is accepted, the distribution of the quality characteristic (as is for example time to failure) can be calculated from the null distribution.

However when prior information is available for the relevant parameter values and a Bayesian procedure is employed, acceptance criteria are not so easy to establish. The relation between a posterior distribution for the parameter and the likely distribution of the quality characteristic of individual items is much more complex than the relation of the individual quality characteristic to a single null parameter value (Bris 2002). However, using Bayes approach we can take into account the cost oriented statistical decision problem demonstrated below, i.e. the loss associated with over-estimation (below specified quality) and the loss associated with under-estimation (higher than specified quality).

3.2 *Description of a common decision problem for quality characteristic*

A common decision theory situation is determining whether a product meets a pre-determined quality specification. The decision is to either accept or reject the product, so superficially it would appear to be a simple hypothesis testing problem. However, both a null and alternate hypotheses are composite, embodying many parameter values each. The true loss is likely to depend on the actual parameter value rather than whether parameter value exceeds some precisely defined specification. (Such problem is of particular interest whenever a reliability demonstration testing of highly reliable products is required).

Suppose the parameter λ cannot exceed some specification limit λ_0. Values less than λ_0 mean better that specified quality; values greater than λ_0 mean increasingly worse quality. The smaller the value of λ bellow the specification limit λ_0, the greater the loss to the manufacturer who has exceeded the quality requirement most likely at some additional cost. The greater the value of λ above the specification limit λ_0, the greater the loss to either the manufacturer or the consumer in terms of low product quality.

In this example, a quadratic loss function would seem to be a reasonable model. The greater the distance from the specification value λ_0, the greater the loss. However the loss associated with over-estimation (below specified quality) is unlikely to be the loss associated with under-estimation (higher than specified quality). In most circumstances, the loss associated with higher than specified quality should be lower than the cost of lower than specified quality. However, there may be exceptions depending on how the specification limit was set and the type of product being tested. Therefore, an asymmetric quadratic loss function would be a reasonable choice for this decision problem.

Let m_1 is the manufacturer's cost of exceeding quality specification, m_2 is manufacturer's cost of a rejected sample, m_3 is manufacturer's cost of an accepted sample which fails to meet quality specification and k is consumer's cost of an accepted sample which fails to meet quality specification.

$$\begin{aligned} \ell(\lambda,d_0) &= m_1(\lambda-\lambda_0)^2; & \lambda \le \lambda_0 \\ &= (m_3+k)(\lambda-\lambda_0)^2; & \lambda > \lambda_0 \end{aligned} \tag{12\n)}$$

$$\begin{aligned} \ell(\lambda,d_1) &= m_1(\lambda-\lambda_0)^2 + m_2; & \lambda \le \lambda_0 \\ &= m_2; & \lambda > \lambda_0 \end{aligned} \tag{13}$$

Schematic demonstration of the idea brings Figure 1.

Suppose that on the basis of some test observation, X, the posterior distribution of λ is known. For each observation X the decision function d must take one of two values: d_0, d_1, which means Accept or Reject. The Bayes risk minimizing decision function is the one which minimizes posterior expected loss at each value of X. The expected posterior loss for each of the two possible actions is:

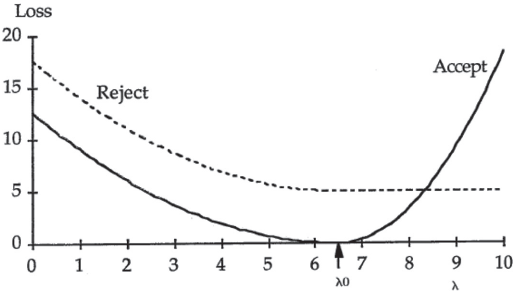


Figure 1. Asymmetric loss function for cost dependent acceptance sampling.

$$\begin{aligned} \mathrm{E}_{\mathrm{h}(\lambda\,|\mathrm{x})}[\ell(\lambda,d_0)] &= \int_{\lambda} \ell(\lambda,d_0)h(\lambda|x)d\lambda = \\ &= m_1 \int_{-\infty}^{\lambda_0} (\lambda-\lambda_0)^2 h(\lambda|x)d\lambda \\ &+ (m_3+k)\int_{\lambda_0}^{+\infty} (\lambda-\lambda_0)^2 h(\lambda|x)d\lambda \end{aligned} \tag{14}$$

$$\begin{aligned} \mathrm{E}_{\mathrm{h}(\lambda\,|\mathrm{x})}[\ell(\lambda,d_1)] &= \int_{\lambda} \ell(\lambda,d_1)h(\lambda|x)d\lambda = \\ &= m_1 \int_{-\infty}^{\lambda_0} (\lambda-\lambda_0)^2 h(\lambda|x)d\lambda + m_2 \end{aligned} \tag{15}$$

The difference between the expected posterior loss of accepting and rejecting the sample is:

$$\begin{aligned} &\mathrm{E}_{\mathrm{h}(\lambda\,|\mathrm{x})}[\ell(\lambda,d_0)] - \mathrm{E}_{\mathrm{h}(\lambda\,|\mathrm{x})}[\ell(\lambda,d_1)] \\ &= (m_3+k)\int_{\lambda_0}^{+\infty} (\lambda-\lambda_0)^2 h(\lambda|x)d\lambda - m_2 \end{aligned} \tag{16}$$

Therefore, the *optimal Bayes Decision Rule (BDR)* is to accept the sample whenever:

$$(m_3+k)\int_{\lambda_0}^{+\infty} (\lambda-\lambda_0)^2 h(\lambda|x)d\lambda \prec m_2 \tag{17}$$

That is, the sample is accepted whenever the expected total cost to both manufacturer and consumer of accepting a sample which fails to meet specifications is less than the manufacturer's cost of discarding any sample, whether or not it meets specifications.

3.3 *Example in case of uniform prior*

If a uniform prior is assumed, the decision rule becomes

$$(m_3+k)p(x)\int_{\lambda_0}^{+\infty}(\lambda-\lambda_0)^2 f(x|\lambda)d\lambda \prec m_2 \tag{18}$$

where

$$p(x)=\left[\int_{\lambda} f(x|\lambda)d\lambda\right]^{-1} \tag{19}$$

Of course the relative magnitudes of the losses, m1, m3, and k as well as the scale factor $p(x)$, are important in determining the optimal decision, but in general the sample will be accepted when the likelihood of x for those parameter values λ > λ0, associated with an unacceptable sample, is small.

4 EXAMPLE FROM INDUSTRY

4.1 *Reliability demonstration testing the failure rate of an exponential distribution*

We give good motivation example from industry (Bris 2002) for the use of the BDR.

Notation

MLE … Maximum Likelihood Estimator

RDT … Reliability Demonstration Testing

λ_0 … requested (by consumer) limit of failure rate at normal condition

λ_2… failure rate in accelerated condition

i … index for condition given by temperature; $i = 1, 2$

λ_1… failure rate tested at given (normal) temperature condition

t_i… total test time in condition i during which r_i failures have occurred

(t_i, r_i) …parameters of acceptance sample plan in condition i

δ^*… posterior risk, Pr $\{\lambda_1 > \lambda_0 \mid \text{passing RDT}\}$

$1 - \delta^*$…consumer's posterior assurance

A …acceleration factor

Consumer asks for following limitation in reliability of delivered Electronic Components (EC): $\lambda_1 \le \lambda_0$. MLE of λ_1 for the time censored data is given by

$$\lambda_1=\frac{r_1}{t_1} \tag{20}$$

However, total time t_1 is usually very large even when $r_1 = 1$ is admitted. Let us perform RDT in accelerated condition:

$$\lambda_2=\frac{r_2}{t_2} \tag{21}$$

but $\lambda_2 = A\ \lambda_1$ and $\lambda_1 \le \lambda_0$, so that the condition for r_2, t_2 is given by

$$\frac{r_2}{t_2}\le A\,\lambda_0 \tag{22}$$

Using the Bayes approach we consider failure rates and acceleration factor to vary randomly according to some prior distributions. To meet the condition (22), possible acceptance sample plans for RDT are presented in the Table 1, given real data: λ_0 = 8.76e-4 /year, A = 13.09.

Conditioned (by t_2, r_2) posterior distribution of λ_2, derived in (Bris 2000), is given as follows:

$$h(\lambda_2|\mathrm{t}_2,r_2)\propto \lambda_2^{r_2}\exp\left(-\mathrm{t}_2\lambda_2\right)\int_0^{\infty}\frac{u^b}{(c+u)^{2b}}\exp\left(-a\lambda_2 u\right)du,$$

where

$a = 7.247e6$

$b = 2.1063$

$c = 0.001435$

(23)

Having the knowledge of posterior, we can optimize not only the variance $Var\{\lambda_2|t_2, r_2\}$ (demonstrated in Figure 2, in units [hour^{-2}]) and mean $E\{\lambda_2|t_2, r_2\}$, moreover we are able to quantify the consumer's posterior assurance $1 - \delta^*$, given by the following relationship:

$$1-\delta^* = \Pr\{\lambda_1 \le \lambda_0 \mid \text{passing RDT}\} \tag{24}$$

as demonstrated in Figure 3.

Using the criterion of small variance of λ_2 at acceptable test time for RDT, sample plan *P2* seems to be optimal. The process of optimization is demonstrated in the Figure 2 which corresponds with the results in Table 1. Having the variance of 7.6 × e − 13/hour^2 (i.e. 0.5832 e-4/year^2) requires minimal total test time t_2 = 1.53 × e6 test hours at maximally 2 allowed failures (r_2 = 2). Calculation of posterior assurance that $\lambda_1 \le \lambda_0$, $1 - \delta^*$, in a test that has passed is 77% (Figure 3).

Above derived optimal Bayes decision rule is to accept the sample whenever:

$$\begin{gathered}(m_3+k)\int_{\lambda_0}^{+\infty}(\lambda_2-\lambda_0)^2 h(\lambda_2|t_2,r_2)d\lambda_2 \prec m_2\\ \Rightarrow \int_{\lambda_0}^{+\infty}(\lambda_2-\lambda_0)^2 h(\lambda_2|t_2,r_2)d\lambda_2 \prec \frac{m_2}{(m_3+k)}\end{gathered} \tag{25}$$

where $h\{\lambda_2 \mid t_2, r_2\}$ is given by (23). Computing the value on the left side of the last equation we can obtain clear limitation for cost in given example of

Table 1. Sample plans for RDT with corresponding characteristics.

	t_2	r_2	$E\{\lambda_2 \mid t_2, r_2\}$	$Var\{\lambda_2/t_2,r_2\}$	$\sqrt{Var\{\lambda_2 / t_2, r_2\}}$
Sample plan	10^6 test hours		e-2/year	e-4/year^2	e-2/year
P_1	2.3	3	1.253	0.4036	0.635
P_2	1.53	2	1.297	0.5832	0.764
P_3	0.76	1	1.402	1.036	1.018
P_4	5.0	6	1.113	0.077	0.277

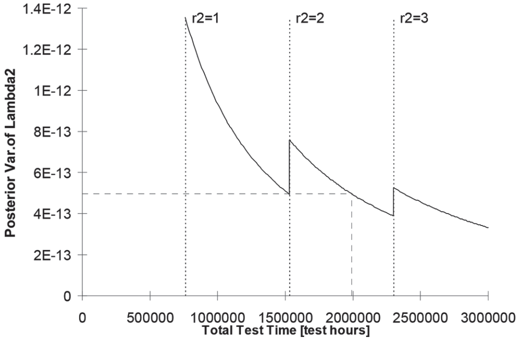


Figure 2. The optimization process for t_2.

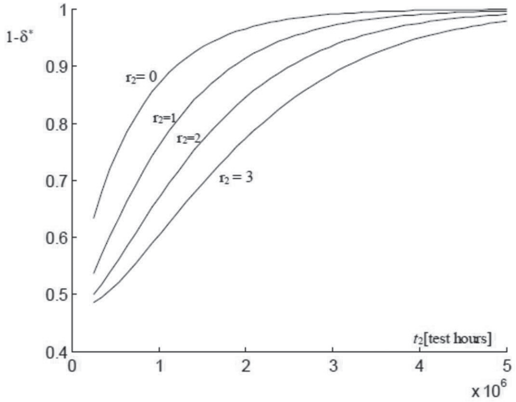


Figure 3. Posterior assurance that $\lambda_1 \leq \lambda_0$ ($\lambda_0 = 8.76\text{e-}4$ / year) for passing tests in dependence on total test time t_2.

RDT. For example, taking into account the sample plan P_2, the value is about 2e-4.

In all practical production situations usually is valid the following relation:

$$(m_3 + k) \succ\succ m_2 \\ \Rightarrow 0 \prec \frac{m_2}{(m_3 + k)} \prec\prec 1 \tag{26}$$

Of course, the later a newly developed product (with a covert fail) is taken off the market, the more expensive it is (both from producer and also consumer point of view). We can remember many examples from previous time from industry concerning the problem (car industry, producers of electronic devices, etc.). The findings of the paper just answer the question, which relation referring the cost should be observed in acceptance sampling plans to make an optimal decision. We usually know that the fraction in the relationship (26) is far less than 1, but we do not know in practice how close to zero it should be. In our example, the optimal decision is made, whenever:

$$2\text{e-}4 \prec \frac{m_2}{(m_3 + k)} \prec\prec 1 \\ \Rightarrow (m_3 + k)2\text{e-}4 \prec m_2$$

5 CONCLUSIONS

Bayes framework for hypothesis testing was developed and proposed to use in acceptance sampling plans. Asymmetric cost oriented loss functions are proposed to allow optimal decision making in acceptance sampling plans.

Optimal Bayes decision rule considering the loss associated with both higher and lower than specified quality was derived.

Possibilities of use the rule in practice are also demonstrated on given example from electronic industry.

ACKNOWLEDGEMENT

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project “IT4 Innovations excellence in science – LQ1602”.

REFERENCES

Berger, J.O. 1993. *Statistical Decision Theory and Bayesian Analysis,* 3rd Edition, Springer, New York.

Briš, R. 2000. Bayes approach in RDT using accelerated and long-term life data; *Reliability Engineering and System Safety* 2000 **67**: 9–16, ELSEVIER.

Briš, R. 2002. Using Posterior Information About the Distribution of Failure Time Parameter in Acceptance Sampling Schemes; *λμ 13 – ESREL 2002 European Conference on System Dependability and Safety*, Lyon, France, March 18–21, 2002, *Proceedings of the Conferences II.*: 627–630.

Dodge, H.F. & Roming, H.G. 1959. *Sampling Inspection Tables*, 2nd Edition, John Wiley & Sons, New York.

Godfrey, A.B. & Mundel, A.B. (1984). Guide for selection of an acceptance sampling plan, *Journal of Quality Technology* 16, 50–55.

Norstrom J.G. 1996. The Use of Precautionary Loss Functions in Risk Analysis; *IEEE Transactions on Reliability*, 1996, Vol. **45**, No. 3, 1996 September.

O'Hagan, Anthony. 1994. *Kendall's Advanced Theory of Statistics*, Volume 2B—Bayesian Inference; ISBN 0340529229, 1994 University Press Cambridge.

Weerahandi, S. 1995. *Exact Statistical Methods for Data Analysis*; Springer Verlag 1995, ISBN 0–387–94360–9.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

High-dimensional simulation experiments with particle filter and ensemble Kalman filter

P. Bui Quang & V.-D. Tran
Hoa Sen University, Ho Chi Minh City, Vietnam

ABSTRACT: Particle filter and ensemble Kalman filter are two Bayesian filtering algorithms adapted to nonlinear state–space models. The problem of nonlinear Bayesian filtering is challenging when the state dimension is high, since approximation methods tend to degrade as dimension increases. We experimentally investigate the performance of particle filter and ensemble Kalman filter as the state dimension increases. We run simulations with two different state dynamics: a simple linear dynamics, and the Lorenz–96 nonlinear dynamics. In our results, the approximation error of both algorithms grows at a linear rate when the dimension increases. This linear degradation appears to be much slower for ensemble Kalman filter than for particle filter.

1 INTRODUCTION

Bayesian filtering consists in computing the conditional distribution of the unobserved state of a dynamical system w.r.t. available observations (the posterior distribution), in order to estimate the state. When the state dynamics or the observation model are linear with additive Gaussian noise, the Kalman filter yields optimal state estimation. Otherwise, approximation methods must be used. The Particle Filter (PF) and the Ensemble Kalman Filter (EnKF) are two algorithms that compute an approximation of the posterior in the form of a random sample of points. These points are weighted in PF, whereas in EnKF they are not attached to a weight.

Approximation problems are known to be more difficult when the state dimension is large. This phenomenon, referred to as "curse of dimensionality" (Bengtsson, Bickel, & Li 2008, Daum & Huang 2003), has been observed in particle filtering. In (Bengtsson, Bickel, & Li 2008, Bickel, Li, & Bengtsson 2008, Bui Quang, Musso, & Le Gland 2010, Snyder, Bengtsson, Bickel, & Anderson 2008), the authors argue that the particle sample size must increase exponentially with the dimension to avoid the curse of dimensionality. A simple linear target tracking problem is analyzed in (Bui Quang, Musso, & Le Gland 2010), and a theoretical study is made in (Bengtsson, Bickel, & Li 2008, Bickel, Li, & Bengtsson 2008, Snyder, Bengtsson, Bickel, & Anderson 2008) under particular model assumptions. The problem occurring in PF when the dimension is large is that the particle approximation tends to collapse to a single Dirac measure, i.e. all the points but one have a zero weight and only one point has a nonzero weight. In this case, the particle approximation of the posterior, and therefore the state estimation, is very poor. A strategy to avoid this problem has recently been proposed in (Beskos, Crisan, & Jasra 2014). In EnKF however, the sample points are not weighted, so that this weight degeneracy phenomenon cannot occur (Le Gland, Monbet, & Tran 2011). EnKF is a popular algorithm for models in which the state dimension is very large, such as models describing geophysical systems (Evensen 2009, van Leeuwen 2009).

In this paper, we lead two simulation experiments where we compare the performance of PF and EnKF when the state dimension increases. Algorithm performance is assessed in terms of the mean squared error between the approximated posterior mean and the true posterior mean. Numerical results clearly indicate that the error increases linearly with the state dimension for both algorithms. This increase is slower for EnKF than for PF.

The outline of the paper is as follows. In Section 2, we present the problem of Bayesian filtering in state–space models. The PF and EnKF algorithms, and the type of models to which they apply, are then described in Section 3. In Section 4, we describe how we assess the performance of algorithms. We define the algorithm error, which is related to the algorithm approximation, and the model error, which is related to the intrinsic model uncertainty and is independent of the algorithm. Simulation experiments are lead in Section 5 and conclusion is drawn in Section 6.

Throughout the paper, we use the following notations: $a_{1:n} = (a_1,\ldots,a_n)$, I_n is the $n \times n$ identity matrix, $0_{n,p}$ is the $n \times p$ zero matrix.

The (MATLAB) computer code used for simulation experiments is available from the first author.

2 STATE–SPACE MODELS AND BAYESIAN FILTERING

A state–space model is a time series model describing the observation of a dynamical system. The system is characterized by a hidden (unobserved) state variable taking values in $\mathbb{R}^m$. State–space models are hidden Markov models where the state variable is continuous (Cappé, Moulines, & Ryden 2005).

The state dynamics is a Markov process with transition kernel

$$X_k \mid X_{k-1} = x_{k-1} \sim q_k(x_{k-1}, \cdot), \tag{1}$$

with the initial distribution $X_0 \sim q_0$ (in terms of densities). The observation model is

$$Y_k \mid X_k = x_k \sim g_k(x_k), \tag{2}$$

where g_k is the likelihood, i.e. $g_k(x_k)$ is the density of the conditional probability $\mathbb{P}[Y_k \in dy_k \mid X_k = x_k]$. The observation sequence $\{Y_k\}_{k \geq 0}$ verifies

$$\mathbb{P}[Y_k \in dy_k \mid X_{0:n}] = \mathbb{P}[Y_k \in dy_k \mid X_k].$$

The problem of Bayesian filtering consists in computing sequentially the conditional distribution of the hidden state X_k w.r.t. to past observations $Y_{0:k-1}$, called the predictor, and the conditional distribution of X_k w.r.t. to all available observations $Y_{0:k}$, called the posterior, at each time step $k \geq 0$. The predictor and the posterior, respectively denoted $p_{k|k-1}$ and p_k, obey to the recursive relation

$$p_{k|k-1}(x) = \int_{\mathbb{R}^m} p_{k-1}(z) q_k(z, x) dz, \tag{3}$$

$$p_k(x) = \frac{g_k(x) p_{k|k-1}(x)}{\int_{\mathbb{R}^m} g_k(x) p_{k|k-1}(x) dx}, \tag{4}$$

with the convention $p_{0|-1} = q_0$ (in terms of densities). Equation (3) is the prediction step and Equation (4) is the update step. Note that Equation (4) is the Bayes formula, where the prior is the predictor.

3 PARTICLE FILTER AND ENSEMBLE KALMAN FILTER

We present now the two nonlinear Bayesian filtering techniques we consider in the paper: particle filter and ensemble Kalman filter. Comprehensive presentations of these algorithms can be found, for example, in (Arulampalam, Maskell, Gordon, & Clapp 2002, Cappé, Moulines, & Ryden 2005) for particle filter and in (Evensen 2003, Evensen 2009) for ensemble Kalman filter.

3.1 *Particle Filter*

Particle Filters (PF), also known as sequential Monte Carlo methods, are Bayesian filtering algorithms for general state–space models, as described in Equations (1)–(2). They are based on the principle of importance sampling. They recursively approximate the predictor and the posterior by a weighted sample of points, called "particles", in the state space, i.e. a weighted sum of Dirac measures $\Sigma_{i=1}^{N} w_k^i \delta_{\xi_k^i}$, where ξ_k^i denote the particle position and w_k^i denote the particle weight.

A common problem with particle filtering is weight degeneracy, which occurs when most of the particles have a weight that is numerically zero, whereas a few particles have a nonzero weight. Weight degeneracy yields poor particle approximation, and it is particularly severe when the state dimension is high (Bengtsson, Bickel, & Li 2008, Bickel, Li, & Bengtsson 2008, Bui Quang, Musso, & Le Gland 2010, Daum & Huang 2003, Snyder, Bengtsson, Bickel, & Anderson 2008, van Leeuwen 2009). A common strategy to (partially) avoid it is to resample the particles according to their weights, i.e. the probability that a particle is drawn equals its weight. This multinomial resampling tends to discard particles with a low weight and to duplicate particles with a large weight. Weight degeneracy can be quantified by the effective sample size $N_{\text{eff}} = 1 / \Sigma_{i=1}^{N} (w_k^i)$. N_{eff} verifies $0 \leq N_{\text{eff}} \leq N$, and it is small is case of weight degeneracy.

There exists many versions of particle filters. The first practical implementation of PF has been proposed in (Gordon, Salmond, & Smith 1993). We give in Algorithm 1 below the implementation of the most classical particle filter, called SIR (for sequential importance resampling), where particles are propagated according to the Markov dynamics explicitly given by Equation (1), and resampled if the effective sample size is too small.

3.2 *Ensemble Kalman Filter*

The Ensemble Kalman Filter (EnKF) is applicable to state–space models where the observation model is linear with an additive noise, i.e.

$$Y_k = H_k X_k + V_k, \tag{5}$$

where H_k is a $d \times m$ matrix and $V_k \sim p_k^V$.

In EnKF, the predictor and the posterior are approximated by a sample of points, like in PF. This sample is referred to as "ensemble" here.

Algorithm 1 SIR particle filter (with prior proposal)

```
k = 0 {initialization}
for $i = 1 \cdots N$ do
  sample $\xi_0^i \sim q_0$
  compute weight $w_0^i \propto g_0(\xi_0^i)$
end for
normalize weights
loop
  $k \leftarrow k+1$ {time iteration}
  compute $N_{\text{eff}} = \frac{1}{\sum_{i=1}^N (w_k^i)^2}$
  if $N_{\text{eff}} \geq N_{\text{th}}$ then
    for $i = 1 \cdots N$ do
      sample $\xi_k^i \sim q_k(\xi_{k-1}^i, \cdot)$
      compute weight $w_k^i \propto w_{k-1}^i g_k(\xi_k^i)$
    end for
  else if $N_{\text{eff}} < N_{\text{th}}$ then
    for $i = 1 \cdots N$ do
      sample $\xi_{k-1}'^i \sim \sum_{i=1}^N w_{k-1}^i \delta_{\xi_{k-1}^i}$
      sample $\xi_k^i \sim q_k(\xi_{k-1}'^i, \cdot)$
      compute weight $w_k^i \propto g_k(\xi_k^i)$
    end for
  end if
  normalize weights
end loop
```

Algorithm 2 Ensemble Kalman filter

```
$k = 0$ {initialization}
for $i = 1 \cdots N$ do
  sample $\xi_{0|-1}^i \sim q_0$
  sample $V^i \sim p_0^V$
  compute $Y^i = H_0 \xi_{0|-1}^i + V^i$
end for
compute $R^N = \frac{1}{N}\sum_{i=1}^N (V^i - \frac{1}{N}\sum_{i=1}^N V^i)(V^i - \frac{1}{N}\sum_{i=1}^N V^i)^T$
compute $K_0^N = P_{0|-1}^N H_0^T (H_0 P_{0|-1}^N H_0^T + R^N)^{-1}$
for $i = 1 \cdots N$ do
  compute $\xi_0^i = \xi_{0|-1}^i + K_0^N (Y_0 - Y^i)$
end for
loop
  $k \leftarrow k+1$ {time iteration}
  for $i = 1 \cdots N$ do
    sample $\xi_{k|k-1}^i \sim q_k(\xi_{k-1}^i, \cdot)$
    sample $V^i \sim p_k^V$
    compute $Y^i = H_k \xi_{k|k-1}^i + V^i$
  end for
  compute $R^N = \frac{1}{N}\sum_{i=1}^N (V^i - \frac{1}{N}\sum_{i=1}^N V^i)(V^i - \frac{1}{N}\sum_{i=1}^N V^i)^T$
  compute $K_k^N = P_{k|k-1}^N H_k^T (H_k P_{k|k-1}^N H_k^T + R^N)^{-1}$
  for $i = 1 \cdots N$ do
    compute $\xi_k^i = \xi_{k|k-1}^i + K_k^N (Y_k - Y^i)$
  end for
end loop
```

Ensemble members are not weighted as in PF. Instead, each point is moved according to an affine transformation that mimics the update step of the Kalman filter. This transformation involves a gain matrix which depends on the predictor covariance matrix. The predictor covariance matrix can be approximated using ensemble members, as

$$P_{k|k-1}^N = \frac{1}{N}\sum_{i=1}^N (\xi_{k|k-1}^i - m_{k|k-1}^N)(\xi_{k|k-1}^i - m_{k|k-1}^N)^T \tag{6}$$

with $m_{k|k-1}^N = \frac{1}{N}\sum_{i=1}^N \xi_{k|k-1}^i$, where $\{\xi_{k|k-1}^i\}_i$ is a sample approximation of the predictor. The gain matrix is then

$$K_k^N = P_{k|k-1}^N H_k^T (H_k P_{k|k-1}^N H_k^T + R^N)^{-1}, \tag{7}$$

where R^N is the sample covariance matrix of the i.i.d. sample $V^1,\dots,V^N$ from the noise distribution p_k^V.

EnKF is displayed in Algorithm 2 below. This algorithm has been developed for data assimilation problems arising in geophysics (especially meteorology and oceanography). Seminal papers on EnKF are (Burgers, van Leeuwen, & Evensen 1998, Evensen 1994).

In practice, the sample covariance matrix $P_{k|k-1}^N$ defined in Equation (6) does not need to be computed nor stored. Indeed, the gain formula in Equation (7) shows that only the matrix product $P_{k|k-1}^N H_k^T$ is required. Firstly, $P_{k|k-1}^N H_k^T$ has size $m \times d$, whereas $P_{k|k-1}^N$ has size $m \times m$. Secondly,

$$\begin{aligned} P_{k|k-1}^N H_k^T &= \frac{1}{N}\sum_{i=1}^N (\xi_{k|k-1}^i - m_{k|k-1}^N)(\xi_{k|k-1}^i - m_{k|k-1}^N)^T H_k^T \\ &= \frac{1}{N}\sum_{i=1}^N (\xi_{k|k-1}^i - m_{k|k-1}^N)\left[H_k(\xi_{k|k-1}^i - m_{k|k-1}^N)\right]^T \\ &= \frac{1}{N}\sum_{i=1}^N (\xi_{k|k-1}^i - m_{k|k-1}^N) h_k^i, \end{aligned}$$

where $h_k^i = \left[H_k(\xi_{k|k-1}^i - m_{k|k-1}^N)\right]^T$. To get $P_{k|k-1}^N H^T$, the following operations (scalar addition or multiplication) are performed for each ensemble member:

- d scalar products of m–dimensional vectors to compute h_k^i, requiring $O(dm)$ operations,
- one (matrix) multiplication of a m–dimensional column–vector with a d–dimensional row–vector, to compute $(\xi_{k|k-1}^i - m_{k|k-1}^N)h_k^i$, requiring $O(dm)$ operations, yielding $O(Ndm)$

operations in total. On the other hand, to get $P^N_{k|k-1}$, a m–dimensional column–vector is multiplied with a m–dimensional row–vector to compute $(\xi^i_{k|k-1} - m^N_{k|k-1})(\xi^i_{k|k-1} - m^N_{k|k-1})^T$, requiring $O(m^2)$ operations for each ensemble member, yielding $O(Nm^2)$ operations in total. Thus, when $d \ll m$, as it is the case in many practical geophysical models (van Leeuwen 2009), it is less computationally demanding to store and compute $P^N_{k|k-1}H^T_k$ than $P^N_{k|k-1}$.

3.3 *Linear formulation of observation model*

State–space models with linear observation are a rather general family of models. In particular, state–space models of the form

$$X_k = F_k(X_{k-1}) + W_k, \tag{8}$$

$$Y_k = H_k(X_k) + V_k, \tag{9}$$

where F_k and H_k are nonlinear mappings, and W_k and V_k are noise, can be casted in this family, and therefore can be handled by EnKF.

Consider the state (column) vector augmented by the observation vector $X'_k = \begin{pmatrix} X^T_k & Y^T_k \end{pmatrix}^T$, and the dynamics noise (column) vector augmented by the observation noise vector $U_k = \begin{pmatrix} W^T_k & V^T_k \end{pmatrix}^T$, taking values in $\mathbb{R}^{m+d}$. Let $E_m = \begin{pmatrix} I_m & 0_{m,d} \end{pmatrix}$ and $E_d = \begin{pmatrix} 0_{d,m} & I_d \end{pmatrix}$ be matrices with respective size $m \times (m+d)$ and $d \times (m+d)$. Then, the state dynamics

$$X'_k = F'_k(X'_{k-1}, U_k), \tag{10}$$

where

$$F'_k(x,u) = \begin{pmatrix} F_k(E_m x) + E_m u \\ H_k(F_k(E_m x) + E_m u) + E_d u \end{pmatrix},$$

is a Markov process. Besides, the observation model

$$Y_k = H' X'_k, \tag{11}$$

where $H' = E_d$, is linear.

We can readily use EnKF to perform Bayesian filtering in the state–space model defined by Equations (10) and (11), as the observation model is linear. Note that there is no observation noise here, i.e. the observation noise variance is zero. EnKF can handle such a model, since $\operatorname{rank}(H') = d$ (so that the matrix $H' P^N_{k|k-1} H'^T + 0_{d,d}$ is invertible and the gain matrix defined in Equation (7) can be computed). Let p'_k be the conditional density of X'_k w.r.t. $Y_{0:k}$. Then p_k, the conditional density of X_k w.r.t. $Y_{0:k}$ (see Section 2), is obtained by marginalizing p'_k as

$$\begin{aligned} p_k(x) &\\ &= p_k(x_1, \dots, x_m) \\ &= \int_{\mathbb{R}^d} p'_k(x_1, \dots, x_m, x_{m+1}, \dots, x_{m+d}) \\ &\times dx_{m+1} \cdots dx_{m+d}. \end{aligned}$$

In terms of sample approximation, this marginalization consists in removing the d last vector components of the ensemble members, i.e. $\xi^i_k = \begin{pmatrix} \varsigma^i_{k,1} \cdots \varsigma^i_{k,m} \end{pmatrix}^T$, where $\varsigma^1_k, \dots, \varsigma^N_k$ is the sample approximating p'_k and $\varsigma^i_k = \begin{pmatrix} \varsigma^i_{k,1} \cdots \varsigma^i_{k,m+d} \end{pmatrix}^T$. The predictor approximation (conditional density of X_k w.r.t. $Y_{0:k-1}$) is obtained similarly by marginalizing the conditional density of X'_k w.r.t. $Y_{0:k-1}$.

PF cannot be applied when the observation noise is zero. Indeed, the likelihood associated with the observation model (11) takes positive values over the linear subspace $\{x \in \mathbb{R}^m : Y_k = H'x\}$ only. In this case, weighting yields (almost surely) a zero weight for each particle and the algorithm degenerates at the first time iteration. PF however can readily be used when the state–space model is in the form of Equations (8)–(9), there is no need to put the observation model in a linear form.

4 PERFORMANCE ASSESSMENT OF BAYESIAN FILTERS

The problem of Bayesian filtering consists in computing the posterior (and the predictor) at each time iteration. From the posterior, one can compute Bayesian estimators of the hidden state. Two classical Bayesian estimators are: the Maximum A Posteriori (MAP) $\operatorname*{argmax}_{x \in \mathbb{R}^m} \{p_k(x)\}$, and the posterior mean $\mathbb{E}[X_k \mid Y_{0:k}]$. The MAP is not readily available in PF or EnKF, because the posterior approximation is in the form of a sample, not a smooth density. The posterior mean however can straightforwardly be approximated by averaging the particles, i.e. $\mathbb{E}[X_k \mid Y_{0:k}] \approx \sum_{i=1}^N w^i_k \xi^i_k$ in PF and $\mathbb{E}[X_k \mid Y_{0:k}] \approx \frac{1}{N} \sum_{i=1}^N \xi^i_k$ in EnKF. Besides, the posterior mean has the minimum mean squared error among unbiased estimators, i.e. for all estimator of the hidden state $\hat{X}_k$ such that $\mathbb{E}[\hat{X}_k] = \mathbb{E}[X_k]$ we have that

$$\mathbb{E}\left[|\hat{X}_k - X_k|^2\right] \geq \mathbb{E}\left[|X_k - \mathbb{E}[X_k \mid Y_{0:k}]|^2\right]$$

(where $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^m$). Hence, in this paper we use the posterior mean as state estimator in PF and EnKF.

A natural way to assess the performance of filtering algorithms is to compute the difference between the estimated state $\hat{X}_k$ and the “true” state X^*_k. The true state is defined as the state value from which the observations are generated, i.e.it verifies

$Y_k \sim g_k(X_k^*)$. In a simulation framework, the true state is known since the observations are generated in the experiment. The difference between $\hat{X}_k$ and X_k^* can be decomposed in two terms, as

$$\hat{X}_k - X_k^* = (\hat{X}_k - \mathbb{E}[X_k \mid Y_{0:k}]) - (X_k^* - \mathbb{E}[X_k \mid Y_{0:k}]),$$

which are related to different sources of error. The first term $\hat{X}_k - \mathbb{E}[X_k \mid Y_{0:k}]$ is related to the algorithm approximation of the posterior mean, hence we refer to it as *algorithm error*. The second term $X_k^* - \mathbb{E}[X_k \mid Y_{0:k}]$ is related to the model and to the choice of the posterior mean as state estimator, hence we refer to it as *model error*.

In the simulations presented in this paper, we analyze separately these two error terms by computing their squared mean at final time step $k = n$. To do so, we generate R i.i.d. true state sequences $X_{0:n}^{*1}, \ldots, X_{0:n}^{*R}$, from which we generate R observation trajectories $Y_{0:n}^1, \ldots, Y_{0:n}^R$, where $Y_k^r \sim g_k(X_k^{*r})$ for all $r \in \{1,\ldots,R\}$ and $k \in \{0,\ldots,n\},$. The (total) Mean Squared Error (MSE), the algorithm MSE, and the model MSE, are then respectively approximated as

$$\begin{aligned}\text{MSE} &= \mathbb{E}\left[\mid \hat{X}_n - X_n^* \mid^2\right] \\ &\approx \frac{1}{R}\sum_{r=1}^{R} |\hat{X}_n^r - X_n^{*r}|^2,\end{aligned}$$

$$\begin{aligned}\text{algorithm MSE} &= \mathbb{E}\left[\mid \hat{X}_n - \mathbb{E}[X_n \mid Y_{0:n}] \mid^2\right] \\ &\approx \frac{1}{R}\sum_{r=1}^{R} |\hat{X}_n^r - \mathbb{E}[X_n | Y_{0:n}^r]|^2,\end{aligned}$$

$$\begin{aligned}\text{model MSE} &= \mathbb{E}\left[\mid X_n^* - \mathbb{E}[X_n \mid Y_{0:n}] \mid^2\right] \\ &\approx \frac{1}{R}\sum_{r=1}^{R} |X_n^{*r} - \mathbb{E}[X_n | Y_{0:n}^r]|^2,\end{aligned}$$

where $\hat{X}_n^r$ is the algorithm approximation of the posterior mean using observations $Y_{0:n}^r$.

5 HIGH-DIMENSIONAL SIMULATIONS

5.1 *Linear dynamics*

We firstly consider a very simple state–space model where the state dynamics and the observation model are linear with additive Gaussian white noise,

$$\begin{aligned} X_k &= X_{k-1} + W_k, \\ Y_k &= H X_k + V_k, \end{aligned}$$

where $W_k \sim \mathcal{N}(0,Q)$ and $V_k \sim \mathcal{N}(0,\rho^2)$, and where $H = \begin{pmatrix}1 & 0 & \cdots & 0\end{pmatrix}$ is a $1 \times m$ matrix. The initial state X_0 follows the distribution $\mathcal{N}(0,Q_0)$.

The true state sequence is simulated thanks to the state dynamics, i.e. $X_0^{*r} \sim \mathcal{N}(0,Q_0)$ and $X_k^{*r} \mid X_{k-1}^{*r} = x_{k-1}^{*r} \sim \mathcal{N}(x_{k-1}^{*r},Q)$ for all $r \in \{1,\ldots,R\}$ and $k \in \{1,\ldots,n\}$. Consequently, the model MSE $\mathbb{E}\left[\mid X_n^* - \mathbb{E}[X_n \mid Y_{0:n}] \mid^2\right]$ is equal to $\mathbb{E}\left[\mid X_n - \mathbb{E}[X_n \mid Y_{0:n}] \mid^2\right]$, which is the trace of the posterior covariance matrix.

In a linear Gaussian model, the exact posterior mean and posterior covariance matrix are given by the Kalman filter at each time step, so that the model MSE can easily be computed, without approximation.

The model parameters are set as follows: $Q_0 = I_m$, $Q = 10^{-6} I_m$, $\rho = 1$. The number of time iterations is set to $n = 100$.

5.2 *Lorenz–96 nonlinear dynamics*

We secondly consider a state–space model where the state dynamics is the highly nonlinear Lorenz–96 dynamics and the observation model is linear. The Lorenz–96 dynamics is a standard dynamical model in geophysical data assimilation (Hoteit, Pham, Triantafyllou, & Korres 2008, Nakano, Ueno, & Higuchi 2007).

The Lorenz–96 dynamics is a deterministic time–continuous multidimensional dynamics. It is defined by the nonlinear ordinary differential equation

$$\dot{x}_{t,j} = (x_{t,j+1} - x_{t,j-2}) x_{t,j-1} - x_{t,i} + f \tag{12}$$

for $j \in \{1,\ldots,m\}$, where $x_t = \left(x_{t,1} \cdots x_{t,m}\right)^T \in \mathbb{R}^m$ and f is a scalar constant parameter. By convention, $x_{t,-1} = x_{t,m-1}, x_{t,0} = x_{t,m}, x_{t,m+1} = x_{t,1}$. Equation (12) is discretized with the (fourth order) Runge–Kutta method to get the discrete–time dynamics $x_{t+h} = F(x_t)$, where h is the discretization step. The state dynamics in the state–space model is then

$$X_k = F(X_{k-1}) + W_k,$$

where $W_k \sim \mathcal{N}(0,Q)$ and $X_0 \sim \mathcal{N}(m_0,Q_0)$.

The observation is the first component of vector X_k disrupted by an additive Gaussian noise with variance ρ^2, i.e. the observation model is the same as in Section 5.1 above.

The true state sequence is generated thanks to the discretized Lorenz–96 dynamics, without dynamics noise, with a fixed initial condition x_0, i.e. $X_0^{*r} = x_0$ and $X_k^{*r} = F(X_{k-1}^{*r})$ for all $r \in \{1,\ldots,R\}$ and $k \in \{1,\ldots,n$. Thus, the true state sequence is deterministic here and needs to be generated only once.

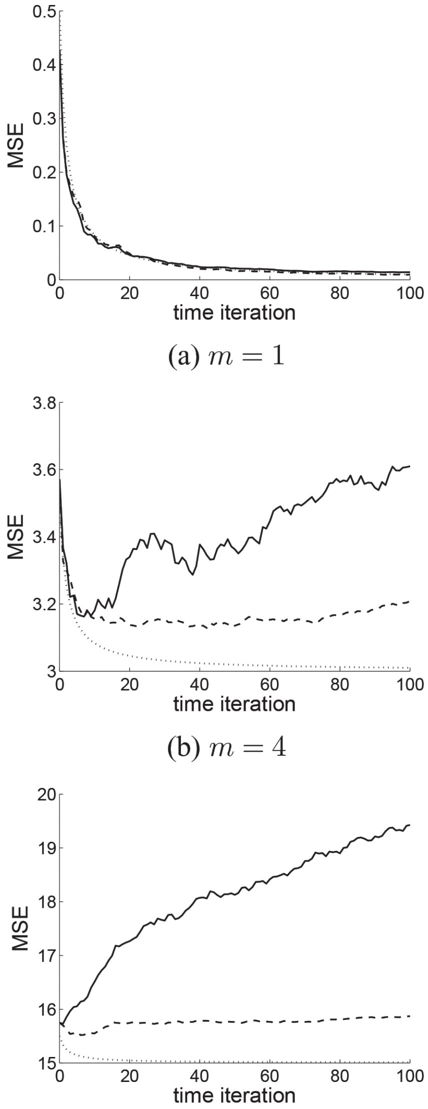
Note that, unlike in the linear Gaussian model in Section 5.1, the exact posterior mean cannot be computed here. We must therefore approximate it accurately to get a reference value to compute the MSEs.

The model parameters are set as follows: $m_0 = (0 \cdots 0)^T$, $Q_0 = 64 I_m$, $Q = 0.25 I_m$, $\rho = 1$, $f = 8$, $h = 0.005$. The number of time iterations is set to $n = 2000$.

5.3 *Numerical results*

We run PF (Algorithm 1) and EnKF (Algorithm 2) to perform Bayesian filtering in the two models described above. In both algorithms, we set the sample size to $N = 100$. In PF, the threshold for multinomial resampling is set to $N_{\text{th}=2/3N}$. The number of simulation runs is $R = 100$.

Results for the linear dynamics model from Section 5.1 are presented in Figures 1 and 2. Figure 1 displays the time evolution of the total MSE for PF, EnKF, and (optimal) Kalman filter, when the state dimension is $m = 1$, 4, and 16. The performance of PF and EnKF reaches optimality when m = 1, but it diverges from optimality when m increases. This divergence is more severe for PF than for EnKF. Figure 2 displays the evolution of the model MSE and the algorithm MSE, for PF and EnKF, at final time step ($k = n$), when the state dimension increases. The model MSE and the algorithm MSE increase at a linear rate with dimension. The increase of algorithm MSE is faster for PF than for EnKF. Figure 2 illustrates that PF is less robust to high dimensionality than EnKF for this linear model.

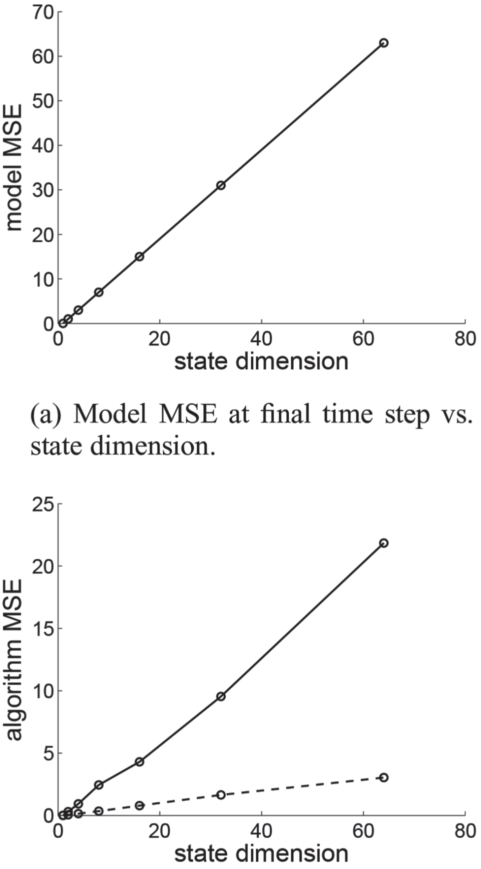


(a) $m = 1$

(b) $m = 4$

(c) $m = 16$

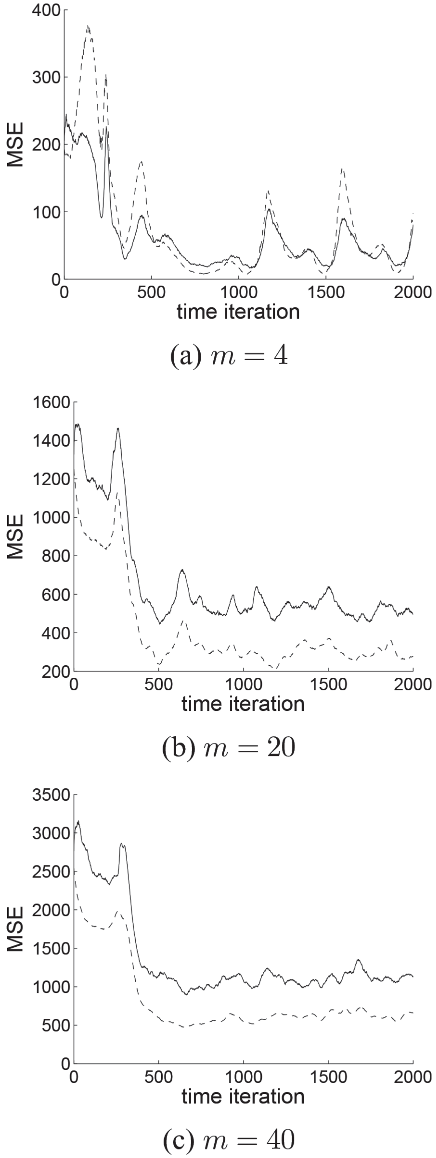
Figure 1. Total MSE vs. time for PF (solid line), EnKF (dashed line) and Kalman filter (dotted line), for different state dimensions (linear dynamics model).



(a) Model MSE at final time step vs. state dimension.

(b) Algorithm MSE at final time step vs. state dimension for PF (solid line) and EnKF (dashed line).

Figure 2. Model MSE and algorithm MSE for PF and EnKF when the state dimension increases (linear dynamics model).

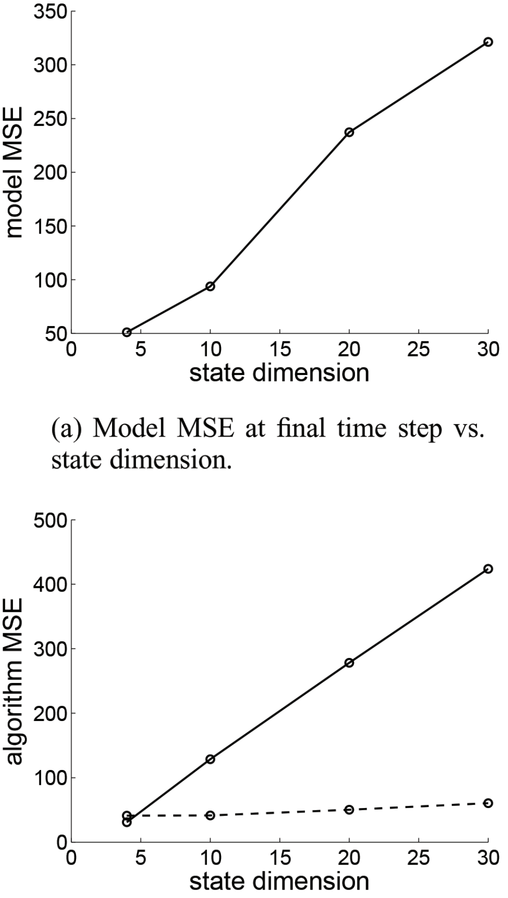


(a) $m = 4$

(b) $m = 20$

(c) $m = 40$

Figure 3. Total MSE vs. time for PF (solid line) and EnKF (dashed line), for different state dimensions (Lorenz–96 nonlinear dynamics model).



(a) Model MSE at final time step vs. state dimension.

(b) Algorithm MSE at final time step vs. state dimension for PF (solid line) and EnKF (dashed line).

Figure 4. Model MSE and algorithm MSE for PF and EnKF when the state dimension increases (Lorenz–96 nonlinear dynamics model).

Results for the nonlinear Lorenz–96 dynamics model from Section 5.2 are presented in Figures 3 and 4. Figure 3 displays the time evolution of the total MSE for PF and EnKF, when the state dimension is $m = 4$, 20, and 40. Figure 4 displays the evolution of the model MSE and the algorithm MSE, for the two filters, at final time step, when the state dimension increases. The same observations than for the linear model can be made here: PF is less robust to high dimension than EnKF, although the algorithm MSE of both algorithms increases at a linear rate.

To get the results presented in Figures 3 and 4, the reference approximation of the posterior mean (required to compute the model and algorithm MSEs) is computed thanks to a particle filter with a large number of particles ($N = 10^5$). We use the optimal particle filter (for this type of model), described in (Le Gland, Monbet, & Tran 2011, Section 6), that differs from the SIR implementation presented in Algorithm 1 in Section 3. When the sample size N is large, we preferably use PF because EnKF has been proven asymptotically biased, i.e. the ensemble members distribution does not converge to the true posterior distribution as $N \to \infty$ (Le Gland, Monbet, & Tran 2011).

6 CONCLUSION

High dimensional nonlinear Bayesian filtering is a difficult approximation problem. In this paper, we study how the performance of two popular nonlinear Bayesian filters, PF and EnKF, is degraded when the state dimension increases.

Regarding PF, several authors argue that, when the dimension increases, the particle sample size must grow at an exponential rate to maintain the approximation quality constant (Bengtsson, Bickel, & Li 2008, Bickel, Li, & Bengtsson 2008, Bui Quang, Musso, & Le Gland 2010, Snyder, Bengtsson, Bickel, & Anderson 2008). EnKF, on the other hand, is widely applied to data assimilation problems in geophysics (Evensen 2009). The models involved in such problems have often a very large dimension (van Leeuwen 2009).

In this paper, we lead simulation experiments to quantify the degradation of PF and EnKF as the state dimension increases. We consider two models with two different state dynamics: a simple linear dynamics, and the nonlinear Lorenz–96 dynamics. The observation model is linear in the two models. We assess the performance of PF and EnKF in terms of the algorithm MSE (the MSE between the approximated posterior mean and the true posterior mean) and the model MSE (the MSE between the true posterior mean and the true state value).

In our simulations, it appears that the algorithm MSE of both algorithms increases linearly with the state dimension. In PF, the algorithm MSE is proportional to $1/N$ (Cappé, Moulines, & Ryden 2005), so that it can be maintained constant if the number of particles grows linearly with the dimension. This empirical result differs from previous results in the literature, showing the need for further analysis to describe the phenomenon. Besides, in our simulations, the algorithm MSE of EnKF increases at a (linear) rate much slower than that of PF. This justifies that EnKF is preferable to PF in high dimensional models.

REFERENCES

Arulampalam, M., S. Maskell, N. Gordon, & T. Clapp (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing 50*.

Bengtsson, T., P. Bickel, & B. Li (2008). Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In *Probability and statistics: Essays in honor of David A. Freedman*.

Beskos, A., D. Crisan, & A. Jasra (2014). On the stability of sequential monte carlo methods in high dimensions. *The Annals of Applied Probability 24*.

Bickel, P., B. Li, & T. Bengtsson (2008). Sharp failure rates for the bootstrap particle filter in high dimensions. In *Pushing the limits of contemporary statistics: Contributions in honor of Jayanta K. Ghosh*.

Bui Quang, P., C.Musso, & F. Le Gland (2010). An insight into the issue of dimensionality in particle filtering. In *Proceedings of 13th International Conference on Information Fusion*, Edinburgh.

Burgers, G., P. van Leeuwen, & G. Evensen (1998). Analysis scheme in the ensemble kalman filter. *Monthly Weather Review 126*.

Cappé, O., E. Moulines, & T. Ryden (2005). *Inference in hidden Markov models*. New York: Springer.

Daum, F. & J. Huang (2003). Curse of dimensionality and particle filters. In *Proceedings of IEEE Aerospace Conference*, Big Sky, MT.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasigeostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research 99*.

Evensen, G. (2003). Ensemble kalman filter: theoretical formulation and pratical implementations. *Ocean Dynamics 53*.

Evensen, G. (2009). *Data assimilation, the ensemble Kalman filter*. Second edition. Berlin: Springer.

Gordon, N., D. Salmond, & A. Smith (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings–F 140*.

Hoteit, I., D.T. Pham, G. Triantafyllou, & G. Korres (2008). Particle kalman filtering for data assimilation in meteorology and oceanography. In *Proceedings of 3rd WCRP International Conference on Reanalysis*, Tokyo.

Le Gland, F., V. Monbet, & V.D. Tran (2011). Large sample asymptotics for the ensemble kalman filter. In D. Crisan and B. Rozovskii (Eds.), *Handbook on Nonlinear Filtering*. Oxford University Press.

Nakano, S., G. Ueno, & T. Higuchi (2007). Merging particle filter for sequential data assimilation. *Nonlinear Processes in Geophysics 14*.

Snyder, C., T. Bengtsson, P. Bickel, & J. Anderson (2008). Obstacles to high-dimensional particle filtering. *Monthly Weather Review 136*.

van Leeuwen, P. (2009). Particle filtering in geophysical systems. *Monthly Weather Review 137*.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

The Prior Probability in classifying two populations by Bayesian method

V.V. Tai
Department of Mathematics, Can Tho University, Can Tho, Vietnam

C.N. Ha
Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

N.T. Thao
Division of Computational Mathematics and Engineering, Institute for Computational Science, Ton Duc Thang University, Ho Chi Minh City, Vietnam
Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

ABSTRACT: This paper considers the Prior Probability (PP) in classifying two populations by Bayesian approach. Specially, we establish the probability density functions for the ratio and the distance between two PPs that be supposed to have Beta distributions. Also, we build the posterior distribution for PPs when knowing the prior Beta distributions of them. From established probability density functions, we can calculate some typical parameters such as mean, variance and mode. According to these parameters, we can survey and also estimate the prior probabilities of two populations to apply for practical problems. The numerical example in one synthetic data set, four bench mark data sets and one real data set not only illustrate for the proposed theories but also present the applicability and feasibility of the researched problem.

1 INTRODUCTION

Classification is an important problem in multivariate statistical analysis and applied in many fields, such as economics, physics, sociology, etc. In literature, there were many different methods proposed to perform the classification problem like logistic regression method, Fisher method, Support Vector Machine method, Bayesian method, etc., in which Bayesian approach was especially interested Mardia, Kent, & Bibby 1979, Webb 2003, Ghosh, Chaudhuri, & Sengupta 2006). In classifying by Bayesian approach, we often study the case of two populations because it can be applied in many practical problems and it is also the theoretical foundation for the case of more than two populations. We suppose to have two populations W_i, $i = 1, 2$ with q_i is the prior probability and $f_i(x)$ is the Probability Density Function (pdf) of the variable X for ith population, respectively. Acording to Pham-Gia, (Pham-Gia, Turkkan, & Vovan 2008), classifying a new observation x_0 by Bayesian method was performed by the rule: if $\max\{q_i f_i(x_0)\} = q_1 f_1(x_0)$ then x_0 is assigned to W_1, in contrast, we assign it to W_2. Pham-Gia (Pham-Gia, Turkkan, & Vovan 2008) also identified the misclassification in this approach that be called as the Bayes error and be calculated by formula

$$\begin{aligned} Pe &= \int_{R^n} \min\{q_1 f_1(x), q_2 f_2(x)\} \mathrm{d}x \\ &= 1 - \int_{R^n} g_{max}(x)\, \mathrm{d}x, \end{aligned}$$

in which $g_{max}(x) = \max\{q_1 f_1(x), q_2 f_2(x)\}$. Therefore, in Bayesian approach, classifying a new observation and computing its error depend on two factors: pdfs and PPs. From the given data, we have many methods to determine the pdfs. This problem was studied excitedly in theoretical aspect and had many good applications with real data (Pham-Gia, Turkkan, & Vovan 2008, Vo Van & Pham-Gia 2010). In fact, when knowing the exact pdfs, determining suitable PPs is a significant factor to improve the performance in Bayesian classification. Normally, depending on known information about the researched problem or the training data, we can determine the prior probabilities. If there is none of information, we usually choose prior probabilities by uniform distribution. When basing on training data, the prior probabilities are often estimated by two main methods: Laplace method: $q_i = \frac{n_i+1}{N+n}$ and ratio of samples method: $q_i = \frac{n_i}{N}$, where n_i is the number of elements in W_i, n is the number of dimensions and N is the number of all objects in training data (James 1978, Everitt 1985). There were many authors who studied and applied these results, such as (McLachlan & Basford 1988, Inman & Bradley Jr 1989, Miller, Inkret, Little, Martz, & Schillaci 2001). Besides, determining specified distributions for PPs is also interested in case of two populations. We can list some researches about this problem such as (McLachlan & Basford 1988, Jasra, Holmes, & Stephens 2005, Pham-Gia, Turkkan, & Bekker 2007). To inheritance their ideal, this article studies the PPs of two populations by building pdfs for the ratio and distance between PPs. According to prior information and sampling data, we can establish posterior pdfs for ratio and distance between two PPs. Then, we can estimate and test the differences between two prior probabilities. Because the sum of PPs is equal to 1, we can survey and determine the PPs for two populations when knowing their ratio or distance.

The remainder of this article is organized as follows. Section 2 presents the theories about the prior pdfs and posterior pdfs for ratio and distance between two PPs. In this section q is assumed to have Beta prior distribution and the posterior distribution of q is updated from the sample information. Section 3 discusses some relations of established pdfs in Section 2 and the computational problems in practices. Section 4 examines three numerical examples to illustrate proposed theories and compare the obtained results with those of existing methods. The final section is the conclusion.

2 THE RATIO AND THE DISTANCE BETWEEN TWO PRIOR PROBABILITIES

Give the variable X and two populations W_1, W_2, with pdfs are $f_1(x)$ and $f_2(x)$, respectively. q is the PP for W_1 and $1-q$ is the PP for W_2. Let $y=\frac{q}{1-q}, z=|1-2q|$. If q is a random variable, then y and z are also random variables. In this section, we build the prior pdfs for y and z when q has Beta distribution. The posterior distributions of y and z are also established when we consider the data samples. When posterior pdfs are computed, we will have a general look about the difference between two PPs and also find them via some representing parameters of y and z (e.g., mean or mode). This ideal can be also performed in a similar way when q has other distributions in [0,1].

2.1 *Distribution of the ratio between two prior probabilities*

Theorem 1 Assuming that q have the prior distribution Beta (α,β), we have following results for pdf of variable y.

a. The prior pdf of y is determied as follows:

$$f_{pri}(y)=\frac{1}{B(\alpha,\beta)}\frac{y^{\alpha-1}}{(y+1)^{\alpha+\beta}} \tag{1}$$

where

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \Gamma(\alpha)=\int_0^\infty x^{\alpha-1}e^{-x}\mathrm{d}x.$$

b. If Beta (α,β) is the prior distribution of q and m is the number of observations belonging to W_1 when collecting n elements, the posterior pdf of y is determined as follows:

$$f_{pos}(y)=\frac{1}{B\left(\hat{\alpha},\hat{\beta}\right)}\frac{y^{\hat{\alpha}-1}}{(y+1)^{\hat{\alpha},\hat{\beta}}} \tag{2}$$

where

$$\hat{\alpha}=\alpha+m, \hat{\beta}=\beta+n-m.$$

Proof

a. Because q has distribution Beta (α,β), pdf of q is:

$$f_{pri}(q)=\frac{1}{B(\alpha,\beta)}\left[q^{\alpha-1}(1-q)^{\beta-1}\right], \quad 0\le q\le 1.$$

Clearly, $q=y/(y+1)$ and the derivative of q is $1/(y+1)^2$. Thus,

$$\begin{aligned}f_{pri}(y)&=\tfrac{1}{(y+1)^2}f_{pri}\left(\tfrac{y}{y+1}\right)\\&=\tfrac{1}{B(\alpha,\beta)}\left(\tfrac{y}{y+1}\right)^{\alpha-1}\\&\qquad\left(1-\tfrac{y}{y+1}\right)^{\beta-1}\tfrac{1}{(y+1)^2}\\&=\tfrac{1}{B(\alpha,\beta)}\tfrac{y^{\alpha-1}}{(y+1)^{\alpha+\beta}}.\end{aligned}$$

b. We call A to be the event for obtaining m observations of W_1 when collecting n observations:

$$P(A)=\binom{m}{n}q^m(1-q)^{n-m}.$$

Then we have

$$\begin{aligned}M(q)&=\tfrac{1}{B(\alpha,\beta)}q^{\alpha-1}(1-q)^{\beta-1}.\\&\qquad\binom{m}{n}q^m(1-q)^{n-m}\\&=\frac{\binom{m}{n}}{B(\alpha,\beta)}q^{\alpha+m-1}(1-q)^{\beta-1+n-m}.\end{aligned}$$

The posterior pdf of q is

$$f_{pos}(q)=\frac{M(q)}{\int_0^1 M(q)\mathrm{d}q}=\frac{q^{\hat{\alpha}-1}(1-q)^{\hat{\beta}-1}}{B\left(\hat{\alpha},\hat{\beta}\right)}. \tag{3}$$

Doing like cases a) with $f_{pos}(q)$ in (3), we obtain (2).

2.2 *Distribution of the distance between two prior probabilities*

Theorem 2 Let z be distance between q and $(1-q)$, $z=|1-2q|$. We have following results for pdf of z

a. If q has distribution Beta (α,β), the pdf of z $(0\le z\le 1)$ is

$$\begin{aligned}g_{pri}(z)=C_1\Big[&(1-z)^{\alpha-1}(1+z)^{\beta-1}\\&+(1+z)^{\alpha-1}(1-z)^{\beta-1}\Big]\end{aligned} \tag{4}$$

where

$$C_1=\frac{1}{2^{\alpha+\beta-1}B(\alpha,\beta)}.$$

b. If q has prior distribution $Beta(\alpha,\beta)$ and m is the number of observations belonging to W_1 when choosing n elements, the posterior pdf of z is determined as follows:

$$\begin{aligned}g_{pos}(z)=C_2\Big[&(1-z)^{\hat{\alpha}-1}(1+z)^{\hat{\beta}-1}\\&+(1+z)^{\hat{\alpha}-1}(1-z)^{\hat{\beta}-1}\Big]\end{aligned} \tag{5}$$

where

$$C_2=\frac{1}{2^{\hat{\alpha}+\hat{\beta}-1}B\left(\hat{\alpha},\hat{\beta}\right)}.$$

Proof

a. We have prior pdf of q is

$$f_{pri}(q)=\frac{1}{B(\alpha,\beta)}q^{\alpha-1}(1-q)^{\beta-1}.$$

When $q\le 1/2$, $z=1-2q$ or $q=(1-z)/2$. In this range, pdf of z is determined by

$$g_1(z)=\frac{1}{2}f_{pri}\left(\frac{1-z}{2}\right)$$

where $f_{pri}\left(\frac{1-z}{2}\right)$ is given by

$$\frac{1}{B(\alpha+\beta)}\left(\frac{1-z}{2}\right)^{\alpha-1}\left(1-\frac{1-z}{2}\right)^{\beta-1}.$$

So we have

$$g_1(z)=C_1\left[(1-z)^{\alpha-1}(1+z)^{\beta-1}\right]. \tag{6}$$

For $q>1/2$, $z=2q-1$ or $q=(z+1)/2$, using the similar way to establish in (6), we have

$$g_2(z)=C_1\left[(1+z)^{\alpha-1}(1-z)^{\beta-1}\right]. \tag{7}$$

Clearly, from (6) and (7), we have (4).

b. We have

$$\begin{aligned}M(q)&=\frac{\binom{m}{n}q^{\alpha-1}(1-q)^{\beta-1}q^m(1-q)^{n-m}}{B(\alpha,\beta)}\\&=\frac{\binom{m}{n}q^{\alpha+m-1}(1-q)^{n-m+\beta-1}}{B(\alpha,\beta)}.\end{aligned}$$

So posterior pdf of q is

$$\begin{aligned}f_{pos}(q)&=\frac{M(q)}{\int_0^1 M(q)\mathrm{d}q}\\&=\frac{q^{\hat{\alpha}-1}(1-q)^{\hat{\beta}-1}}{B\left(\hat{\alpha},\hat{\beta}\right)}.\end{aligned}$$

It is easy to proof (5) by using the same way to (a).

Moreover, if we set $w=1-2q$, $-1\le w\le 1$, and the conditions in part (b) (theorem 2) are unchanged, we can proof the following result:

$$g_{pos}(w)=\frac{1}{2^{\hat{\alpha}+\hat{\beta}-1}B\left(\hat{\alpha},\hat{\beta}\right)}(1-w)^{\hat{\alpha}-1}(1+w)^{\hat{\beta}-1}. \tag{8}$$

3 SOME RELATIONS

3.1 *Some surveys*

i. Because q is in the interval [0,1], y is in the interval $[0,+\infty)$. If $q=1/2$ then $y=1$. When $q\to 0$, we have $y\to 0$. Similarly, when $q\to 1$, we have $y\to+\infty$. When $y\to 0$ or $y\to+\infty$ we receive the best classification. We also have z is a random variable whose value change from 0 to 1. When $q=1/2$, we have $z=0$ and when $q=0$ or $q=1$ we receive $z=1$. In the classification problem, if there is none of prior information, we often choose $y=1$ or $z=0$. When $y\to 0$, $y\to+\infty$ or $z=1$, we obtain the best classification.

ii. The k th moment for posterior of y and z is determined by

$$E_{pos}\left(y^k\right) = \frac{1}{B\left(\hat{\alpha},\hat{\beta}\right)}\int_0^1 \frac{y^{k+\hat{\alpha}-1}}{(y+1)^{\hat{\alpha}+\hat{\beta}}}\mathrm{d}y. \tag{9}$$

$$\begin{aligned} E_{pos}\left(z^k\right) = C_2\int_0^1 z^k \Big[&(1-z)^{\hat{\alpha}-1}(1+z)^{\hat{\beta}-1} \\ &+ (1+z)^{\hat{\alpha}-1}(1-z)^{\hat{\beta}-1}\Big]\mathrm{d}z. \end{aligned} \tag{10}$$

According to above equations, we can compute easily the means and the variances of y and z (Berg 1985) by the help of some mathematical software packages in Matlab, Maple, etc.

iii. When having the posterior pdfs of y and z, we can compute the highest posterior density (hpd) regions for them. The hpd credible interval $I_{1-\alpha}$ is often numerically computed although tables exist for some distributions (Isaacs 1974). Berger (Berger 1985) had proposed the algorithm to determine hpd and Turkan and Pham-Gia (Pham-Gia, Turkkan, & Eng 1993) written a program to determine hpd in different cases of distributions.

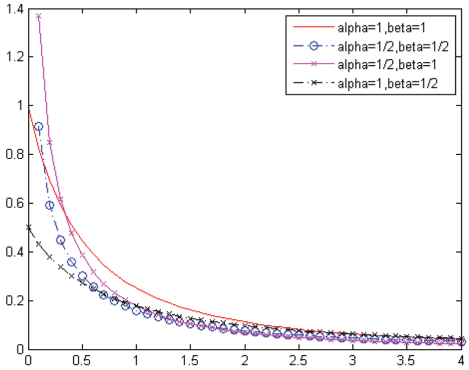


Figure 1. The prior pdf of y when q has distribution Beta(1/2,1/2), Beta(1/2,1), Beta(1,1), Beta(1,1/2).

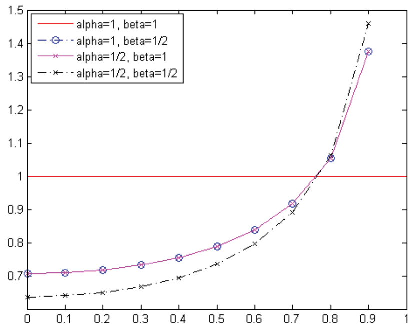


Figure 2. The prior pdf of z when q has distribution Beta(1/2,1/2), Beta(1/2,1), Beta(1,1), Beta(1,1/2).

iv. The building pdfs determined by (1), (2), (4) and (5) depends on the prior pdf of q. In practice, this distribution is not easy to survey. It really depends on the known information about the research. Although there have been a lot of authors discussing about this problem such as (McLachlan & Basford 1988), (Inman & Bradley Jr 1989), (Miller, Inkret, Little, Martz, & Schillaci 2001) none of optimal solution for all cases. According to (Pham-Gia, Turkkan, & Eng 1993) there are at least two of prior information that be often used for ratio between two Beta distributions. These are uniform distribution (or Beta(1,1)) and Jeffreys prior (or Beta(1/2, 1/2)) see Figures 1–2.

3.2 *The computational problem*

Because the features of populations are often discrete, we must estimate their pdfs before running Bayesian method. There are some proposed methods to solve this problem; however, kernel function method is the most popular one in practice. In this method,the choices of smoothing parameter and kernel function has effects on the result. Although (Scott 1992), (Martinez & Martinez 2007), (Vo Van & Pham-Gia 2010), etc. have had many discussions about this problem, the optimal choice is not found. Here, the smoothing parameter is chosen by Scott and the kernel function is the Gaussian.

When the prior probabilities and the pdfs have been identified, we have to find the maximum function $g_{\max}(x)$ to compute the Bayes error. In the unidimensional case, we can use the specified expression to compute the maximum function of pdfs and the Bayes error (Pham-Gia, Turkkan, & Vovan 2008). In the multidimensional case, this calculation is really complicated. Vovan and Pham-Gia (Vo Van & Pham-Gia 2010) and some other researchers have mentioned about this problem. In this case, the Bayes error is estimated by the Monte-Carlo simulation with the help of some mathematical software packages in Maple, Matlab, etc.

In this article, the programs used for estimating pdf, computing the Bayes error, are coded by Matlab software.

4 THE NUMERICAL EXAMPLE

This section examines three examples to illustrate the proposed theories in Section 2 and 3. Example 1 considers a synthetic data set containing 20 observations in two populations. Population I includes 9 observations and 11 ones for population II. We survey this simple example to test the theoretical results in Section 2 and compare the performance of the proposed method with those of other choices, which compute prior probabilities according to Uniform distribution, ratio of sample method and Laplace method.

Example 2 compares the results of surveying methods throughout four bench mark data sets including Seed, Thyroid, User and Breast Tissue. For each data set, we choose randomly two populations for experiment. These popular data sets are often studied in recognized statistics. When there is a new method that relates to classification problem, these data are also used to compare the result of the new method with traditional ones. In the third example, we resolve a practical issue in Vietnam: appraising the ability to repay loans of the bank costumers in BinhPhuoc province. In this section, the prior probability chosen by Uniform distribution, ratio of sample method, Laplace method, y and z are respectively denoted by BayesU, BayesP, BayesL, BayesR and BayesD. In cases of BayesR and BayesD, from the posterior pdf of y or z, we calculate the mean value and use it as the prior probability of population.

Example 1. Give the studied marks (scale 10 grading system) of 20 students, in which 9 students have marks being smaller than 5 (W_1: fail the exam) and 11students have marks being larger or equal to 5 (W_2: pass the exam). The data set is given by the Table 1. Assuming that we need to classify the ninth object and the prior probability q of W_1 is a random variable having distribution Beta(10,5). The training set presents that the total observations $N = 19$ and the number of observations in W_1 is $n = 8$. Then, we have:

The mean, variance and mode of posterior distribution of y are 1.2, 0.1886, 1.0, respectively. The 95% hpd credible interval of y: (1.2845, 1.6739).

The mean, variance and mode of posterior distribution of z are 0.1438, 0.0112, 0, respectively. The 95% hpd credible interval of z: (0.0962, 0.1914).

Using the mean value of y and z, we have the prior probabilities of two populations respectively are (0.5455;0.4545) and (0.5719;0.428). According to the prior probabilities and those of the existing methods, we classify the ninth element. The results presented by the following Table 2. It can be seen that only BayeR and BayesD give the right classification for the ninth object. These methods also have the smaller Bayes error than those of others. This result presents that if we choose the suitable prior probability for q, using proposed method, we can be received a better classification than other traditional methods.

Table 1. The studied marks of 20 students and the actual result.

Objects	Marks	Group	Objects	Marks	Group
1	0.6	W_1	11	5.6	W_2
2	1.0	W_1	12	6.1	W_2
3	1.2	W_1	13	6.4	W_2
4	1.6	W_1	14	6.4	W_2
5	2.2	W_1	15	7.3	W_2
6	2.4	W_1	16	8.4	W_2
7	2.4	W_1	17	9.2	W_2
8	3.9	W_1	18	9.4	W_2
9	4.3	W_1	19	9.6	W_2
10	5.5	W_2	20	9.8	W_2

Example 2. In this example, BayesU, BayesP, BayesL, BayesR and BayesD will be used to classify some bench mark data sets that include Thyroid, Seeds, User, and Breast Tissue. In each data set, we choose randomly two populations. The summary of data features is presented by Table 3. The detailed data sets are given by [http://archive.ics.uci.edu/ml].

The survey of bench mark data whose sizes, dimensions are various will show the effectiveness and the stability of new method. Assuming that the prior probability q of W_1 is a random variable having distribution Beta ([N/2],[N/2]) (N is the number of elements in data set). Applying (2) and (5) with n and m got from training data, we compute the posterior pdfs of y and z. Also, using the mean value for each case, we calculate the prior probabilities of populations. In this example, the ratio between training and test set is 1:1. The results that we received when running 10 times randomly are summarized in Table 4. Table 4 shows that BayesR, BayesD are more stable than the existing ones. In almost of data sets, BayesR and BayesD have the smaller errors than other methods.

Example 3. In bank credit operation, determining the repay ability of customers is really important. If the lending is too easy, the bank may have bad debt problems. In contrast, the bank will miss good opportunities to lend. Therefore, in recent years, the classification of credit applicants has been especially studied in Vietnam. In this example, the data including 27 cases of bad debt and 33 cases of good debt of a bank in BinhPhuoc province, Viet-

Table 2. The result when classifying the ninth object.

Method	Prior	$g_{\max}(x_0)$	Population	Bayes error
BayesU	(0.5;0.5)	0.0353	2	0.0538
BayesB	(0.421;0.579)	0.0409	2	0.0558
BayesL	(0.429;0.571)	0.0403	2	0.0557
BayesR	(0.5545;0.454)	0.0365	**1**	0.0517
BayesD	(0.572;0.428)	0.0383	**1**	0.0503

Table 3. Summary of four bench mark data sets.

Data	No of objects	No of dimensions
Thyroid	185	5
Seed	140	7
Breast	70	9
Users	107	5

Table 4. Summary five Bayesian methods of bench mark data.

Data	Method	Empirical error (%)
Thyroid	BayesU	1.304
	BayesP	1.196
	BayesL	1.196
	BayesR	**0.979**
	BayesD	1.195
Breast	BayesU	8.284
	BayesP	**7.427**
	BayesL	**7.427**
	BayesR	7.713
	BayesD	**7.427**
Seeds	BayesU	**3.715**
	BayesP	4.001
	BayesL	4.001
	BayesR	3.857
	BayesD	**3.715**
Users	BayesU	12.643
	BayesP	15.661
	BayesL	15.661
	BayesR	**12.264**
	BayesD	**12.264**

Table 5. The results of five Bayesian methods for Example 3.

	Empirical error (%)	Bayes error
Method	(X, Y)	(X, Y)
BayesU	**20**	0.1168
BayesP	20.33	0.1170
BayesL	20.33	0.1170
BayesR	21	0.1170
BayesD	**20**	0.1170

nam will be considered. The objects in data who are bank borrowers are immigrants. Two independent features are X (years of schooling) and Y (years of immigration). Because of the sensitive problem, authors have to conceal the detailed data. In this example, the choice prior for Beta distribution, the ratio between training and test set, the way to perform experiment are similar to Example 2. The results are presented in Table 5. It can be seen that the result in this example is quite similar to two previous ones. Especially, BayesD ensure the reasonable for all cases and give the best result.

5 CONCLUSION

This article establishes the prior and posterior distributions for the ratio and the distance between two prior probabilities having Beta distribution in Bayesian classification. From related pdfs that have been built, we can survey, compute typical parameters for prior probabilities of populations themselves. The numerical examples proved that if we have good prior information, choose the reasonable prior distributions of prior probabilities, we can determine the prior probabilities which give the better results in the comparison to traditional methods. In the coming, we will use these results to apply in different real data.

REFERENCES

Berg, A.C. (1985). SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

Everitt, B.S. (1985). Mixture Distributions. *Encyclopedia of statistical sciences* (5), 559–569.

Ghosh, A.K., P. Chaudhuri, & D. Sengupta (2006). Classification Using Kernel Density Estimates. *Technometrics 48*(1).

Inman, H.F. & E.L. Bradley Jr (1989). The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics-Theory and Methods 18*(10), 3851–3874.

Isaacs, G.L. (1974). *Tables for Bayesian statisticians*. Number 31. University of Iowa.

James, I.R. (1978). Estimation of the mixing proportion in a mixture of two normal distributions from simple, rapid measurements. *Biometrics*, 265–275.

Jasra, A., C.C. Holmes, & D.A. Stephens (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 50–67.

Mardia, K.V., J.T. Kent, & J.M. Bibby (1979). *Multivariate analysis*. Academic press.

Martinez, W.L. & A.R. Martinez (2007). *Computational statistics handbook with MATLAB*. CRC press.

McLachlan, G.J. & K.E. Basford (1988). Mixture models: Inference and applications to clustering. *Applied Statistics*.

Miller, G., W.C. Inkret, T.T. Little, H.F. Martz, & M.E. Schillaci (2001). Bayesian prior probability distributions for internal dosimetry. *Radiation protection dosimetry 94*(4), 347–352.

Pham-Gia, T., N. Turkkan, & A. Bekker (2007). Bounds for the Bayes error in classification: a Bayesian approach using discriminant analysis. *Statistical Methods and Applications 16*(1), 7–26.

Pham-Gia, T., N. Turkkan, & P. Eng (1993). Bayesian analysis of the difference of two proportions. *Communications in Statistics-Theory and Methods 22*(6), 1755–1771.

Pham-Gia, T., N. Turkkan, & T. Vovan (2008). Statistical discrimination analysis using the maximum function. *Communications in StatisticsSimulation and Computation 37*(2), 320–336.

Scott, D.W. (1992). Multivariate Density Estimation: Theory practice and visualization.

Vo Van, T. & T. Pham-Gia (2010). Clustering probability distributions. *Journal of Applied Statistics 37*(11), 1891–1910.

Webb, A.R. (2003). *Statistical pattern recognition*. John Wiley & Sons.

Efficient methods to solve optimization problems

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Estimation of parameters of Rikitake systems by SOMA

T.D. Nguyen
Vietnam Aviation Academy, Ho Chi Minh, Vietnam

T.T.D. Phan
HCMC University of Food Industry, Ho Chi Minh, Vietnam

ABSTRACT: This paper aims to present the combination of chaotic signal and Self-Organizing Migrating Algorithm (SOMA) to estimate the unknown parameters in chaos synchronization system via the active — passive decomposition method. The unknown parameters were estimated by self-organizing migrating algorithm. Based on the results from SOMA, two Rikitake chaotic systems were synchronized.

1 INTRODUCTION

Chaos theory is one of the most important achievements in nonlinear system research. Chaos dynamics are deterministic but extremely sensitive to initial conditions. Chaotic systems and their applications to secure communications have received a great deal of attention since Pecora and Carroll proposed a method to synchronize two identical chaotic systems (Pecora & Carroll 1990). The high unpredictability of chaotic signal is the most attractive feature of chaos based secure communication. Several types of synchronization have been considered in communication systems. The Active Passive Decomposition (APD) method was proposed by Kocarev and Parlitz (1995), it was known as one of the most versatile schemes, where the original autonomous system is rewritten as controlled system with the desired synchronization properties. Many of the proposed solutions focused on synchronization-based methods for parameter estimation (Shen & Wang 2008, Ge & Chen 2005), among others. In (Parltiz & Junge 1996), the parameters of a given dynamic model were estimated by minimizing the average synchronization error using a scalar time series.

Recently, a new class of stochastic optimization algorithm called Self-Organizing Migrating Algorithm (SOMA) was proposed in literature (Zelinka 2004 & Zelinka 2008). SOMA wors on a population of potential solutions called specimen and it is based on the self-organizing behavior of groups of individuals in a ''social environment''. It was proven that SOMA has ability to escape the traps in local optimal and it is easy to achieve the global optimal. Therefore, SOMA has attracted much attention and wide applications in different fields mainly for various continuous optimization problems. However, to the best of our knowledge, there is no research on SOMA for estimation of the parameter of chaos synchronization via ADP method.

Motivated by the aforementioned studies, this paper aims to present the combination of chaotic signal and the unknown parameters in chaos synchronization system were estimated via ADP method. Based on the results from SOMA algorithm, the estimated parameters were used to synchronize two chaotic systems.

2 PROBLEM FORMULATION

2.1 *The active-passive decomposition method*

Kocarev & Parlitz (1995) proposed a general drive response scheme named as Active Passive Decomposition (APD). The basic idea of the active passive synchronization approach consists in a decomposition of a given chaotic system into an active and a passive part where different copies of the passive part synchronize when driven by the same active component. In the following, we explain the basic concept and terminology of the active passive decomposition.

Consider an autonomous n-dimensional dynamical system, which is chaotic as

$$\dot{\mathbf{u}} = \mathbf{g}(\mathbf{u}) \tag{1}$$

The system is rewritten as a non-autonomous system:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{s}) \tag{2}$$

where x is a new state vector corresponding to **u** and **s** is some vector valued function of time given by

$$\mathbf{s} = \mathbf{h}(\mathbf{x}) \tag{3}$$

The pair of functions **f** and **h** constitutes a decomposition of the original vector field **g**, and are chosen such that any system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y},\mathbf{s}) \tag{4}$$

given by the same vector field **f**, the same driving signal **s**, but different variables **y**, synchronizes with the original system. Here, **x** constitutes the active system while **y** is the passive one.

The synchronization of the pair of identical systems. (2) and (4) occurs if the dynamical system describing the evolution of the difference $||y_k - x_k||$ $\rightarrow 0$ for $k \rightarrow \infty$.

2.2 *The parameter estimation*

When estimating the parameters, suppose the structure of the system is known in advance, the transmitter (driver) system is set with original parameters and the parameter in receiver (response) system is unknown. Therefore, the problem of parameter estimation can be formulated as the following optimization problem:

$$\text{Cost function}=\sqrt{\frac{1}{M}\sum_{k=1}^{M}\left\|y_k - x_k\right\|^{2}} \tag{5}$$

where M denotes length of data used for parameter estimation, the parameter can be estimated by minimum the Cost function (5).

Because of the unstable dynamic behavior of chaotic systems, the parameter estimation for chaotic systems is a multidimensional continuous optimization problem, the parameters are not easy to obtain. In addition, there are often multiple variables in the problem and multiple local optimums in the landscape of Cost function, so traditional optimization methods are easy to trap in local optima and it is difficult to achieve the global optimal parameters. Therefore, SOMA was chosen because it has been proven that the algorithm has the ability to converge toward the global optimum.

3 SELF ORGANIZING MIGRATING ALGORITHM

SOMA is the evolutionary algorithms which imitates nature process of wildlife migration. The method was established in 1999, developed by Prof. Ivan Zelinka at the University of Tomas Bata, Zlín. SOMA is a stochastic optimization algorithm that is modeled on the social behavior of cooperating individuals. The approach is similar to that of genetic algorithms, although it is based on the idea of a series of "migrations" by a fixed set of individuals, rather than the development of successive generations. It can be applied to any cost-minimization problem with a bounded parameter space, and is robust to local minima. SOMA works on a population of candidate solutions in loops called migration loops. The population is initialized randomly distributed over the search space at the beginning of the search. In each loop, the population is evaluated and the solution with the highest fitness becomes the leader L. Apart from the leader, in one migration loop, all individuals will traverse the in put space in the direction of the leader. Mutation, the random perturbation of individuals, is an important operation for Evolutionary Strategies (ES). It ensures the diversity amongst the individuals and it also provides the means to restore lost information in a population. Mutation is different in SOMA compared with other ES strategies. SOMA uses a parameter called PRT to achieve perturbation. This parameter has the same effect for SOMA as mutation has for GA.

The novelty of this approach is that the PRT vector is created before an individual starts its journey over the search space. The PRT vector defines the final movement of an active individual in search space.

The randomly generated binary perturbation vector controls allowed dimensions for an individual. If an element of the perturbation vector is set to zero, then the individual is not allowed to change its position in the corresponding dimension.

An individual will travel a certain distance (called the Path Length) towards the leader in n steps of defined length. If the Part Length is chosen to be greater than one, then the individual will overshot the leader. This path is perturbed randomly.

There are specified following parameters of SOMA algorithm:

Cost function: determines how to evaluate individuals.

Specimen: describes a form of individuals.

Population size: The number of individuals in the population which is contained in one migration.

Migrations: The maximum number of migrations to complete.

Step: The step size of individual during migration.

Part Length: duration of path which use individuals for migration.

PRT: perturbation of migration.

Minimal diversity: diversity of evolutionary process.

More detailed description of SOMA can be found in e.g. (Zelinka 2004).

There are many of SOMA variations which are differentiated by way of migration. In our case, SOMA-All-To-One variation has been chosen, in which individuals migrate past the best one.

4 RIKITAKE SYSTEM'S

In this section, we apply the ADP technique to achieve the synchronization between two identical Rikitake systems. The mathematical description of Rikitake system is as follows 0:

$$u=\begin{cases}\dot{x}=-\mu x+zy\\ \dot{y}=-\mu y+(z-a)x\\ \dot{z}=1-xy\end{cases}\tag{6}$$

where x, y and z are the state variables, and **μ** and **a** are the positive real constants. The Rikitake system (4) exhibits a chaotic attractor for **μ** = 2 and **a** = 5 as shown in Figure 1.

To illustrate the synchronization of two identical Rikitake systems, we consider different active-passive decompositions of the drive system with the denote X and the response system with the denote Y.

The identical drive system X (**μ** = 2 and **a** = 5) is given by:

$$X=\begin{cases}\dot{x}_d=-2x_d+z_d s(t)\\ \dot{y}_d=-2y_d+(z-5)x_a\\ \dot{z}_d=1-x_d y_d\end{cases}\tag{7}$$

The response system Y is described by the following equations:

$$X=\begin{cases}\dot{x}_r=\mu x_r+z_r s(t)\\ \dot{y}_r=-\mu y_r+(z-a)x_j\\ \dot{z}_r=1-x_r y_r\end{cases}\tag{8}$$

where **a**, **μ** are unknown parameter of response system, s(t) is the transmitted signal.

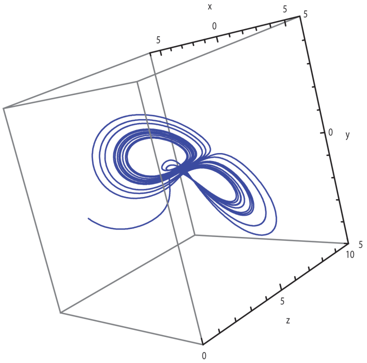


Figure 1. The Rikitake chaotic attractor.

Table 1. SOMA parameter setting.

Parameter	Value
Population size	20
Migrations	50
Step	0.11
Path length	3
Perturbation	0.1
Minimal diversity	–1

Subtracting system (7) from system (8) yields the error dynamical system between system (7) and system (8) $e_k = ((x_r, y_r, z_r)_k - (x_d, y_d, z_d)_k)$ were used to create a cost function **CF** representing the Root Mean Square Error (RMSE) of synchronization between **X** and **Y**:

$$\text{CF}=\sqrt{\frac{1}{M}\sum_{k=1}^{M}\left\|Y_k(x_r,y_r,z_r)-X_k(x_d,y_d,z_d)\right\|^2}\tag{9}$$

The parameter estimation can be formulated as a multidimensional nonlinear problem to minimize the cost function **CF**. SOMA is used to find a suitable parameter **a**, **μ** such that the cost function **CF** can be asymptotical approach to minimum point. The minimum value of cost function guarantee of the best solution with suitable parameters. Systems are asymptotically (and globally) synchronized.

In our simulations, the transmitted signal is chosen **s**(t) = y_d. The initial states of the drive system (7) and the response system (8) are taken as $x_d(0) = 6$, $y_d(0) = 0$, $z_d(0) = 0$ and $x_r(0) = 1$, $y_r(0) = 2$, $z_r(0) = 1$, respectively. Hence the error system has the initial values $e_1(0) = 5$, $e_2(0) = -2$ and $e_3(0) = -1$. SOMA-All-To-One is used to solve the systems, which the control parameters setting are given in Table 1. Simulation was implemented using Mathematica programming language and executed on Pentium D 2.0 G, 2 GB personal computer.

4.1 *Case study 1: simulation on one-dimensional parameter estimation*

In this case, we consider one-dimensional parameter estimation. That mean one parameter **a** (or **μ**) is known with the original value, one of **μ** (or **a**) are unknown and need to be estimated.

a. When **a** = 5 is known in advance, the initial guesses are in the range [0,5] for **μ**, the control parameter was set as Table 1. SOMA-All-To-One has found the best result was collected with **μ** = 1.97209 as shown in 3D cost function (Fig. 3). After 3 migrations, both the worst and the best values of the cost function approaches minimum value 0.634789 quickly as shown in Figure 2,

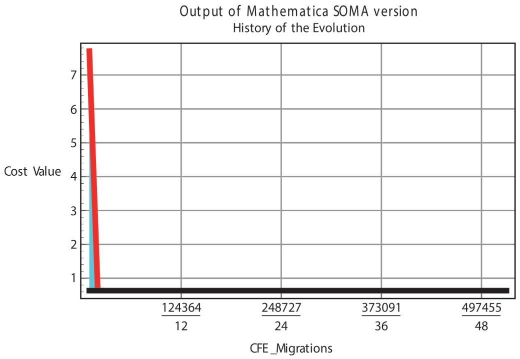


Figure 2. The worst and the best values of cost function (1a).

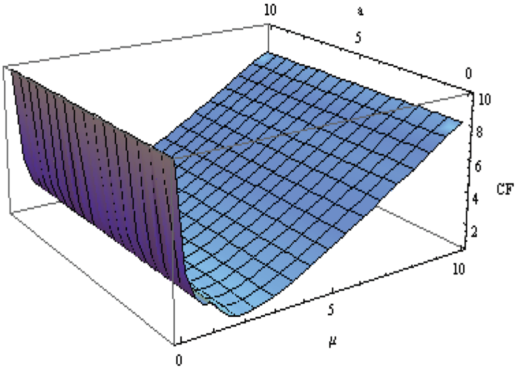


Figure.3. 3D cost function (**a** = 5).

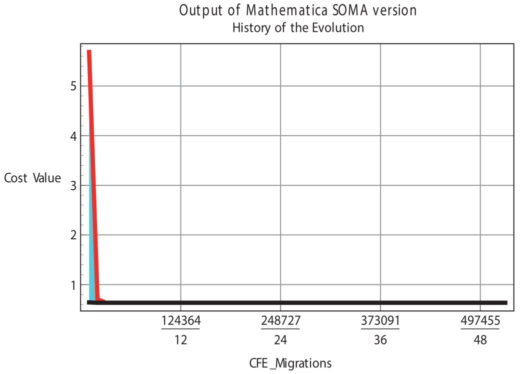


Figure 4. The worst and the best values of cost function (1b).

b. When **μ** = 2 is known in advance, the initial guesses are in the range [0, 10] for **a**, the control parameter was set as Table 1. SOMA-All-To-One has found the best result was collected with a = 4.99441 as shown in 3D cost function (Fig. 5). Both the worst and the best values of the cost function approaches minimum quickly as shown in Figure 4 (CF = 0635367).

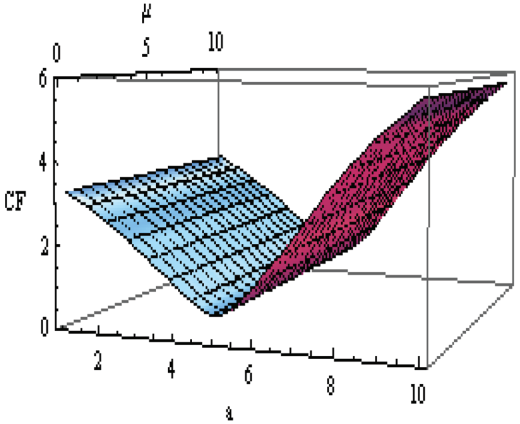


Figure 5. 3D cost function (**μ** = 2).

4.2 *Case study 2: simulation on two-dimensional parameter estimation*

In this case, we consider two-dimensional parameter estimation. Both two parameter **a**, and **μ** are unknown and need to be estimated. The initial guesses are in the range [0, 5] for **μ** and [0, 10] for **a**. SOMA-All-To-One has found the best result (CF = 0.634578) was collected with parameters **μ** = 1.9658 and **a** = 4.98468 as shown in 3D cost function (Fig. 7) Both the worst and the best values of cost function approaches minimum gradually after 24 migrations as shown in Figure 6.

The final estimated value are **μ** = 1.9658 and **a** = 4.98468. Thus, the actual parameters were fully identified. As shown in Figure (2–7), the values of cost function always approach to original minimum value CF = 0.635398, and the estimated parameter obtained by SOMA are also very close to the true value of original parameters. So, it's proven that SOMA is effective to estimate parameters for chaos synchronization system.

Base on the values were estimated by SOMA (**μ** = 1.9658 and **a** = 4.98468), the response system was constructed. The effective of the estimated value on the synchronization errors of driver systems X (4,4,1) and response system Y(−1,−1,−1) via ADP were demonstrated as shown in Figure 8.

Without ADP, the synchronization between two systems were not identified totally as shown in Figure.8 (a,e,i), and the trajectories of e(t) were unpredicted as shown in Figure.8 (c,g,k). In the opposite,

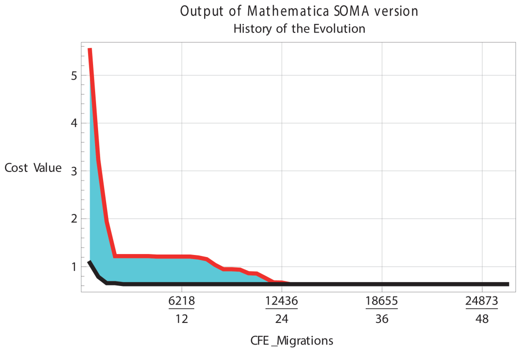


Figure 6. The worst and the best values of cost function (2).

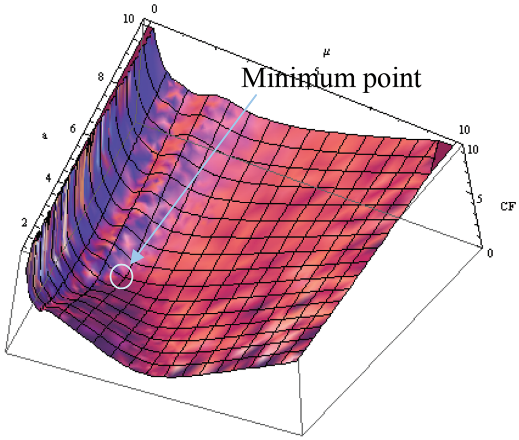


Figure 7. 3D cost function.

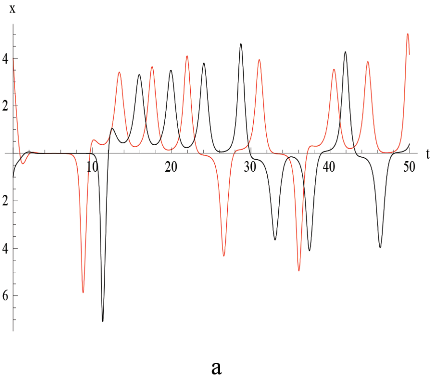


Figure 8a. Synchronization of x_d and x_r without using ADP.

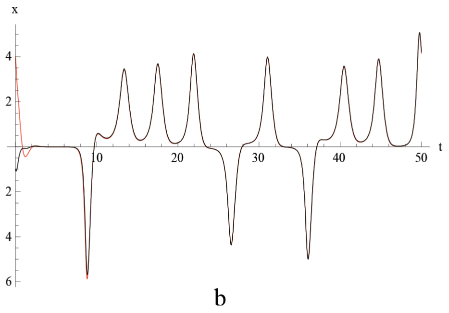


Figure 8b. Synchronization of x_d and x_r with ADP.

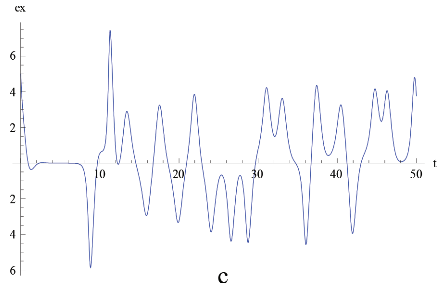


Figure 8c. Difference of x_d and x_r without using ADP.

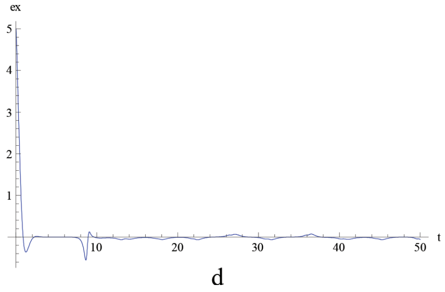


Figure 8d. Difference of x_d and x_r with ADP.

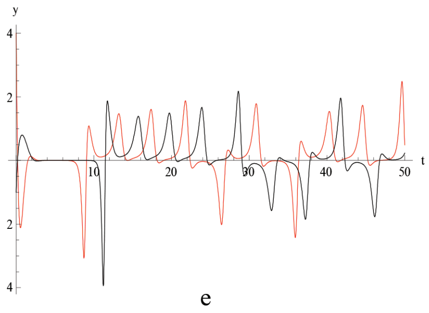


Figure 8e. Synchronization of y_d & y_r without using ADP.

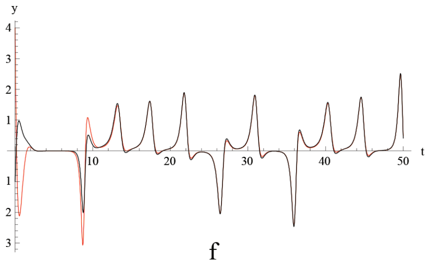


Figure 8f. Synchronization of y_d and y_r with ADP.

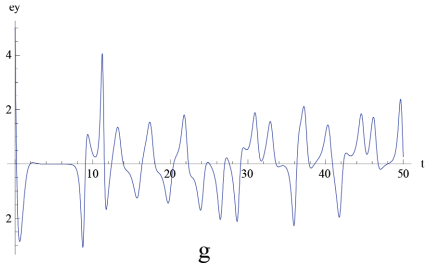


Figure 8g. Difference of y_d and y_r without using ADP.

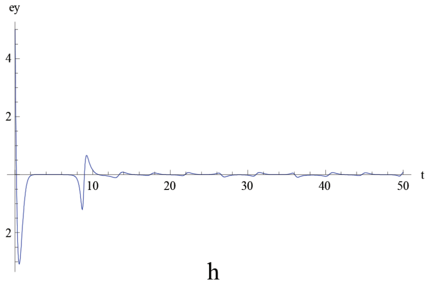


Figure 8h. Difference of y_d and y_r with ADP.

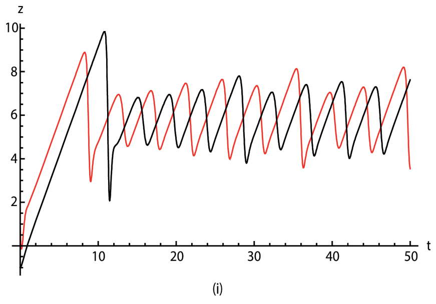


Figure 8i. Synchronization of z_d and z_r without using ADP.

Figure. 8 (d,h,l) displays that the trajectories of e(t) tends to zero after t > 12, and trajectories of x_r,y_r,z_r converged to x_d,y_d,z_d absolutely when ADP was applied as shown in Figure. 8 (b,f,j). It's proven that the estimated values and ADP method are effective to synchronize for two chaotic systems.

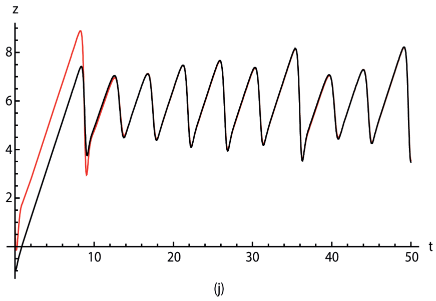


Figure 8j. Synchronization of z_d and z_r with ADP.

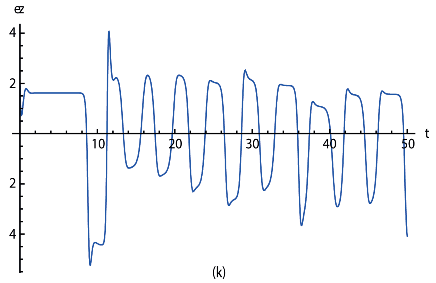


Figure 8k. Difference of z_d and z_r without using ADP.

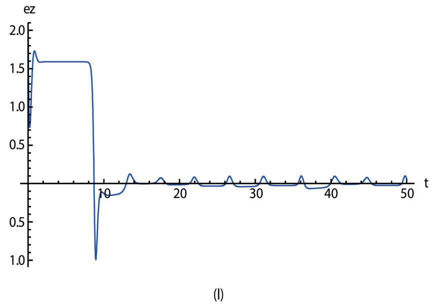


Figure 8l. Difference of z_d and z_r with ADP.

5 CONCLUSIONS

In this paper, the ADP method is applied to synchronize two identical Rikitake chaotic systems. Parameter estimation for chaotic system was formulated as a multidimensional optimization problem. Self-Organizing Migration Algorithm (SOMA) was used to find the unknown values of chaotic parameters. Based on the results from

SOMA algorithm, two chaotic systems were synchronized absolutely.

REFERENCES

Cuomo, K.M. & A.V. 1993. Oppenheim. Circuit implementation of synchronized chaos with applications to communications. *Phys. Rev. Lett.,71. 1:65–68*.

Ge Z.M, Cheng J.W. 2005. Chaos synchronization and parameter identification of three time scales brushless DC motor system. *Chaos, Solitons & Fractals, 24,* 597–616.

Guan X.P., Peng H.P., Li L.X. & Wang Y.Q. 2001. Parameter identification and control of Lorenz system. *Acta Phys Sin, 50,* 26–29.

Huang L.L., Wang M. & Feng R.P. 2005. Parameters identification and adaptive synchronization of chaotic systems with unknown parameters. *Phys Lett A, 342,* 299–304.

Keisuke Ito. 1980. Chaos in the Rikitake two-disc dynamo system, *Earth and Planetary Science Letters Vol. 51, 2*, 451–456.

Kocarev, L. & U. Parlitz. 1995. General Approach for Chaotic Synchronization with Applications to Communication. *Phys. Rev. Lett., 74,* 5028–5031.

Li R.H., Xu W. & Li S. 2007. Adaptive generalized projective synchronization in different chaotic systems based on parameter identification. *Phys Lett A*, *367,* 199–206.

Liao, X., Chen, G. & O. Wang. 2006. On Global synchronization of chaotic systems, *Dynamics of continuous discrete and impulsive systems, Vol 1*.

McMillen. 1999. The shape and dynamics of the rikitake attractor. *The Nonlinear Journal Vol.1, 1–10*.

Parlitz U, Junge L. 1996. Synchronization based parameter estimation from times series. *Phys Rev E, 54*, 6253–9.

Parltiz U. 1996. Estimating model parameters from time series by autosynchronization. *Phys Rev Lett,76,*1232–5.

Pecora, L.M. & T.L. Carroll. 1990. Synchronization in chaotic systems. *Phys. Rev. Lett*., *64(8),* 821–824.

Schuster, H.G. & W. Just. 2005. Deterministic Chaos. *Wiley VCH*.

Shen L.Q, Wang M. 2008. Robust synchronization and parameter identification on a class of uncertain chaotic systems. *Chaos, Solitons & Fractals, 38*, 106–111.

Wu X.G., Hu H.P. & Zhang B.L. 2004. Parameter estimation only from the symbolic sequences generated by chaos system. *Chaos, Solitons & Fractals, 22*, 359–366

Xiaoxin Liao, Guanrong Chen & Hua O Wang. 2002. On global synchronization of chaotic systems. *AK, May*.

Zelinka, I. 2008. Real-time deterministic chaos control by means of selected evolutionary techniques., *Engineering Applications of Artificial Intelligence*, *10*.

Zelinka, I. 2004. SOMA - Self-Organizing Migrating Algorithm," In: B.V. Babu, G. Onwubolu Eds., *New optimization techniques in engineering, Springer-Verlag, chapter 7*.

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Clustering for probability density functions based on Genetic Algorithm

V.V. Tai
Department of Mathematics Can Tho University, Can Tho, Vietnam

N.T. Thao
Division of Computational Mathematics and Engineering, Institute for Computational Science
Ton Duc Thang University, Ho Chi Minh City, Vietnam
Faculty of Mathematics and Statistics Ton Duc Thang University, Ho Chi Minh City, Vietnam

C.N. Ha
Faculty of Mathematics and Statistics Ton Duc Thang University, Ho Chi Minh City, Vietnam

ABSTRACT: Basing on the L^1-distance between Probability Density Functions (pdfs) in a cluster and its representing pdf, the L^1-distances between representing pdfs of different clusters, this article proposes two new internal validity measures for clustering for pdfs. Then, we apply Genetic Algorithm coded for solving integer optimization problems to minimize these internal validity measures so that establish the suitable clusters. The numerical examples in both synthetic data and real data will show that the proposed algorithm gives the better results than those of existing ones while testing by internal validity measures and external validity measures.

1 INTRODUCTION

In recent years, because of the fast development of networking, data storage, and the data collection capacity, there has been an increasing on data that we receive and exchange everyday, especially on big data. According to Wu et al. (Wu, Zhu, Wu, & Ding 2014), in every day, 2.5 quintillion bytes of data have been created and 90% of data in the world have been produced from 2009 to 2011. Therefore, how to analyse effectively the big data that have a huge volume and be received from many uncertain sources is a challenge for many researchers in data mining and statistics (George, Haas, & Pentland 2014, Wu, Zhu, Wu, & Ding 2014). Clustering that can partition unknown large data into groups so that elements in each group have the similar properties is a basic method in data mining and statistics. It is an important step to understand the data before performing further analysis. Therefore, the clustering problem has been researched extensively in many areas such as physics, biology, economics, engineering, sociology and any field that needs to group the similar elements together.

There are many algorithms can resolve the Clustering for Discrete Elements (CDE). These algorithms were summarized by Fukanaga and Webb (Keinosuke 1990, Webb 2003). However, because of the various strategies in CDE, the clustering results are also different. Therefore, how to evaluate these different results is an interesting question for many researchers. Generally, there are two main types of validity measures used to evaluate the quality of the clustering result: external validity measure and internal validity measure. Some popular external validity measures are Rand index (Rand 1971), F-index (Larsen & Aone 1999), Jaccard index, all of them evaluate the clustering throughout some specific references. Therefore, an external evaluation is impossible when we do not have any reference. The internal criteria considers some metrics which are based on data set and the clustering schema (analyze intrinsic characteristics of a clustering), so it can be performed for all cases. There were a large number of popular internal validity measures proposed in both non-fuzzy and fuzzy clustering as Intra index (MacQueen 1967), Xie-Beni index (S index) (Xie & Beni 1991), Dunn Index (Dunn 1973), DB index (Davies & Bouldin 1979). Most of them evaluate the quality of the clustering result by the compactness and the separation of clusters. Basing these internal validity measures, a lot of algorithms have been proposed to search the optimal value of these measures, so that the compactness and separation of established clusters are optimized. We can list a lot of studies using Genetic algorithm for CDE as (Falkenauer 1992, Jain, Murty, & Flynn 1999, Hruschka 2003, Agustn-Blas, Salcedo-Sanz, Jiménez-Fernández, Carro-Calvo, Del Ser, & Portilla-Figueras 2012). Besides, some other evolutionary approaches were also applied to resolve the clustering problems as Particle Swarm Optimization (Das, Abraham, & Konar 2008), Ant Colony algorithms (Jiang, Yi, Li, Yang, & Hu 2010), Artificial Bee Colony algorithms (Zhang, Ouyang, & Ning 2010). All of them have supplied a novelty approach to establish clusters and improve the quality of result.

The clustering for probability density functions (CDF) that be necessary for big data has been interested by many researchers recently. We can find some important studies in literature as Matusita (Matusita 1967), Glick (Glick 1972) which proposed some standard measures to compute the similarity of two or more pdfs (Vo Van & Pham-Gia 2010) which established the cluster width criterion and applied it for hierarchical approach and non-hierarchical approach in CDF, (Goh 2008, Montanari & Calò 2013) which proposed some novelty methods to build clusters in CDF and (Chen & Hung 2014) which introduced a method called "automatic clustering algorithm" to find the number of clusters and then establish optimal result. However, the validity measures used in above studies are external validity measures and none of previous studies proposed an internal validity measures in CDF. Therefore, it is impossible to perform the CDF when we do not have any reference. In addition, it cannot apply evolutionary approaches for optimizing the clusters without internal validity measures. Although Chen and Hung (Chen & Hung 2014) had proposed the automatic clustering algorithm but we cannot evaluate their result is whether really optimal or not. Furthermore, automatic clustering algorithm is easy to merge all of pdfs to a single cluster (number of clusters $k = 1$) when the pdfs have a high overlapping degree. Basing on the idea that optimize the compactness and separation of established clusters, this article proposes two internal validity measures in CDF. From them, we apply the Genetic Algorithm coded for solving integer optimization problem (Deep, Singh, Kansal, & Mohan 2009) to minimize these internal validity measures. Hence, the suitable clusters are established. The above algorithm is integrated in Global Optimization Toolbox in Matlab Software and is easy to use. The numerical examples in this article will show the proposed method can find the optimal internal validity measure. The results with optimal internal measures will be re-tested when using external measure (with references). It can be seen that the proposed algorithm improve significantly the performance of CDF.

This article is organized as follows. In section 2, we summarize some issues relating to L^1-distance, the representing pdf of cluster and propose two new internal validity measures. Section 3 reviews the Genetic Algorithm called as MI-LXPM presented in (Deep, Singh, Kansal, & Mohan 2009). Section 4 is the numerical examples that use MI-LXPM to optimize the proposed internal validity measures in Section 2. It will demonstrate our algorithm can improve the performance of CDF. Section 5 is the conclusion.

2 SOME RELATIONS

2.1 *L^1-distance and representing pdf*

Let $F = \{f_1(x), f_2(x), \ldots, f_N(x)\}, N > 2$ is the set of pdfs for k clusters, $C = (C_1, C_2, \ldots, C_k), k \geq 2$. Definition 1. L^1- distance of F is defined as follows: For $N > 2$,

$$\|f_1, f_2, \ldots, f_N\|_1 = \int_{R^n} f_{\max}(x)\mathrm{d}x - 1. \tag{1}$$

And for $N = 2$,

$$\|f_1, f_2\|_1 = \int_{R^n} | f_1(x) - f_2(x) | \mathrm{d}x, \tag{2}$$

where $f_{\max}(x) = \max f_1(x), f_2(x), \ldots, f_N(x)$. From (1), it is easy to prove that $\|f_1, f_2, \ldots, f_N\|_1$ is a non-decreasing function of N, with $0 \leq \|f_1, f_2, \ldots, f_N\|_1 \leq N - 1$. Equality on the left occurs when all f_i are identical and on the right when f_i have disjoint supports. From (2), we have

$$\frac{1}{2}\|f_1, f_2\|_1 = \int_{R^n} f_{\max}(x)\mathrm{d}x - 1.$$

2.2 *The representing probability functions of clusters*

Definition 2. Give the set of pdfs, $F = (f_1, f_2, \ldots, f_N)$, $N \geq 2$ which be separated to k clusters, $C = (C_1, C_2, \ldots, C_k)$, $k \geq 2$. The representing pdf for cluster C_1 is defined by

$$fv_i = \frac{\sum_{f_i \in C_i} f_i}{n_i}, \tag{3}$$

where n_i is the number of pdfs in cluster C_i. We also have $fv_i \geq 0$ for all x and $\int_{R^n} fv_i \mathrm{d}x = 1$.

2.3 *Two new proposed internal validity measures*

In this section, we propose IntraF and SF index to evaluate the quality of the established clusters in CDF. Two internal validity measures are presented as follows:

IntraF index

$$IntraF = \frac{1}{n}\sum_{i=1}^{k}\sum_{f\in C_i}\left\|f - fv_i\right\|^2 \qquad (4)$$

where $\left\|f - fv_i\right\|$ is the L^1 - distance between f and fv_i and n is the number of all pdfs.

The more similar between pdfs in cluster to their representing pdf are, the smaller IntraF is. Therefore, IntraF index reflects the compactness of established clusters and at first, we can see that it is suitable to evaluate the clusters quality.

SF index

The IntraF index can compute the compactness of clusters but cannot assess the separation between different clusters. Therefore, we propose the new index to measure this separation. This index is called as SF and it is defined as follows:

$$SF = \frac{\sum_{i=1}^{k}\sum_{f\in C_i}\left\|f - fv_i\right\|^2}{n\min\left(\left\|fv_i - fv_j\right\|^2\right)} = \frac{IntraF}{\min_{i\neq j}\left(\left\|fv_i - fv_j\right\|^2\right)} \qquad (5)$$

where $\left\|fv_i - fv_j\right\|$ is the L^1 distance between representing pdfs of cluster i and cluster j.

The SF index compute the pairwise L^1 - distance between all representing pdfs of all clusters. Then their minimum is considered as the separation measurement. The more separate between the clusters are, the larger denominator is and the smaller SF is. Thus, the smallest SF indeed indicates a valid optimal partition which consider both compactness and separation of clusters.

3 GENETIC ALGORITHM FOR SOLVING INTEGER OPTIMIZATION PROBLEM

Firstly, we have had to encode the solution in clustering problem to the chromosome before applying Genetic Algorithm to optimize the internal validity measures. Each individual is presented by a chromosome having the same length with the number of pdfs. The value l_j in each gene in the chromosomes represents the label of cluster to which jth pdf is assigned. For example, the clustering result with $C_1 = \{f_1, f_4\}, C_2 = \{f_2, f_5, f_7\}, C_3 = \{f_3, f_6\}$ is presented by the chromosomes: **1 2 3 1 2 3 2**.

The Genetic Algorithm for solving the integer optimization problems (Deep, Singh, Kansal, & Mohan 2009) is called MI-LXPM and presented as follows.

Crossover

The crossover operator used in (Deep, Singh, Kansal, & Mohan 2009) is the Laplace crossover. Give two individual $x^1 = \left(x_1^1, x_2^1, \ldots, x_n^1\right)$ and $x^2 = \left(x_1^2, x_2^2, \ldots, x_n^2\right)$, their offsprings $y^1 = \left(y_1^1, y_2^1, \ldots, y_n^1\right)$ and $y^2 = \left(y_1^2, y_2^2, \ldots, y_n^2\right)$ are generated in following way:

$$y_i^1 = x_i^1 + \beta_i \,|\, x_i^1 - x_i^2 \,|, y_i^2 = x_i^2 + \beta_i \,|\, x_i^1 - x_i^2 \,|, \qquad (6)$$

In (6), β_i satisfies the Laplace distribution and is generated as

$$\beta = \begin{cases} a - b\log(u_i) & if \quad r_i \leq \frac{1}{2} \\ a + b\log(u_i) & if \quad r_i > \frac{1}{2} \end{cases}$$

where a is location parameter and $b > 0$ is scaling parameter, $u_i, r_i \in [0,1]$ are uniform random numbers. For CDF problem, in each above individual, n is the number of pdfs and $2 \leq x_i \leq k$ with k is the number of clusters.

Mutation

The mutation operator used in MI-LXPM is the Power mutation. By it, a solution x is created in the vicinity of a parent solution $\overline{x}$ in the following manner.

$$x = \begin{cases} \overline{x} - s\left(\overline{x} - x^l\right) & if \quad t < r \\ \overline{x} + s\left(x^u - \overline{x}\right) & if \quad t \geq r \end{cases}$$

In above equation, s is a random number having power distribution and calculated by $s = \left(s_1\right)^p$, where s_1 is chose randomly in interval $[0,1]$ and p called the index of mutation is an integer number; $t = \frac{\overline{x} - x^l}{x^u - \overline{x}}$ where x^l and x^u be the lower and upper bounds on the value of the decision variable (in CDF $x^l = 2$ and $x^u = k$); r is a random number between 0 and 1.

Truncation procedure for integer restriction

In order to ensure that after crossover and mutation operations have been performed, the integer restrictions are satisfied, the following truncation procedure is applied. For all $i = 1, \ldots, n, x_i$ is truncated to integer value $\overline{x_i}$ by therule: If x_i is integer then $\overline{x_i} = x_i$, otherwise $\overline{x_i}$ is equal to $[x_i]$ or $[x_i]+1$ with the probability is 0.5, $[x_i]$ is the integer part of $\overline{x_i}$.

Selection

MI-LXPM use the tournament selection that presented by Goldberg and Deb (Goldberg & Deb 1991).

The above part presents the detailed MI-LXPM algorithm. This algorithm then be applied to optimize the SF index that has been proposed in Section 2 for solving problem of CDF. We call this hybrid algorithm, which be presented by below five step, as MI-LXPM-CDF:

St. 1 Starting with a randomly clustering solutions presented by chromosomes.
St. 2 Evaluating SF index for each clustering solution.
St. 3 Performing the genetic operations such as, selection, crossover, and mutation, on the current clustering solutions to introduce new ones.
St. 4 Replace the current clustering solution with the new ones having smaller SF index.
St. 5 If some criterion is met then stop, else go to St. 2.

The main ideal of MI-LXPM-CDF is that: throughout each iteration, from existing clustering solutions, we create some new ones and choose a determined numbers of best ones for the next iteration. In the end, we have the solution with the internal validity measures is optimized. Because MI-LXPM is an algorithm to find the global optimum, the new approach increase the chance to avoid trapping in local solution in the comparison with some hill climbing algorithms, such as k-means or non-hierarchical approach. The above algorithm is named as MI-LXPM-CDF whose suitability, feasibility, applicability will be tested by the numerical example in following section.

4 NUMERICAL EXAMPLES

In this section, we conduct three experiments to compare the proposed algorithm with Van and Pham-Gias non-hierarchical (Vo Van & Pham-Gia 2010) and the Automatic clustering of Chen and Hung (Chen & Hung 2014). In the first example, we consider seven univariate normal probability densities whose variances are the same and means are different. This is a simple examples presented in (Vo Van & Pham-Gia 2010). We review this example to illustrate the theoretical results, test the suitability of the proposed algorithm. The more complicate synthetic example researched in (Chen & Hung 2014) will be review in Example 2. This example contains 100 uniform distribution pdfs with dynamic parameter and separate into two groups with 50 pdfs in each group. In the final example, we apply the proposed algorithm for images recognition that be an interesting problem for many researchers in data mining with big data. We take 26 images from Caltech 101 dataset (Fei-Fei 2004). These 26 images contain 2 categories (lotus and sunflowers) with 13 images each. In each example, we apply MI-LXPM-CDF to optimize the internal validity measure, then comparing external measure (the error in the comparison with the truth) of the new algorithm with the existing ones in (Vo Van & Pham-Gia 2010, Chen & Hung 2014). The detailed results are shown as follows.

Example 1: We supposed to have seven populations with univariate normal pdfs, with specific parameters:

$$\sigma_1 = \sigma_2 = \cdots = \sigma_7 = 1;$$
$$\mu_1 = 0.3; \mu_2 = 4.0; \mu_3 = 9.1;$$
$$\mu_4 = 1.0; \mu_5 = 5.5; \mu_6 = 8.0; \mu_7 = 4.8.$$

Form Figure 1, it can be seen that the suitable separation for these pdfs is

$$C_1 = \{f_1, f_4\},\ C_2 = \{f_2, f_5, f_7\},\ C_3 = \{f_3, f_6\}$$

The clustering results of MI-LXPM-CDF and other algorithms are presented in Table 1.

It can be observed that all of methods are absolutely accurate in the comparison with the remark mentioned before. In fact, this is a simple and easy example. This be only used as the first test for our algorithm. The result verifies that our algorithm is

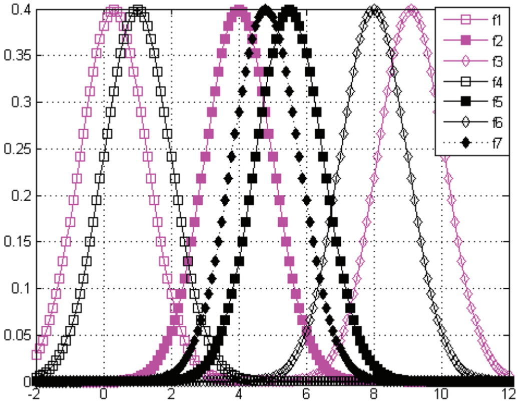


Figure 1. The pdfs of 7 univariate normal distributions.

Table 1. The results of MI-LXPM-CDF and existing algorithms.

	Misclustering rate (%)	SF index
Vo Van & Pham-Gia 2010	0	0.0493
Chen & Hung 2014	0	0.0493
MI-LXPM-CDF	**0**	**0.0493**

suitable at first and need to retest in more complicate example as follows.

Example 2: In this example, we review the synthetic data studied in (Chen & Hung 2014). The data consist of two classes f_1 and f_2 with 100 uniform pdfs on the interval $[0,1000]$ (see Figure 2). The pdfs of these two classes are defined as follows:

$$f_{1,i} = U(a_i, b_i),\ f_{2,i} = U(c_i, d_i),\ i = \overline{1,50},$$

where

$$a_i = 4(i-1) + \lambda_1, b_i = 195 + 5i + \lambda_2$$
$$c_i = 805 - 5j - \lambda_3, d_i = 1004 - 4j - \lambda_4$$

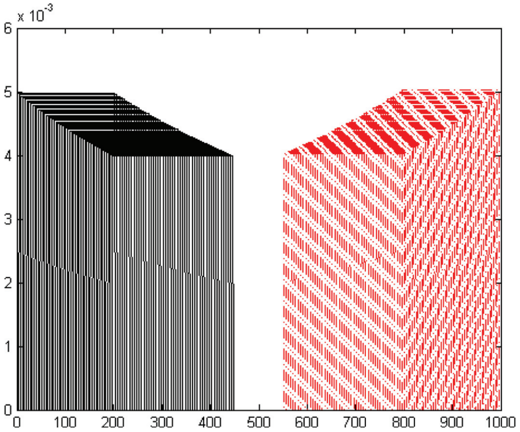


Figure 2. Two classes of pdfs f_1 (black) and f_2 (red).

with $U(a_i, b_i)$ and $U(c_i, d_i)$ denote the uniform distribution on the interval (a_i, b_i) and (c_i, d_i), respectively, $\lambda_1, \ldots, \lambda_4$ are drawn from $U(0,4)$.

We next create the mixtures of the above uniform distributions. Considering the first class $g_1 = f_1$ and the second class $g_2 = \lambda f_1 + (1-\lambda) f_2$, where $\lambda \in [0,1]$ and receive the value as 0, 0.1, 0.2, 0.3, respectively in this paper. Figure 3 shows the classes g_1 (black) and g_2 (red) in the cases $\lambda = 0.1$ and $\lambda = 0.3$. The clustering results of MI-LXPM-CDF and other methods for four cases of λ are presented in Table 2.

In this example, the larger λ is, the more overlapping degree between pdfs and the more complicated of problem are. It can be seen that our method improves significantly the performance of non-hierarchical approach in (Vo Van & Pham-Gia 2010) for all cases. Especially, MI-LXPM-CDF have the maximal accuracy in three cases $\lambda = 0$, $\lambda = 0.2$ and $\lambda = 0.3$. In case of $\lambda = 0.1$, the proposed method makes error being 5%, the reason is the algorithm cannot find the global optimum. It shows that our algorithm, especially mutation operator need some improvement in coming so that it can escape the local optimum and gives the better results. Anyway, the clustering results in Table 2 demonstrate that MI-LXPM-CDF is feasibility and improve the performance of clustering when evaluating by whether internal or external validity measures.

Example 3 In this example, we apply our algorithm in a more challenger problem being recognized images. The data set of images collected from 101 Objects database will be considered. We take 26 images in 2 categories (lotus and sunflowers) with 13 images each. The detail will be shown in Figure 4. We use the raw colour data in Grayscale

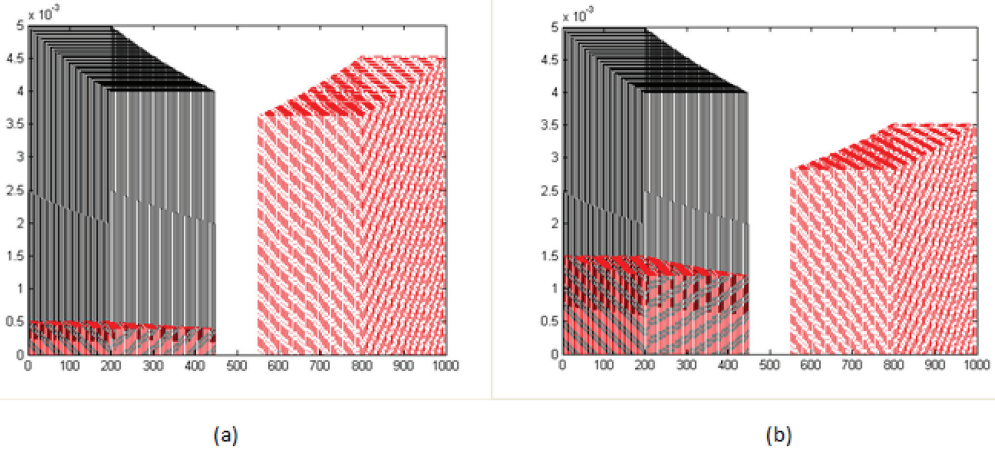


Figure 3. Two classes of pdfs g_1 (black) and g_2 (red) in cases of $\lambda = 0.1$ and $\lambda = 0.3$.

Table 2. The results of MI-LXPM-CDF and existing algorithms in each case of λ.

	Misclustering rate %			
	$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$
Vo Van & Pham-Gia 2010	9.2	9	8.8	13.4
Chen & Hung 2014	0	0	0	0
MI-LXPM-CDF	**0**	**5**	**0**	**0**

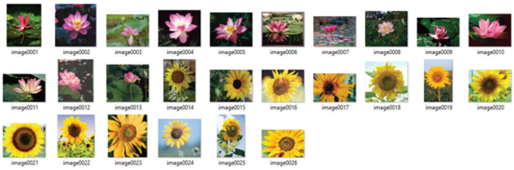


Figure 4. The detail of images data.

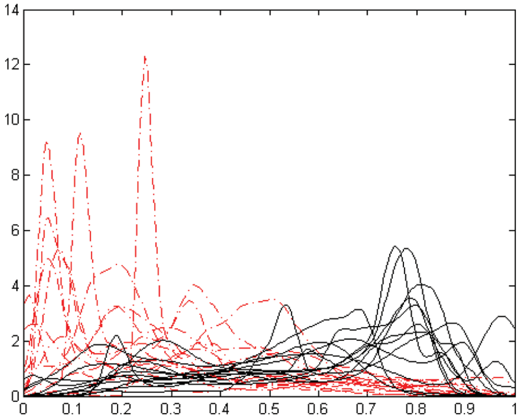


Figure 5. Two classes of pdfs: lotus (red), sunflowers (black).

Table 3. The results of MI-LXPM-CDF and existing algorithms.

Methods	Misclustering rate %
Vo Van & Pham-Gia 2010	31.61
Chen & Hung 2014	50
MI-LXPM-CDF	**11.54**

colour space for these images and estimate the pdfs for each image by the Grayscale distribution of image pixels.

Figure 5 shows the pdfs estimated from two classes of images with the red pdfs for lotus and the black pdfs for sunflowers. Each researched method is run 10 times and the average of misclustering rates (%) of all methods are showed in Table 3. In this example, the disadvantage of Automatic clustering be shown when it gives a single cluster with all pdfs ($k = 1$), therefore the misclustering rate of this method is 50%. The non-hierarchical approach gives the result with the average misclustering rate is 31.61% while MI-LXPM-CDF is the best with the error is 11.54%. It proves that the proposed algorithm can improve significantly the clustering performance and can be well applied in many practical problem in data mining with big data.

5 CONCLUSION

Basing on the L^1 - distance, the representing pdf of cluster and some related problems, this article proposes two internal validity measures named IntraF and SF index to evaluate the clustering results. The SF index be used as the object function needed to minimized. Further more, this article applies the Genetic Algorithms named MI-LXPM that be coded for solving integer optimization problem to find the optimal value of SF index in CDF. The proposed algorithm, MI-LXPM-CDF, is tested by external validity measure in many synthetic and real data sets. The numerical examples show MI-LXPM-CDF not only has good effects on simulation problems but also improve the clustering performance in practical problems, such as images recognition. Clearly, in the era of big data has arrived, with the uncertain and large volume data, this research and others which focus on the improvement of performance of clustering and classification problem are really necessary. The coming, MI-LXPM-CDF will be researched and improved some operators to increase it ability in searching the global optimal internal validity measure. Besides, some problems in data mining with big data such as, images, sound and video recognition will be researched.

REFERENCES

Agustn-Blas, L., S. Salcedo-Sanz, S. Jiménez-Fernández, L. Carro-Calvo, J. Del Ser, & J. Portilla-Figueras (2012). A new grouping genetic algorithm for clustering problems. *Expert Systems with Applications 39*(10), 9695–9703.

Chen, J. H. & W. L. Hung (2014). An automatic clustering algorithm for probability density functions. *Journal of Statistical Computation and Simulation* (ahead-of-print), 1–17.

Das, S., A. Abraham, & A. Konar (2008). Automatic kernel clustering with a multi-elitist particle swarm optimization algorithm. *Pattern recognition letters 29*(5), 688–699.

Davies, D.L. & D.W. Bouldin (1979). A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (2), 224–227.

Deep, K., K.P. Singh, M. Kansal, & C. Mohan (2009, June). A real coded genetic algorithm for solving integer and mixed integer optimization problems. *Applied Mathematics and Computation 212*(2), 505–518.

Dunn, J.C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.

Falkenauer, E. (1992). The grouping genetic algorithms widening the scope of the GAs. *Belgian Journal of Operations Research, Statistics and Computer Science 33*(1), 2.

Fei-Fei, R. (2004). L. and Fergus and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR Workshop on Generative-Model Based Vision*.

George, G., M.R. Haas, & A. Pentland (2014). Big data and management. *Academy of Management Journal 57*(2), 321–326.

Glick, N. (1972). Sample-based classification procedures derived from density estimators. *Journal of the American Statistical Association 67*(337), 116–122.

Goh, A. (2008). Unsupervised Riemannian Clustering of Probability Density Functions. pp. 377–392.

Goldberg, D.E. & K. Deb (1991). A comparative analysis of selection schemes used in genetic algorithms. *Foundations of genetic algorithms 1*, 69–93.

Hruschka, E.R. (2003). A genetic algorithm for cluster analysis. *Intelligent Data Analysis 7*(1), 15–25.

Jain, A.K., M.N. Murty, & P.J. Flynn (1999). Data clustering: a review. *ACM computing surveys (CSUR) 31*(3), 264–323.

Jiang, H., S. Yi, J. Li, F. Yang, & X. Hu (2010). Ant clustering algorithm with K-harmonic means clustering. *Expert Systems with Applications 37*(12), 8679–8684.

Keinosuke, F. (1990). Introduction to statistical pattern recognition. *Academic Press, Boston*.

Larsen, B. & C. Aone (1999). Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 16–22. ACM.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 281–297. Oakland, CA, USA.

Matusita, K. (1967). On the notion of affinity of several distributions and some of its applications. *Annals of the Institute of Statistical Mathematics 19*(1), 181–192.

Montanari, A. & D.G. Calò (2013). Model-based clustering of probability density functions.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association 66*(336), 846–850.

Vo Van, T. & T. Pham-Gia (2010). Clustering probability distributions. *Journal of Applied Statistics 37*(11), 1891–1910.

Webb, A.R. (2003). *Statistical pattern recognition*. John Wiley & Sons.

Wu, X., X. Zhu, G.-Q.Wu, & W. Ding (2014). Data mining with big data. *Knowledge and Data Engineering, IEEE Transactions on 26*(1), 97–107.

Xie, X.L. & G. Beni (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis & Machine Intelligence* (8), 841–847.

Zhang, C., D. Ouyang, & J. Ning (2010). An artificial bee colony approach for clustering. *Expert Systems with Applications 37*(7), 4761–4767.

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Optimization of truss structures with reliability-based frequency constraints under uncertainties of loadings and material properties

V. Ho-Huu, T. Vo-Duy & T. Nguyen-Thoi
Division of Computational Mathematics and Engineering, Institute for Computational Science, Ton Duc Thang University, Ho Chi Minh City, Vietnam
Faculty of Civil Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam

L. Ho-Nhat
Hochiminh University of Technology, Ho Chi Minh City, Vietnam

ABSTRACT: In this paper, the Reliability Based Design Optimization (RBDO) problem of truss structures with frequency constraints under uncertainties of loading and material properties is presented. Moreover, a double loop method with a new combination of an improved differential evolution algorithm which is proposed recently and an inverse reliability analysis is used for solving this problem. Three numerical examples including a welded beam, a 10-bar and 52-bar trusses are considered for evaluating the efficiency and application ability of the proposed approach.

1 INTRODUCTION

In the engineering discipline, the optimization always plays a very important role in designing of structures. The optimal design would help significantly reduce the cost and also improve the performance of the structures. However, engineering design problems are often related to uncertainties which derive from various sources like manufacturing process, material properties and operating environments, etc. These uncertainties may cause structures to suffer different working conditions from the initial design. Sometimes, this results in risks to structures. Therefore, considering influence of uncertain factors during the designing process is really necessary.

The optimization of truss structures with frequency constraints is to minimize the whole weight of the structures while frequency constraints must be satisfied. The design variables are the elements areas or/and nodal coordinates. For this kind of optimization problems, the frequency constraints play a important role for avoiding resonance phenomenon of structures (Grandhi 1993), but in mathematical aspect it is not easy to solve because of their highly nonlinear, non-convex and implicit function properties. Therefore, despite of being introduced in (Bellagamba & Yang 1981) and being presented in more details in (Kaveh & Zolghadr 2014), the structural optimization with frequency constraints still has a lot of rooms for improvement and attracts certain attention from researchers. Besides, the Reliability-Based Design Optimization (RBDO) for truss structures with frequency constrains has not yet been considered in the literature. Thus, it is really necessary to develop efficient tools for optimization of truss structures with reliability-based frequency constraints under uncertainties of loadings and material properties.

In solving the RBDO problems, the most direct approach is a double loop method in which the optimization loop (outer loop) is a deterministic optimization process; it repeatedly calls the reliability analysis loop (inter loop) in each cycle (Chen et al. 2013; Valdebenito & Schuëller 2010). The reliability analysis loop is a separate optimization problem which can be evaluated using direct methods such as the reliability index approach (Chen et al. 2013) or inverse methods such as inverse reliability strategy (Du et al. 2007; Du et al. 2004). In the double loop method, choosing a optimization algorithm in the optimization loop is of crucial important for solving a particular RBDO problem (Valdebenito & Schuëller 2010). For example, the gradient-based optimization methods such as Sequential Quadratic Programming (SQP), Generalized Reduced Gradient algorithm (GRG), etc. can be quite efficient for the optimization problems with explicit, convex and continuous objective functions, but they will be inefficient for the optimization problems with implicit, non-convex and discontinuous objective functions. This is because these methods use gradient information in searching progress. In contrast, the global optimization methods such as Genetic Algorithm (GA), Differential Evolution (DE), Particle Swarm

Optimization (PSO), etc. search solutions on the whole design space with only objective function information. Therefore, they could easily deal with various optimization problems. However, these methods are still costly in computational source for searching the global solution.

Recently, Ho-Huu et al. 2016 has proposed an adaptive elitist Differential Evolution (aeDE) algorithm for truss optimization with discrete variables. The aeDE is the newly improved version of the Differential Evolution (DE) algorithm based on three modifications. The effectiveness and robustness of the aeDE are verified through six optimization problem of truss structures. The numerical results demonstrated that the aeDE is more efficient than the DE and some other methods in the literature in terms of the quality of solution and convergence rate.

This paper hence tries to fill the above mentioned research gaps by solving the RBDO problem for truss with frequency constraint for the first time. For solving this problem, the double loop procedure is employed. In particular, for the optimization loop, the aeDE is employed while for the reliability analysis loop an inverse reliability strategy (Du et al. 2004) is used. Three numerical examples including a welded beam, a 10-bar and 52-bar trusses are considered for evaluating the efficiency and application ability of the proposed approach.

2 FORMULATION OF RBDO PROBLEM

A typical RBDO problem is defined by

$$\begin{array}{ll} \text{Minimize} & : f(\mathbf{d},\mathbf{x},\mathbf{p}) \\ \text{Design Variables:} & \text{DV} = \{\mathbf{d},\mu_x\} \\ \text{Subject to} & : \text{Prob}\{g_i(\mathbf{d},\mathbf{x},\mathbf{p}) \le 0\} \ge r_i,\ i = 1,2,\ldots,m. \end{array} \tag{1}$$

where $f(\mathbf{d},\mathbf{x},\mathbf{p})$ is the objective function; $\mathbf{d}$ is the vector of deterministic design variables; $\mathbf{x}$ is the vector of random design variables; $\mathbf{p}$ is the vector of random parameters; $g_i(\mathbf{d},\mathbf{x},\mathbf{p})$ ($i = 1,2,\ldots,m$) are constraint functions; r_i ($i = 1,2,\ldots,m$) $= \Phi(\beta_i)$ are desired probabilities of constraint satisfaction; m is the number of probabilistic constraints; $\Phi(.)$ is the standard cumulative function of the normal distribution; β_i is the target reliability index of the ith probabilistic constraint; $\boldsymbol{\mu}_\mathbf{x}$ is the mean of the random design variables $\mathbf{x}$.

3 INVERSE RELIABILITY ASSESSMENT METHOD

In conventional reliability analysis, the probabilistic constraint is checked by finding the probability of the constraint function g_i such that this probability is greater than or equal to a desired probability given by user. In presence of multiple constraints, however, some constraints may never be active and consequently their reliabilities are extremely high (approaching 1.0). Although these constraints are the least critical, the evaluations of these reliabilities will unfortunately dominate the computational effort in probabilistic optimization (Du et al. 2004). To overcome this drawback, Du et al. 2004 proposed an inverse reliability strategy in which the reliability assessment of the constraint function g_i is implemented only up to the necessary level. The brief description of the inverse strategy is summary as follows:

A percentile performance is defined as

$$g^{\alpha} \le 0 \tag{2}$$

where g^{α} is the α-percentile performance of $g(\mathbf{d},\mathbf{X},\mathbf{P})$, namely,

$$\text{Prob}\left\{g\left(\mathbf{d},\mathbf{X},\mathbf{P}\right) \le g^{\alpha}\right\} = \alpha \tag{3}$$

It has been shown in (Du et al. 2004) that $g \le 0$ indicates that $\text{Prob}\{g(\mathbf{d},\mathbf{X},\mathbf{P}) \le 0\} \ge \alpha$. Therefore, the original constraint is now converted to evaluate the α-percentile performance. More details of this strategy and method for evaluating the α-percentile performance may be found in (Du et al. 2004).

4 ADAPTIVE ELITIST DIFFERENTIAL EVOLUTION ALGORITHM

Among a variety of global optimization algorithms, the Differential Evolution (DE) algorithm first proposed by performance better than some other methods in the literature (Civicioglu & Besdok 2013; Vesterstrom & Thomsen 2004). However, it still requires high computational cost during the searching process. One of the main reasons for this restriction is that the DE did not keep the trade-off between the global and local search capabilities. Hence, in our previous work, we have proposed the adaptive elitist Differential Evolution (aeDE) algorithm which ensures the balance between the global and local search capabilities of the DE. Through six numerical examples for truss structures, it was shown that the aeDE better than the DE in terms both quality of solution and convergence rate. For more details of the aeDE, readers can refer to (Ho-Huu et al. 2016). In this paper, the aeDE is extended to solve the RBDO problem.

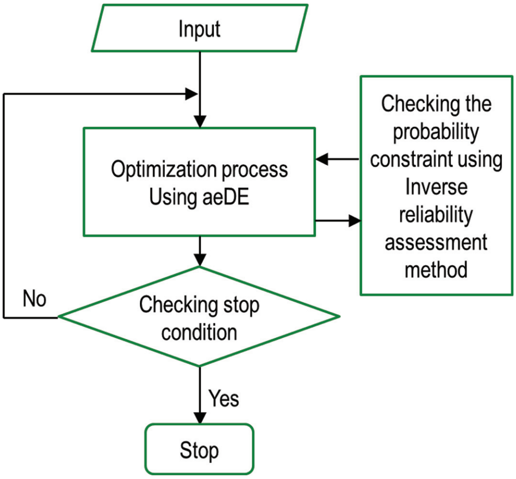

Input
Optimization process Using aeDE
Checking the probability constraint using Inverse reliability assessment method
Checking stop condition
No
Yes
Stop

Figure 1. Flow chart of the double-loop method.

5 A GLOBAL INTEGRATED FRAMEWORK FOR SOLVING RBDO PROBLEM

The aeDE and the inverse reliability strategy is integrated into the double loop procedure. This integration is named DLM-aeDE and is summarized in the flow chart of Figure 1.

6 NUMERICAL EXAMPLES

In this section, two numerical examples consisting of a welded beam and a 10-bar truss are considered. Because the RBDO for truss structures with frequency constraints has not been provided in the literature, a welded beam, a benchmark problem in the RBDO field, is presented as the first example to validate the accuracy of the implementation codes. Then, a 10-bar truss structure is carried out. The parameters of the aeDE including the population size *NP*, *threshold*, *delta* and *MaxIter* are set to 20, 10^{-3}, 10^{-6} and 1000, respectively. In this study, all codes including finite element analysis of the beam and the truss and the aeDE are written in Matlab.

6.1 *Welded beam*

The first example is a welded beam as shown in Figure 2. This beam was previous solved by Cho & Lee 2011 using a CL-SORA method, Hyeon & Chai 2008 using a moment-based RBDO method and Ho-Huu et al. 2016 using a SORA-ICDE method. The objective function is the welding cost. Five probabilistic constraints are related to physical quantities such as shear stress, bending stress, buckling, and displacement constraint.

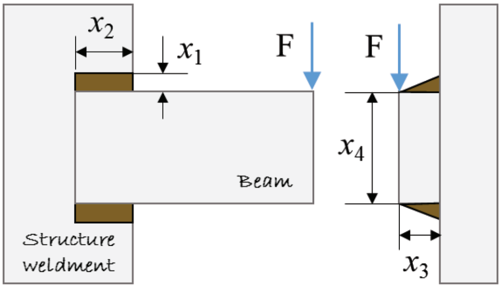


Figure 2. Flow chart of the double-loop method.

The RBDO problem has four random variables (x_1, x_2, x_3, x_4) which are statistically independent and follow normal distribution. The RBDO model of the welded beam problem is given by

$$\begin{aligned}&\text{find} \quad \mathbf{d} = [d_1,d_2,d_3,d_4]^{\mathrm{T}}\\&\text{minimize} \quad f(\mathbf{d},\mathbf{z}) = \mathrm{c}_1\, d_1^2 d_2 + \mathrm{c}_2\, d_3 d_4 (z_1 + d_2)\\&\text{subject to} \quad \mathrm{Prob.}\{g_j(\mathbf{x},\mathbf{z})<0\} \geq \Phi(\beta_j^t), j = 1,...,5\end{aligned} \tag{4}$$

where

$$g_1(\mathbf{x},\mathbf{z}) = \tau(\mathbf{x},\mathbf{z}) / z_6 - 1;\ g_2(\mathbf{x},\mathbf{z}) = \sigma(\mathbf{x},\mathbf{z})/z_7 - 1$$
$$g_3(\mathbf{x},\mathbf{z}) = x_1/x_4 - 1;\quad g_4(\mathbf{x},\mathbf{z}) = \delta(\mathbf{x},\mathbf{z})/z_5 - 1$$
$$g_5(\mathbf{x},\mathbf{z}) = 1 - P_{\mathrm{c}}(\mathbf{x},\mathbf{z})/z_1;$$
$$\tau(\mathbf{x},\mathbf{z}) = \{t(\mathbf{x},\mathbf{z})^2 + 2t(\mathbf{x},\mathbf{z})tt(\mathbf{x},\mathbf{z})\mathrm{X}_2/2\mathrm{R}(\mathbf{x}) + tt(\mathbf{x},\mathbf{z})^2\}^{1/2}$$
$$t(\mathbf{x},\mathbf{z}) = \frac{z_1}{\sqrt{2}x_1x_2};\quad tt(\mathbf{x},\mathbf{z}) = M(\mathbf{x},\mathbf{z})R(\mathbf{x})/\mathrm{J}(\mathbf{x})$$
$$M(\mathbf{x},\mathbf{z}) = z_1\left(z_2 + \frac{x_2}{2}\right);\quad R(\mathbf{x}) = \frac{\sqrt{x_1^2 + (x_1 + x_1)^2}}{2}$$
$$J(\mathbf{x}) = \sqrt{2}x_1x_2\{x_2^2 / 12 + (x_1 + x_3) / 4\}$$
$$\sigma(\mathbf{x},\mathbf{z}) = \frac{6z_1z_2}{x_3^2x_4};\quad \delta(\mathbf{x},\mathbf{z}) = \frac{4z_1z_2^3}{z_3x_3^3x_4}$$
$$x_i \sim N(d_i, 0.1693^2) \quad \text{for } i = 1,2$$
$$x_i \sim N(d_i, 0.0107^2) \quad \text{for } i = 3,4$$
$$\beta_1^t = \beta_2^t = \beta_3^t = \beta_4^t = \beta_5^t = 3;$$
$$3.175 \leq d_1 \leq 50.8;\quad 0 \leq d_2 \leq 254;$$
$$0 \leq d_3 \leq 254;\quad 0 \leq d_4 \leq 50.8$$
$$z_1 = 2.6688 \times 10^4(\mathrm{N});\ z_2 = 3.556 \times 10^2(mm);$$
$$z_3 = 2.0685 \times 10^5(\mathrm{MPa});\ z_4 = 8.274 \times 10^4(\mathrm{MPa});$$
$$z_5 = 6.35(\mathrm{mm});\ z_6 = 9.377 \times 10^1(\mathrm{MPa})$$
$$z_7 = 2.0685 \times 10^2(\mathrm{MPa});\ c_1 = 6.74135 \times 10^{-5}(\$ / \mathrm{mm}^3);$$
$$c_2 = 2.93585 \times 10^{-6}(\$ / \mathrm{mm}^3)$$

The obtained results of the DLM-aeDE are listed in Table 1 in comparison with those obtained by moment-based RBDO, SORA-ICDE and other methods. It can be seen that the results obtained

Table 1. Optimization results for welded beam problem.

		Hyeon & Chai 2008	Ho-Huu et al. 2016	This work
Design variable (mm)		Moment	SORA-ICDE	DLM-aeDE
x_1		5.729	5.730	5.728
x_2		200.59	201.00	201.089
x_3		210.59	210.63	210.610
x_4		6.238	6.240	6.239
Cost ($)		2.5895	2.5926	2.5923
	β_1	3.01	3.01	3.01
	β_2	3.52	3.29	3.07
Reliability index	β_3	3.01	3.00	3.01
	β_4	Infinite	Infinite	Infinite
	β_5	3.31	3.12	3.01

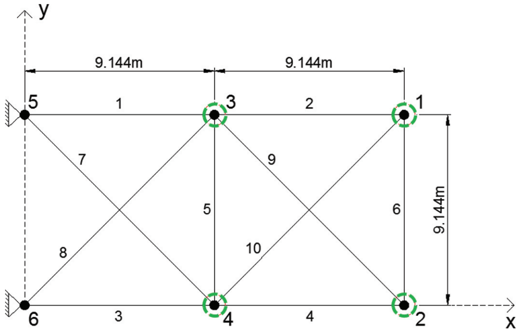


Figure 3. Model of a 10-bar planar truss structure.

Table 2. Data for the 10-bar planar truss structure.

Parameters (unit)	Value
Modulus of elasticity E (N/m^2)	6.89×10^{10}
Material density ρ (kg/m^3)	2770
Added mass (kg)	454
Allowable range of cross-sections (m^2)	$0.645 \times 10^{-4} \leq A \leq 50 \times 10^{-4}$
Constraints on first three frequencies (Hz)	$\omega_1 \geq 7$, $\omega_2 \geq 15$, $\omega_3 \geq 20$

by the DLM-aeDE are in good agreement with those gained by other studies. It can also be seen from Table 1 that all reliability levels are satisfied the required reliability indexes. These results demonstrate that the Matlab implementation of the DLM-aeDE is reliable and accurate.

6.2 *10-bar planar truss*

In the second example, a simple 10-bar truss structure, as depicted in Figure 3 is considered. All free nodes are added a non-structural mass of 454 kg. Data for the problem including the material properties, design variable bounds, and frequency constraints are summarized in Table 2. This example was investigated by some researchers such as Kaveh & Zolghadr 2014 utilizing democratic PSO, Zuo et al. 2014 using a hybrid algorithm between optimality criterion and genetic algorithm (OC-GA), etc. However, these studies are limited on solving the deterministic optimization problem in which the cross-sectional areas of bars are assumed to be independent design variables while Young's modulus and mass density of the truss and the added masses are fixed as given in Table 3. In this study, both the cross-sectional areas of bars, Young's modulus, mass density of the truss and the added masses are assumed to be the random design variables which have normal distribution with expected values equal to those of the Deterministic Optimization (DO) problem and standard deviation of 5%. The reliability indexes of all frequency constraints are set to be 3. This is equivalent to assume that the safety level of the structure must be greater than or equal to 99.865%.

The results of the DLM-aeDE are presented in Table 3 in comparison with those obtained by some methods for deterministic optimization. It can be seen that the reliability indexes for all frequency constraints are satisfied the required reliability indexes of the RBDO problem. The best weight obtained by the DLM-aeDE is 665.637 lb corresponding with the probability of safety of 99.865%. The results in Table 3 also show that the for the DO problem, the reliability of the structure is very low (around 50%). This illustrates that the

Table 3. Optimum results for 10-bar space truss structure.

		Kaveh & Zolghadr 2014	This work	
		Deterministic Optimization	Deterministic Optimization	Reliability-based Design Optimization
Design variable (area in^2)		DPSO	aeDE	DLM-aeDE
A_1		35.944	35.775	42.893
A_2		15.53	14.926	19.020
A_3		35.285	34.840	45.926
A_4		15.385	14.252	18.729
A_5		0.648	0.646	0.661
A_6		4.583	4.569	5.714
A_7		23.61	24.632	30.599
A_8		23.599	23.043	30.019
A_9		13.135	11.932	15.320
A_{10}		12.357	12.601	15.883
Weight (lb)		532.39	524.629	665.637
Reliability index (Probability of safety %)	β_1	–	0.00 (50.00%)	3.00 (99.86%)
	β_2	–	2.20 (1.35%)	4.79 (100%)
	β_3	–	0.00 (49.99%)	3.00 (99.86%)
Number of structural analyses		–	3940	774000

Table 4. Eight element group for the 52-bar dome truss structure.

Group number	Elements
1	1–4
2	5–8
3	9–16
4	17–20
5	21–28
6	29–36
7	37–44
8	45–52

safety of the whole truss is enhanced effectively and become more applicable in reality when the influence of uncertain factors during the designing process is taken in to account.

6.3 *52-bar dome truss*

In the last example, a simple 52-bar dome truss structure, as shown in Figure 4 is considered. All of the bars are arranged into eight groups as in Table 4. All free nodes are permitted to move ±2 m in each allowable direction from their initial position but again must guarantee symmetry for the whole structure. Therefore, there are five shape variables and eight sizing variables. The material properties, design variable bounds, and frequency constraints of the problem are given in Table 5. This

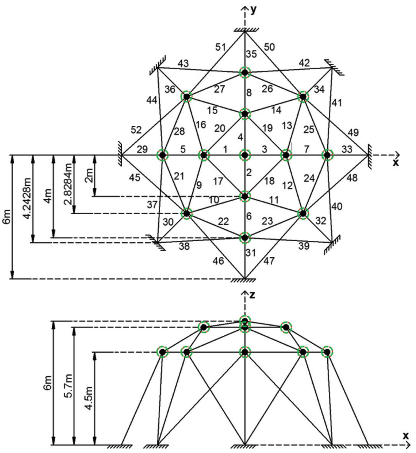


Figure 4. Model of a 52-bar dome truss structure.

Table 5. Data for the 52-bar dome truss structure.

Parameters (unit)	Value
Modulus of elasticity E (N/m^2)	2.1×10^{11}
Material density ρ (kg/m^3)	7800
Added mass (kg)	50
Allowable range of cross-sections (m^2)	$0.0001 \le A \le 0.001$
Constraints on first three frequencies (Hz)	$\omega_1 \le 15.961$, $\omega_2 \ge 28.648$

Table 6. Optimum results for 52-bar dome truss structure.

	Miguel & Fadel Miguel 2012	This work	
	Deterministic Optimization	Deterministic Optimization	Reliability-based Design Optimization
Design variable (area in^2)	FA	aeDE	DLM-aeDE
Z_A	6.4332	5.9889	4.0553
X_B	2.2208	2.2482	2.4973
Z_B	3.9202	3.7658	4.0568
X_F	4.0296	3.9865	3.9998
Z_F	2.5200	2.5005	2.6939
A_1	1.0050	1.0014	1.0106
A_2	1.3823	1.1288	1.0076
A_3	1.2295	1.1843	2.0040
A_4	1.2662	1.4444	2.0138
A_5	1.4478	1.3897	1.5572
A_6	1.0000	1.0002	1.0201
A_7	1.5728	1.5531	2.3314
A_8	1.4153	1.4354	2.2555
Weight (lb)	197.53	193.479	271.036
Reliability index	β_1 –	2.72 (99.67%)	3.00 (99.86%)
(Probability of safety %)	β_2 –	0.00 (50.03%)	3.00 (99.86%)
Number of structural analyses	–	7200	2100261

problem was previously studied by some researchers such as (Gomes 2011), (Miguel & Fadel Miguel 2012), (Khatibinia & Sadegh Naseralavi 2014), etc. However, similar to the 10-bar planar truss, these studies are limited on solving the deterministic optimization problem in which the cross-sectional areas of bars are assumed to be independent design variables while Young's modulus and mass density of the truss and the added masses are fixed as given in Table 5. In this study, both the cross-sectional areas of bars, Young's modulus, mass density of the truss and the added masses are assumed to be the random design variables which have normal distribution with expected values equal to those of the Deterministic Optimization (DO) problem and standard deviation of 5%. The reliability indexes of all frequency constraints are set to be 3 which is equivalent to the safety level of the structure of being greater than or equal to 99.865%.

The results of the problem are provided in Table 6 in comparison with those in (Miguel & Fadel Miguel 2012) for deterministic optimization. From Table 6, it can be seen that for the DO problem, the reliability of the structure is very low (around 50%) for the second constraint. This may lead to dangerousness for the structure when the input parameters are changed. On the other hand, with the results of the RBDO problem, the reliability of the structure may be ensured with the required safety levels.

7 CONCLUSIONS

In this study, the RBDO problem for truss structures with frequency constraints uncertainties of loading and material properties is presented. Moreover, the new double loop approach combining an inverse reliability method and an adaptive elitist differential evolution algorithm (DLM-aeDE) is employed to solve this problem. The proposed method is then applied for a welded beam and a 10-bar truss structure. The results reveal that (1) the DLM-aeDE is good competitor to the other algorithms for solving the RBDO problem; (2) the best solution of the RBDO for 10-bar and 52-bar trusses are found with reliability of 99.865%; (3) the RBDO for truss structures with frequency constraints make designing process of truss structures more practical in reality.

ACKNOWLEDGEMENTS

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 107.99-2014.11.

REFERENCES

Bellagamba, L. & Yang, T.Y. 1981. Minimum-mass truss structures with constraints on fundamental natural frequency. *AIAA Journal*, *19*(11), 1452–1458. http://doi.org/10.2514/3.7875.

Chen, Z., Qiu, H., Gao, L., Su, L. & Li, P. 2013. An adaptive decoupling approach for reliability-based design optimization. *Computers & Structures*, *117*(0), 58–66. http://doi.org/http://dx.doi.org/10.1016/j.compstruc.2012.12.001.

Cho, T.M. & Lee, B.C. 2011. Reliability-based design optimization using convex linearization and sequential optimization and reliability assessment method. *Structural Safety*, *33*(1), 42–50. http://doi.org/10.1016/j.strusafe.2010.05.003.

Civicioglu, P. & Besdok, E. 2013. *A conceptual comparison of the Cuckoo-search, particle swarm optimization, differential evolution and artificial bee colony algorithms. Artificial Intelligence Review* (Vol. 39). http://doi.org/10.1007/s10462-011-9276-0.

Du, X., Guo, J. & Beeram, H. 2007. Sequential optimization and reliability assessment for multidisciplinary systems design. *Structural and Multidisciplinary Optimization*, *35*(2), 117–130. http://doi.org/10.1007/s00158-007-0121-7.

Du, X., Sudjianto, A. & Chen, W. 2004. An Integrated Framework for Optimization Under Uncertainty Using Inverse Reliability Strategy. *Journal of Mechanical Design*, *126*(4), 562. http://doi.org/10.1115/1.1759358.

Gomes, H.M. 2011. Truss optimization with dynamic constraints using a particle swarm algorithm. *Expert Systems with Applications*, *38*(1), 957–968. http://doi.org/http://dx.doi.org/10.1016/j.eswa.2010.07.086.

Grandhi, R. 1993. Structural optimization with frequency constraints—A review. *AIAA Journal*, *31*(12), 2296–2303. http://doi.org/10.2514/3.11928.

Ho-Huu, V., Nguyen-Thoi, T., Le-Anh, L. & Nguyen-Trang, T. 2016. An effective reliability-based improved constrained differential evolution for reliability-based design optimization of truss structures. *Advances in Engineering Software*, *92*, 48–56. http://doi.org/10.1016/j.advengsoft.2015.11.001.

Ho-Huu, V., Nguyen-Thoi, T., Vo-Duy, T. & Nguyen-Trang, T. 2016. An adaptive elitist differential evolution for truss optimization with discrete variables. *Computer & Structures*, *165*, 59–75. http://doi.org/10.1016/j.compstruc.2015.11.014.

Hyeon Ju, B. & Chai Lee, B. 2008. Reliability-based design optimization using a moment method and a kriging metamodel. *Engineering Optimization*, *40*(5), 421–438. http://doi.org/10.1080/03052150701743795.

Kaveh, A. & Zolghadr, A. 2012. Truss optimization with natural frequency constraints using a hybridized CSS-BBBC algorithm with trap recognition capability. *Computers and Structures*, *102–103*, 14–27. http://doi.org/10.1016/j.compstruc.2012.03.016.

Kaveh, A. & Zolghadr, A. 2014. Democratic PSO for truss layout and size optimization with frequency constraints. *Computers & Structures*, *130*(0), 10–21. http://doi.org/http://dx.doi.org/10.1016/j.compstruc.2013.09.002.

Khatibinia, M. & Sadegh Naseralavi, S. 2014. Truss optimization on shape and sizing with frequency constraints based on orthogonal multi-gravitational search algorithm. *Journal of Sound and Vibration*, *333*(24), 6349–6369. http://doi.org/http://dx.doi.org/10.1016/j.jsv.2014.07.027.

Miguel, L.F.F. & Fadel Miguel, L.F. 2012. Shape and size optimization of truss structures considering dynamic constraints through modern metaheuristic algorithms. *Expert Systems with Applications*, *39*(10), 9458–9467. http://doi.org/10.1016/j.eswa.2012.02.113.

Valdebenito, M.A. & Schuëller, G.I. 2010. A survey on approaches for reliability-based optimization. *Structural and Multidisciplinary Optimization*, *42*(5), 645–663. http://doi.org/10.1007/s00158-010-0518-6.

Vesterstrom, J. & Thomsen, R. 2004. A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems. *Evolutionary Computation, 2004. CEC2004. Congress on*. http://doi.org/10.1109/CEC.2004.1331139.

Zuo, W., Bai, J. & Li, B. 2014. A hybrid OC–GA approach for fast and global truss optimization with frequency constraints. *Applied Soft Computing*, *14, Part C*(0), 528–535. http://doi.org/http://dx.doi.org/10.1016/j.asoc.2013.09.002.

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Optimum revenue calculation method to generate competitive hydroelectric power on Hua Na hydropower

Phan T.H. Long & L.Q. Hung
Water Resources University, Hanoi, Vietnam

Phan Dao
Ton Duc Thang University, Ho Chi Minh City, Vietnam

ABSTRACT: The paper introduces the calculate plan with better periods for flood control, calculation methods based on dynamic programming in the way of irregular mesh. The program was applied to calculate the Hua Na Hydropower, with two different operating models. The objective function to generate competitive hydroelectric power suggests the maximum revenue.

1 INTRODUCTION

1.1 *Competitive generating market*

The deregulation and development of electricity market proved that it is the advanced stage of management science in energy field. Electricity market (Figure 1) creates fair competitive business environment for participants and becomes outstanding solution to attract the investment and to raise the efficiency of production and business activities for the power industry. Electricity market has been developed all over the world, not only a single national market but also multi-national one trading the power among countries in the same region. ASEAN member countries such as Singapore, Philippine, Thailand, Malaysia, etc. have positive steps in forming their individual electricity market and toward the ASEAN electricity market in the future. Following the Prime Minister's Decision 63/2013/QD-TTg the competitive electricity wholesale market would be put under pilot implementation from 2016 to 2018 and was expected to begin operations officially from 2019.

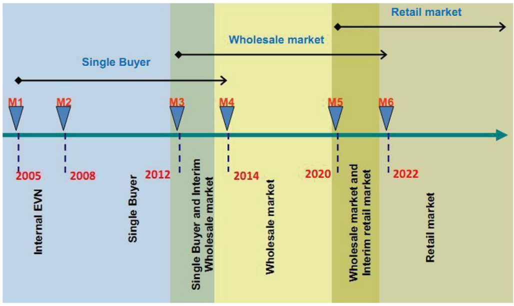


Figure 1. Roadmap of competitive power market in Vietnam.

Table 1. Basic param-eters of Hua Na HPP.

	Unit	Value
Installed Capacity	MW	180
Number of units		2
Maximum Capacity per unit	MW	90
Minimum Capacity per unit	MW	70
Average Annual Generation	GWh	716.6
Full water supply level	m	240
Pre-flood water level	m	235
Dead water level	m	215
Flood control volume	$10^6 m^3$	100
Total volume	$10^6 m^3$	569.35
Active volume	$10^6 m^3$	390.99
Area of reservoir	km^2	5,345
Maximum head	m	118.30
Design head	m	100
Minimum head	m	85.43
Turbine full-gate discharge	$m.s^{-1}$	203.4

1.2 *Hua Na hydropower plant*

Hua Na hydropower plant (basic parameters see in Table 1) is located in Nghe An province Que Phong district, Dong Van commune, on river Chu. With a total investment capital of VND 7,065 trillion, this hydropower plant was the first large-scale project of the Vietnam National Oil and Gas Group.

Following the Prime Minister's Decision 1911/2015/QD-TTg dated 5-Nov-2015 about the operating reservoirs on the Ma river basin, the Hua Na HPP must be moderate with following new conditions such as minimum level reservoir, term of flood control, etc. (unlike the process of design consultants) beside has engaged in competitive electricity markets. This is a problem that needs some solution of calculation to optimize revenue and limit excessive discharge.

2 METHODOLOGY

2.1 *Dynamic programming*

Dynamic programming is a technique used for optimizing a multistage process. It is a "Solution-seeking" concept which replaces a problem of *n* decision variables by *n* sub problems having preferably one decision variable each. Such an approach allows analysts to make decisions stage-by-stage, until the final result is obtained. For operating reservoir, the water levels have been divided from full water supply level to dead water level. For one month, with two values of water level at begin and end of month, the values of discharge, head, power and revenue will be calculated.

2.2 *Dividing power output for peak hours*

According the rules, Vietnam power market has 5 peak hours in working days. Every week has 5 working days. Thus a month has about 108 peak hours. Monthly power output is divided into two parts: One with high price and other with mean price.

2.3 *Selection calculation term*

Following the new conditions, such as flood control, the water level in reservoir always less than 235 m from 01/07 to 30/11; the time to calculate the optimal plan suggests to start from 01/12 years to 30/11 next years. But in Article 11, from 15/10, the water level can be rise to 240 m with condition good forecast hydrology. Two different operating models are store from 16/10 and store from 1/12 every year.

2.4 *Monthly revenue*

Based on dynamic programming calculation, for each month, three values of head, discharge and price have been determined. The prices could be increased, or changed by year or determined by ratio between dry season and rainy season (see Table 2). The result will be better if monthly price has been determined.

2.5 *Irregular meshing*

Two mesh have been applied: Regular mesh with per 1 m from dead water level (215 m) to full supply water level (240 m); Irregular mesh has two parts: per 0.1 m from 230 to 240 and 0.5 m from 215 to 230 m

Table 2. Price (VND) per power (KW) and (KWh).

	Dry season		Rainy season	
Price	KW	KWh	KW	KWh
Peak hour	250	2000	100	1500
Normal hour	250	900	100	500

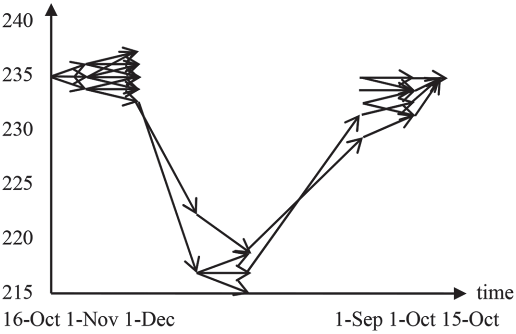


Figure 2. Model dynamic programming applied to one-dimensional storage reservoir with irregular mesh.

2.6 *Step-wise procedure of the algorithm*

Depending on the natural inflow, release capacity, and boundary conditions of reservoir, the maximum value of revenue for all reservoirs (in case of multiple reservoir system) at every time step of operating horizon are found out.

Considering the maximum revenue as in the code Visual Basic 2010 bellows:

```
If (amount(i, j, k) < (amount (i, j - 1, l) +
revenue_per_lapse(Head_turbine(i, j, k), j, Dis-
charge(i, j, k)))) Then
amount (i, j, k) = amount (i, j - 1, l) + reve-
nue_per_lapse (Head_turbine (i, j, k), j, Dis-
charge(i, j, k))
E(i, j, k) = E(i, j - 1, l) + Etb(i, j, k)
End If
```

At the end of period, as 15-Oct or 30-Nov, conventional dynamic programming is run through this corridor (see Figure 2) to find the trajectory, *water_level*, which gives maximum objective function value, *amount*.

3 RESULTS AND DISCUSSIONS

3.1 *Meshing method*

The use of meshing methods increases the amount and calculation time, but will result in more consistent calculations, the meshing with small distance in above and longer distance in the lower part corresponds to the share volume lake which parts are together. Method of finer meshing will gain better calculation results. However, for multi reservoirs, the application of finer meshing, for example distance in 1 cm, has increased rapidly amount of calculation. It is useful when applied to the calculation for multiple reservoirs. Model of irregular mesh is demonstrated in Figure 3.

3.2 *Result for chosen year and observed years*

The annual mean price has been shown on Table 3:

And the real mean price of Hua Na HPP from date operate generating has shown on Figure 4.

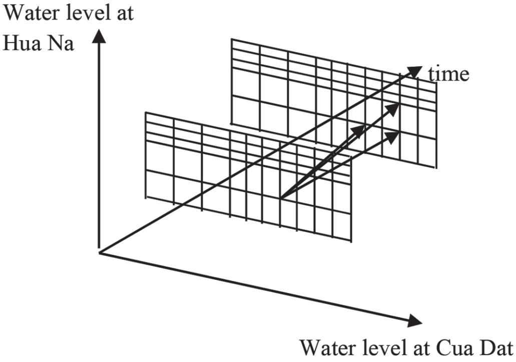


Figure 3. Model dynamic programming applied to two-dimensional storage reservoirs with irregular mesh.

Table 3. Annual mean price.

	Irregular mesh	Regular mesh
Mean price	VND/KWh	VND/KWh
Store since 16/10	1072.222	1071.822
Store since 1/12	1105.223	1105.036

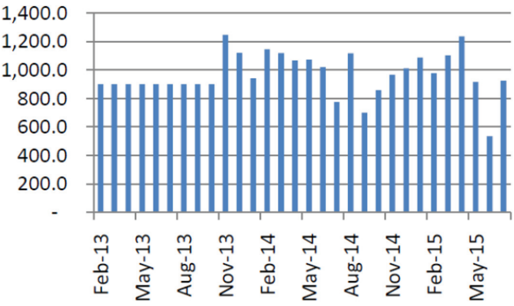


Figure 4. Real mean price of Hua Na HPP (VND/KWh).

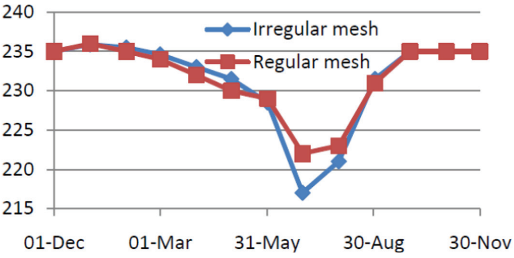


Figure 5. Evolution process upstream water level in chosen year.

Table 4. Annual revenue.

	Irregular mesh	Regular mesh
Setting	Billion VND	Billion VND
Store since 16/10	788.310	787.866
Store since 1/12	772.108	771.688

Table 5. In minimum water level conditions in dry season (Store since 16/10).

Valee	Unit	Value
Annual revenue	bil.VND	827,3478
Annual power out	GWh	679,7706
Annual mean price	VND	1217,099

For irregular mesh, the active volume has been used contrary to regular mesh. Water level at 30-Jun is 217 m in case irregular mesh compare to 222 m in case regular mesh. The comparison has been shown on Figure 5.

In 2015, all of reservoirs of Vietnam have the minimum water level in the dry season. In case Hua Na HPP, the new values have been calculated. Annual revenue results are clearly demonstrated and compared in Tables 4–5.

4 CONCLUSIONS

The operators which comply with the new reservoir may have more difficulty but with putting into operation the competitive electricity market will help businesses sell electricity at times of high prices on water shortages, lack of electricity. And in overall revenue divided by a power unit volume is higher than the fixed price.

The use of irregular meshing method increases the amount and time of calculation, but the result is suitable. The mesh with two parts corresponds to the share volume lake which parts are together.

The selected period will not be calculated depending on the hydrological year, which should be based on operational procedures. The price has been changed in the dry and rainy seasons of the electricity market, which may be different from the distribution season the river is applied.

For power plants competitive, the objective of the investor is the largest revenue, which sometimes will not fit with the calculation of the greatest amount of electricity anymore.

REFERENCES

Hua Na Hydropower JSC. 2015. Report on the "Assessment of change in average power of Hua Na NMTD when operating under inter-reservoir operation on the Ma River basin".

Ministry of Industry and Trade. 2014. The circular "Rules operate competitive electricity market".

Nandalal, K.D.W. & Bogardi J.J. 2013. Dynamic Programming Based Operation of Reservoirs *Applicability and Limits*, in Cambridge.

Prime Minister. 2013. The Decision "Regulation of the roadmap, the conditions and electricity industry structure to form and develop the level of the electricity market in Vietnam".

Prime Minister. 2015. The Decision "Promulgating operation Process reservoir on the Ma River basin".

Maintenance modelling and optimization

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

A dynamic grouping model for the maintenance planning of complex structure systems with consideration of maintenance durations

H.C. Vu, A. Barros & M.A. Lundteigen
Department of Production and Quality Engineering, Norwegian University of Science and Technology, Trondheim, Norway

P. Do
Research Centre for Automatic Control, Lorraine University, Nancy, France

ABSTRACT: In the framework of maintenance optimization of multi-component systems, dynamic grouping has developed and becomes an interesting approach. However, most existing dynamic grouping models assume that the repair time is negligible. This assumption may be not always relevant and limits the application of these models in many real situations. The main objective of this paper is to develop a dynamic grouping model taking into account both preventive and corrective maintenance duration for complex structure systems. Analytical method is developed for the evaluation of total maintenance cost. This analytical method helps to overcome, when compared with simulation methods, the computational time problem which often is a big problem for the maintenance optimization of systems with a large number of components. A numerical example is presented to show how the proposed grouping approach can be used for the maintenance planning of a complex structure system containing 12 components. A link with the coastal highway route E39 project is also discussed.

1 INTRODUCTION

Many industrial systems involve a high number of components where maintenance is necessary to maintain the performance throughout the system lifetime. Maintenance planning that results in the most economical way of grouping maintenance is essential, and failing to do so will expose the system owner to very large costs. This is a question that is of significant interest to the e.g. the Norwegian Road Administration who is currently planning for replacing many ferry crossings along the coastal E-39 in Norway with new strait crossing bridges and submergible tunnels (the so-called "Ferry free E-39 project").

In the last decade, grouping maintenance has developed and becomes an interesting approach in the maintenance optimization framework of multi-component systems (Dekker 1996). The idea of this approach is to take advantage of positive economic dependence to reduce the maintenance cost by jointly maintaining several components at the same time. The positive economic dependence among components implies that costs can be saved when several components are jointly maintained instead of separately (Nicolai & Dekker 2008).

Among many grouping maintenance strategies, the dynamic grouping developed in Wildeman et al. (1997) has drawn much attention, thanks to its ability of taking into account different online information which may occur over time (e.g. a varying deterioration of components, unexpected opportunities, changes of utilization factors of components). This approach has been developed to deal with the different maintenance challenges in many papers, e.g. condition-based model (Bouvard et al. 2011), predictive maintenance model (Van Horenbeek & Pintelon 2013), time limited opportunities (Do et al. 2013), availability constraint under limited access to repairmen (Do et al. 2015). However, these works only deal with series structure systems, in which the economic dependence among components is always positive. Indeed, the maintenance of a group of components can save the setup cost paid for preparation tasks of a maintenance action, and the downtime cost. Recently, to response to the fact that system structures are usually more complex and include redundancy (it could be a mixture of some basic configurations, e.g. series, parallel, series-parallel, k-out-of-n), the dynamic grouping is developed to take into consideration of the complexity due to the system structures (Vu et al. 2014, Vu et al. 2015, Nguyen et al. 2014). In these papers, under impacts of the complex structures, the economic dependence can be either positive or negative, depending on the considered group of components.

In order to facilitate the maintenance modeling, the most of existing works consider that the maintenance durations are negligible. This assumption may be unrealistic and limits the application of these models in real situations. To this end, the dynamic-grouping with taking into account the preventive maintenance durations is developed in Vu et al. (2014), but even so, the corrective maintenance duration (repair time) is still not investigated. Van Horenbeek & Pintelon (2013) has considered the repair time in their predictive grouping model, and used the Monte Carlo simulation to deal with the unpredictability of component failures. Unfortunately, the model is only valid for the systems with a series structure where the stoppage of any component leads to the shutdown of the entire system. The situation is likely to be more complicated in cases of the complex structures where the system can still operate (partially functioning) when some redundant components fail.

In this paper, a dynamic grouping strategy is developed for the complex structure systems with taking into account both preventive and corrective maintenance duration. However, under consideration of maintenance duration, the maintenance model becomes much more complex. In the present paper, an analytical method is developed to find the optimal maintenance planning where maintenance cost is the criterion. When compared with the Monte Carlo simulation, the proposed analytical method can reduce significantly the computational time, and may be applied to systems with a large number of components.

The rest of the paper is organized as follows. Section 2 is devoted to present the maintenance modeling and some general assumptions of this work. The dynamic grouping approach proposed in Vu et al. (2014) and Vu et al. (2015) for the complex structure systems is shortly described in Section 3. The development of the dynamic grouping approach to take into account the durations of repair actions is shown in Section 4. In Section 5, a numerical example is proposed to show how the developed grouping approach can be applied to the maintenance planning of a complex system of 12 components. The link between the research and the Ferry free E-39 project is discussed in Section 6. Finally, concluding remarks are made in Section 7.

2 MAINTENANCE MODELING

2.1 *General assumptions*

During the development of the proposed grouping strategy, the following assumptions are considered.

- Consider a multi-component system with a complex structure where the economical dependence among components could be both positive and negative (i.e. the maintenance cost of a group of components is not equal to the total maintenance cost of all components in the group).
- The system contains n repairable components which have two possible states: operational or failed.
- A time-based model is used to model the time to failure of components. $r_i(t)$ denotes the failure rate of component i, and $r_i(t) > 1$ ($i = 1, ..., n$).
- The logistic supports (e.g. repair teams, spare parts) are sufficient, available, and efficient to ensure that the repair at failures and the preventive replacement can be successfully and quickly carried out.
- The maintenance duration of a preventive maintenance action (denoted by ω_i^p) and a corrective maintenance action (denoted by ω_i^c) are constant and bigger than zero.

According to the complex structure, two kinds of components are here distinguished.

- Critical components: a shutdown of a critical component for whatever reason leads to a shutdown of the whole system.
- Non-critical components: the system can partially work when a non-critical component stops.

2.2 *Maintenance cost structure*

The cost to be paid for a maintenance action (preventive or corrective) of a component contains three following parts (Vu et al. 2014, Vu et al. 2015).

- A setup cost that can be composed by the cost of crew traveling and preparation costs (e.g. erecting a scaffolding or opening a machine).
- A specific cost that is related to the specific characteristics of the component such as spare part costs, specific tools and maintenance procedures.
- A downtime cost that has to be paid if the component is a critical one because the system is not functioning during the maintenance of the component. This downtime cost could be production loss costs, quality loss costs, restart costs, or machine damage costs, etc.

In general, the above costs may be changed overtime, and not the same for every component or every maintenance action. In this paper, in order to simplify the maintenance model, the setup cost and the downtime cost are assumed to be independent from the component characteristics. Moreover, all the costs are constant and depend on the nature of the maintenance action (preventive or corrective).

2.3 *Preventive Maintenance*

The aim of Preventive Maintenance (PM) is to reduce the probability of experiencing a failure of a component/system. In this paper, after a PM action, the maintained component is considered to be a new one. The cost that has to be paid for a PM action of a component i can be written as

$$C_i^p = S^p + c_i^p + \pi_i(t) \cdot c_d^p \cdot \omega_i^p \tag{1}$$

where S^p and c_i^p are the setup cost and the specific cost of a PM action respectively; $\pi_i(t)$ is an indicator function, which presents the criticality of the component i at time t, defined as

$$\pi_i(t) = \begin{cases} 1 & \text{if component } i \text{ is critical at time } t, \\ 0 & \text{otherwise;} \end{cases} \tag{2}$$

$c_d^p \cdot \omega_i^p$ is the downtime cost has to be paid during the replacement of the critical component i; and c_d^p is the mean downtime cost per unit time.

Note that the consideration of both the maintenance durations, and the complexity of the system structure leads to the possibility that a non-critical component can become critical one within the maintenance period of the other components. That is the reason why the indicator function π_i is presented as a function of time (Eq. 2).

2.4 *Corrective Maintenance*

During the system operation, if a component i fails, the component is then immediately repaired. After the minimal repair action, the repaired component is considered to be in the state that it has just before the failure. As with the preventive maintenance, when a Corrective Maintenance (CM) action is carried out on a component i, it requires a CM cost, denoted by C_i^c, which can be expressed as follows

$$C_i^c = S^c + c_i^c + \pi_i(t) \cdot c_d^c \cdot \omega_i^c \tag{3}$$

where S^c, c_i^c, and c_d^c are the setup cost, the specific cost, and the mean downtime cost per time unit related to a CM action of the component i respectively.

In the next section, the dynamic grouping strategy developed in (Vu et al. 2014, Vu et al. 2015) will be briefly presented. Under this strategy, the setup cost and the downtime cost paid for PM can be saved by simultaneously performing some PM actions. Note that, in these papers, the grouping of CM actions is not allowed due to the limitations of the logistic support.

3 DYNAMIC GROUPING APPROACH FOR COMPLEX STRUCTURE SYSTEMS

In a complex structure system, the maintenance grouping of a group of components can have both negative and positive impacts on the system function depending on the criticality of the group and its components. Thus, the consideration of the criticality in grouping optimization is important, and can help to improve the grouping performance (Vu et al. 2014, Vu et al. 2015).

The dynamic grouping model, developed for complex structure systems, contains the four following phases (Fig. 1).

- *Phase 1: System analysis.* In this phase, the solution for determining the criticality of components (π_i) or groups of components (π_{G^k}) at time t with respect to a specific system function is developed. For this purpose, the reliability block diagram is used in this paper.
- *Phase 2: Individual optimization.* This phase is designed to determine the long-term maintenance plan for each component separately by minimizing its long-run expected maintenance cost rate. In this phase, the economic dependency among components is not considered, and the criticality of components is considered to be fixed.
- *Phase 3: Grouping optimization.* A specific short-term horizon, and all the PM actions of components within the horizon are firstly identified based on the individual maintenance plan obtained from phase 2. These PM actions are then grouped to take advantage of the positive economic dependence among components. The grouping solution is found by maximizing the total economic profit within the considered horizon.
- *Phase 4: Update of the grouping solution.* The grouping solution obtained in phase 3 needs to be updated in the two following cases: grouping planning for a new short-term horizon (rolling horizon); occurrences of dynamic contexts such as maintenance opportunities, changes in production planning, changes in operation conditions (Vu et al. 2014).

Above paragraph presents the four phases of the dynamic grouping approach developed for

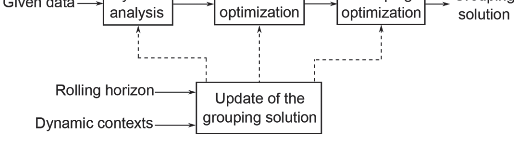


Figure 1. Dynamic grouping for complex structure systems.

complex structure systems. In the next section, we will describe how this approach can be developed to take into account both the complexity of the system structure, and the repair time.

4 DYNAMIC GROUPING APPROACH WITH TAKING INTO ACCOUNT THE REPAIR TIME

The consideration of the repair time does not lead to the complete change in the above grouping approach. Indeed, to take into account the repair time, only phase 2 and phase 3 are developed and presented in this section. The other phases of the approach remain unchanged and can be found in more detail in Vu et al. (2014) and Vu et al. (2015).

4.1 *Individual optimization*

As mentioned above, in this phase, the long-term maintenance plan is separately determined for each component. To do this, age-based replacement strategy (Barlow & Hunter 1960) is usually chosen thanks to its high performance at component level. When the repair time is considered, the replacement decisions based on the component's age will face many difficulties in maintenance modeling and maintenance optimization due to the unpredictability of the failures. For this reason, the calendar-based replacement strategy is used in this paper (Fig. 2).

According to the calendar-based replacement strategy, the component i is replaced at fixed-time intervals T_i and minimally repaired at its failures.

The long-term expected maintenance cost rate of the component i is calculated based on the renewal theory as follows

$$CR_i(T_i) = \frac{C_i^p + C_i^c \cdot \Lambda_i(0, T_i)}{T_i + \omega_i^p} \tag{4}$$

where $\Lambda_i(0,T_i)$ is the mean number of failures of component i on $(0,T_i]$. Under the minimal repair, $\Lambda_i(0,T_i)$ is equal to

$$\Lambda_i(0, T_i) = \int_{\nu_i(0)}^{\nu_i(T_i)} r_i(t)dt = \int_0^{x_i} r_i(t)dt \tag{5}$$

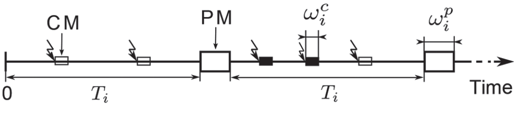


Figure 2. Calendar-based replacement strategy.

where $\nu_i(t)$ is the age (total operating time) of component i at time t. We have $\nu_i(0) = 0$, and $\nu_i(T_i) = x_i$.

Equation 4 can be rewritten as follows

$$CR_i(x_i) = \frac{C_i^p + C_i^c \cdot \int_0^{x_i} r_i(t)dt}{x_i + \omega_i^c \cdot \int_0^{x_i} r_i(t)dt + \omega_i^p} \tag{6}$$

The optimal replacement interval of the component i (denoted by T_i^*) is then determined by minimizing its long-term expected maintenance cost rate.

$$x_i^* = \arg\min_{x_i} CR_i(x_i) \tag{7}$$

and

$$T_i^* = x_i^* + \omega_i^c \cdot \int_0^{x_i^*} r_i(t)dt \tag{8}$$

The corresponding minimal maintenance cost rate is

$$CR_i^* = \frac{C_i^p + C_i^c \cdot \int_0^{x_i^*} r_i(t)dt}{T_i^* + \omega_i^p} \tag{9}$$

4.2 *Grouping optimization*

Individual maintenance plan. Based on the replacement intervals obtained in the previous phase, in this phase, the individual maintenance dates of components in a specific short-term horizon are determined.

In details, consider a planning horizon PH between a and b. The first PM activity of the component i in PH, denoted by t_{i^1}, is then determined as follows

$$t_{i^1} = T_i^* - d_i(a) + a \tag{10}$$

where $d_i(a)$ is the time between a and the last PM activity of the component i before a.

The other PM activities of the component i in PH can be determined as

$$t_{i^j} = t_{i^{j-1}} + \omega_i^p + T_i^* \text{ if } j > 1 \text{ and } t_{i^j} \le b \tag{11}$$

where t_{i^j} denotes the jth PM activity of the component i in PH.

Grouping solution. A partition of $\{1,...,N\}$ is a collection of m mutually exclusive groups $G^1,...,G^m$ which cover all N PM activities in PH.

$$G^l \cap G^k = \varnothing, \forall l \neq k \tag{12}$$

and

$$G^1 \cup G^2 \cup ... \cup G^m = \{1, ..., N\} \tag{13}$$

A grouping solution, denoted by *GS*, is a partition of $\{1, ..., N\}$ such that all PM activities in each group are jointly executed at the same time.

Evaluation of the grouping performance. The cost saving of the grouping maintenance compared to the individual maintenance is used as the only one criterion to evaluate the performance of a grouping solution. The cost saving of a grouping solution *GS* is

$$\begin{aligned} CS(GS) &= \sum_{k=1}^{m} CS(G^k) \\ &= \sum_{k=1}^{m} \Big(U_{G^k} - \Delta H_{G^k}^1 - \Delta H_{G^k}^2 - \Delta H_{G^k}^3 \end{aligned} \tag{14}$$

where

- $CS(G^k)$ is the cost saving when all components of group G^k are jointly maintained.
- U_{G^k} is the saving of the PM setup cost when all n_k components of group G^k are grouped. We consider that only one setup cost has to be paid when a group of components are maintained together.

$$U_{G^k} = (n_k - 1) \cdot S^p \tag{15}$$

- $\Delta H_{G^k}^1$ is the penalty costs due to the changes of maintenance dates from the optimal individual ones t_{i^j} to the group execution date t_{G^k}.

$$\begin{aligned} \Delta H_{G^k}^1 = \sum_{i^j \in G^k} \{ & C_i^c \cdot [\Lambda_i(0, t_{G^k}) \\ & - \Lambda_i(0, t_{i^j})] - CR_i^* \cdot (t_{G^k} - t_{i^j})\} \end{aligned} \tag{16}$$

- $\Delta H_{G^k}^2$ is the cost related to the change in total planned downtime of the system.

$$\Delta H_{G^k}^2 = c_d^p \cdot [\pi_{G^k} \cdot \omega_{G^k} - \sum_{i^j \in G^k} \pi_i \cdot \omega_i^p] \tag{17}$$

where ω_{G^k} is the maintenance duration of group G^k. When the number of repair teams is sufficient, we have $\omega_{G^k} = \max_{i^j \in G^k} \omega_i^p$.

- $\Delta H_{G^k}^3$ is the cost related to the change in total unplanned downtime of the system.

$$\begin{aligned} \Delta H_{G^k}^3 = c_d^c \cdot [& (1 - \pi_{G^k}) \cdot \sum_{q \in Q_{G^k}} \omega_q^p \cdot \Lambda_q(t_{G^k}, t_{G^k} + \omega_{G^k}) \\ & - \sum_{i^j \in G^k} (1 - \pi_i) \cdot \sum_{l \in L_i} \omega_i^p \cdot \Lambda_l(t_{i^j}, t_{i^j} + \omega_i^p)] \end{aligned} \tag{18}$$

where Q_{G^k} and L_i are the sets of non-critical components which become critical ones during the maintenance of group G^k and component *i* respectively.

The following points need to be noted during the calculation of $CS(GS)$.

- The optimal value of t_{G^k} is determined as

$$\begin{aligned} t_{G^k}^* = \underset{t_{G^k}}{\arg\min} \Big\{ & \Delta H_{G^k}^1 + c_d^c \cdot (1 - \pi_{G^k}) \\ & \cdot \sum_{q \in Q_{G^k}} \Lambda_q(t_{G^k}, t_{G^k} + \omega_{G^k}) \Big\} \end{aligned} \tag{19}$$

- The calculation of $CS(GS)$ is mostly based on the mean number of failures of components in different intervals; in other words, it is based on the age of components at different instants. Unfortunately, the determination of the age of components over time becomes a real challenge when the repair time is taken into account. To overcome this problem, numerical simulation methods have been widely used in the literature. The use of simulation methods leads to a high computational time and difficultiesin the case of systems with a large component number. For this reason, in the next paragraph, an analytical method is developed to determine the age of components.

Age analysis method. This analytical method is developed to calculate the age of components at instants $t_{i^j}, t_{i^j} + \omega_i^p$ in the individual maintenance plan, and at instants $t_{G^k}, t_{G^k} + \omega_{G^k}$ in the grouped maintenance plan. The following steps of the proposed method are separately and repeatedly applied to the two above plans.

- *Step 1: Determination of a partition of PH.* A partition of *PH* containing *V* sub-intervals (PH_1, PH_2, ..., PH_V) is determined so that a component can have only one state either under PM or not in each sub-interval.
- *Step 2: Analyze the system structure in a sub-interval* $PH_v = [a_v, b_v]$. This analysis is done in order to determine the following sets of components.
  G_v is the set of components which are preventively maintained in PH_v.
  A_v is the set of components which are not functioning due to the PM of G_v.
  B_v is the set of components which are not in G_v and A_v.
  For a component *i* in B_v, a set of components C_v^i is determined such that the PM of any component in C_v^i leads the component *i* to stop.
- *Step 3: Calculate the age of components at* b_v.

$$v_i(b_v) = 0 \text{ if } i \in G_v.$$

$$v_i(b_v) = v_i(a_v) \text{ if } i \in A_v.$$

For a component i in B_v, $v_i(b_v)$ is determined by solving the following equation system.

$$b_v = a_v + [\nu_i(b_v) - \nu_i(a_v)] + \omega_i^c \cdot \Lambda_i(a_v, b_v) + \sum_{j \in C_v^i} \omega_j^c \cdot \Lambda_j(a_v, b_v), \forall i \in B_v \quad (20)$$

- *Step 4: Return to the step 2 for all sub-intervals from $v = 1$ to V.*

Optimal grouping solution. Based on the above calculation of the grouping performance, we can compare different grouping solutions and determine the optimal one.

$$GS^* = \arg\max_{GS} CS(GS) \quad (21)$$

The finding of the optimal grouping solution is a NP-complete problem because the number of possible grouping solutions increases very quickly with the increasing of the number of PM activities in PH. Consequently, in this paper, the Genetic Algorithm citeHolland1975 is used to search the optimal grouping solution.

5 NUMERICAL EXAMPLE

In order to show how our dynamic grouping approach can be applied to the maintenance planning with taking into account the repair time, a specific system and its data are randomly created. The system contains 12 components with the reliability block diagram shown in Figure 3.

The failure behaviors of the components is described by the Weibull distribution with scale parameter λ_i and shape parameter $\beta_i > 1$. The failure rate of component i is

$$r_i(t) = \frac{\beta_i}{\lambda_i}\left(\frac{t}{\lambda_i}\right)^{\beta_i - 1} \quad (22)$$

The data of components are given in Table 1.

The other costs are $S^p = S^c = 15$, $c_d^p = 80$, and $c_d^c = 120$.

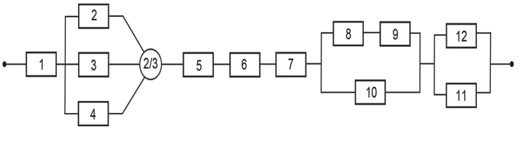


Figure 3. Reliability block diagram.

Table 1. Given data of components.

Components	λ_i	β_i	c_i^p	ω_i^p	c_i^c	ω_i^c
1	253	2.45	155	1	22	0.35
2	205	2.35	225	3	36	0.81
3	117	1.87	785	1	66	0.29
4	190	2.00	245	2	28	0.54
5	119	1.65	375	1	92	0.36
6	284	2.50	300	2	76	0.44
7	297	3.05	345	1	55	0.27
8	108	1.55	555	1	102	0.31
9	200	1.95	190	3	45	0.39
10	125	1.85	350	2	44	0.32
11	189	2.75	460	2	30	0.78
12	275	1.85	130	1	24	0.34

Table 2. Given data of components.

Components	π_i	C_i^p	C_i^c	T_i^*	CR_i^*
1	1	250	79	348.75	1.21
2	0	240	51	352.19	1.18
3	0	800	81	434.42	3.97
4	0	260	43	471.75	1.10
5	1	470	150	310.39	3.83
6	1	475	143	390.06	2.02
7	1	440	102	379.20	1.72
8	0	570	117	444.85	3.61
9	0	205	60	385.20	1.08
10	0	365	59	367.79	2.15
11	0	475	45	371.72	2.02
12	0	145	39	612.40	0.52

5.1 *Individual optimization*

In this phase, the block replacement strategy is used for the maintenance planning at component level. The intermediate results and the optimal maintenance frequencies of components are presented in Table 2.

The long-term maintenance cost rate of the system when all PM activities are individually performed is

$$CR_{sys}^{\overline{G}} = \sum_{i=1}^{12} CR_i^* = 24.41 \quad (23)$$

5.2 *Grouping optimization*

The new system is put into operation at time zero; therefore, we have $\nu_i(0) = 0$, $d_i(0) = 0$, and $t_{i^1} = T_i^*$ for all $i = 1,\ldots,n$. Consider a finite interval $PH = [a,b]$ in which each component is preventively maintained once. We have $a = 0$, and $b = \max_{i=1}^{12} t_{i^1} = 612.40$.

Table 3. Optimal grouping solution.

Group	Components	π_{Gk}	t^*_{Gk}	ω_{Gk}	$CS(G_k)$
1	1, 4, 5, 6, 7, 10	1	367.08	2	326.29
2	2, 9, 11	0	373.38	3	62.77
3	3, 8, 12	0	458.74	1	95.98

The total maintenance costs of the system in *PH* when all PM activities are individually performed, denoted $TC^{\overline{G}}$, is calculated as follows

$$TC^{\overline{G}} = CR^{\overline{G}}_{sys} \cdot (b-a) = 14948.68 \tag{24}$$

In order to find the optimal grouping solution, Genetic Algorithm is used with the following parameters: crossover probability = 0.8, mutation probability = 0.02, population size = 60, number of generations = 500. The program is implemented by the use of Matlab R2015a on a DELL computer (Intel core i7, 2.6 Ghz, 16 GB RAM). The computational time is approximately 40 minutes. The optimal grouping solution obtained by the program is reported in Table 3.

The total economic profit of the optimal grouping solution is

$$CS(GS) = \sum_{k=1}^{3} CS(G^k) = 485.03 \tag{25}$$

The total maintenance costs of the system in *PH* when the PM activities are grouped, denoted TC^G, is calculated as follows

$$TC^G = TC^{\overline{G}} - CS(GS) = 14463.65 \tag{26}$$

From the obtained results, we can conclude that the maintenance grouping helps to reduce the total maintenance costs of the system in *PH* from 14948.68 to 14463.65. The reduction is equal to 3.24% of the total maintenance costs of the system in *PH* when all PM activities are individually performed.

6 THE LINK BETWEEN THE PROPOSED RESEARCH AND THE COASTAL HIGHWAY ROUTE E39 PROJECT

In this section, we will shortly discuss about the link between the presented research and the E39 project. Norway's coastal highway E39 is part of the European trunk road system. The route runs from Kristiansand in the south to Trondheim in central Norway, a distance of almost 1100 km. There are eight wide and deep fjords along the route. The fjord crossings will require massive investments and huge bridges than previously installed in Norway. Figure 4 describes the submerged floating tunnel bridge concept for crossing the Sognefjord, which is both deep and wide, and is considered challenging to cross. The more information about the E39 project can be found on http://www.vegvesen.no/Vegprosjekter/ferjefriE39/.

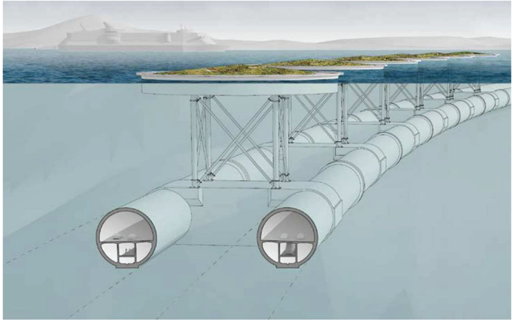


Figure 4. Submerged floating tunnel bridge concept.

The construction and operation of such superstructure system face with many technological challenges included maintenance planning problems. This research is then partially funded by the E39 project, and motivated by the following maintenance planning problems.

- The bridge is a complex system containing a large number of components which are interdependent. More over, the bridge availability is highlighted due to the important impacts of its closures on the traffic, the environment, and the people safety. For this purpose, the developed dynamic grouping approach based on the analytical method can help to improve the bridge availability, and prevent the maintenance optimization from the computational time problem.
- The taking into account of the repair time in the maintenance modeling is necessary when the estimated repair time is considerable.

Given the above efforts, many challenges related to the maintenance planning of the superstructure system still exist such as the uncertainties of the data and their impacts on the maintenance performance, the imperfect preventive maintenance, the component's maintainability, etc. These challenges will be our objectives in the future research.

7 CONCLUSION

In this paper, a dynamic grouping approach is developed for the maintenance planning of the complex structure systems with consideration of both preventive and corrective maintenance duration. To overcome the problem of computational time, an age analytical method is proposed, given the complexity of the maintenance optimization problem when considering maintenance duration. The numerical example describes how the proposed grouping approach can be applied to the maintenance planning of a system of 12 components. The obtained results confirm the advantage of the proposed grouping strategy and show that the computational time is reasonable. Finally, it should be noted that this research is motivated by the real problems that we have to face to within the scope of the E39 project. However, the application of the proposed approach is still limited, and needs to be investigated in the future research.

ACKNOWLEDGMENTS

This research was supported by the Ferry free E39 project and the Department of Production and Quality Engineering, Norwegian University of Science and Technology.

REFERENCES

Barlow, R. & L. Hunter (1960). Optimum preventive maintenance policies. *Operations research 8*(1), 90–100.

Bouvard, K., S. Artus, C. Berenguer, & V. Cocquempot (2011). Condition-based dynamic maintenance operations planning & grouping. application to commercial heavy vehicles. *Reliability Engineering & System Safety 96*(6), 601–610.

Dekker, R. (1996). Applications of maintenance optimization models: a review and analysis. *Reliability Engineering & System Safety 51*(3), 229–240.

Do, P., A. Barros, K. Berenguer, C. Bouvard, & F. Brissaud (2013). Dynamic grouping maintenance with time limited opportunities. *Reliability Engineering & System Safety 120*, 51–59.

Do, P., H.C. Vu, A. Barros, & C. Berenguer (2015). Maintenance grouping for multi-component systems with availability constraints and limited maintenance teams. *Reliability Engineering* & *System Safety 142*, 56–67.

Holland, J. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Ann Arbor, MI, United States: University of Michigan Press.

Nguyen, K.A., P. Do, & A. Grall (2014). Condition-based maintenance for multi-component systems using importance measure and predictive information. *International Journal of Systems Science: Operations & Logistics 1*(4), 228–245.

Nicolai, R. & R. Dekker (2008). Optimal maintenance of multi-component systems: a review. In *Complex system maintenance handbook*, pp. 263–286. London, United Kingdom: Springer.

Van Horenbeek, A. & L. Pintelon (2013). A dynamic predictive maintenance policy for complex multi-component systems. *Reliability Engineering & System Safety 120*, 39–50.

Vu, H.C., P. Do, A. Barros, & C. Be´renguer (2014). Maintenance grouping strategy for multi-component systems with dynamic contexts. *Reliability Engineering & System Safety 132*, 233–249.

Vu, H.C., P. Do, A. Barros, & C. Be´renguer (2015). Maintenance planning and dynamic grouping for multi-component systems with positive and negative economic dependencies. *IMA Journal of Management Mathematics 26*(2), 145–170.

Wildeman, R.E., R. Dekker, & A. Smit (1997). A dynamic policy for grouping maintenance activities. *European Journal of Operational Research 99*(3), 530–551.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

A parametric predictive maintenance decision framework considering the system health prognosis accuracy

K.T. Huynh & A. Grall
ICD, ROSAS, LM2S, Université de technologie de Troyes, UMR 6281, CNRS, Troyes, France

C. Bérenguer
Université Grenoble Alpes, GIPSA-lab, Grenoble, France
CNRS, GIPSA-lab, Grenoble, France

ABSTRACT: Nowadays, the health prognosis is popularly recognized as a significant lever to improve the maintenance performance of modern industrial systems. Nevertheless, how to efficiently exploit prognostic information for maintenance decision-making support is still a very open and challenging question. In this paper, we attempt at contributing to the answer by developing a new parametric predictive maintenance decision framework considering improving health prognosis accuracy. The study is based on a single-unit deteriorating system subject to a stochastic degradation process, and to maintenance actions such as inspection and replacement. Within the new framework, the system health prognosis accuracy is used as a condition index to decide whether or not carrying out an intervention on the system. The associated mathematical cost model is also developed and optimized on the basis of the semi-regenerative theory, and is compared to a more classical benchmark framework. Numerical experiments emphasize the performance of the proposed framework, and confirm the interest of introducing the system health prognosis accuracy in maintenance decision-making.

1 INTRODUCTION

Maintenance involves a wide range of activities, such as inspection, testing, repair, replacement, etc., in order to extend equipment life, improve equipment availability, as well as to retain equipment in its proper condition. Since these activities are costly, maintenance strategies are needed to organize and schedule them in a logical and economical manner. In literature, maintenance strategies have evolved from the naive *breakdown maintenance*, to the blind *time-based maintenance*, and lately towards the sophisticated *Condition-Based Maintenance* (CBM) (Jardine et al. 2006). Nowadays, with the development of *prognostics and health management* technologies, a new trend of maintenance strategies called *predictive maintenance* has been recently emerged (Shin and Jun 2015). Theoretically, the predictive maintenance belongs to the class of CBM strategies; but unlike the conventional ones, it bases the maintenance decisions on the system health prognostic information instead of diagnostic information. Such a strategy anticipates more efficiently the system failures, allows timely interventions on the system, and hence promises better performance than conventional CBM strategies. However, various recent works on the comparison between predictive maintenance and conventional CBM strategies (see e.g., (Khoury et al. 2013, Huynh et al. 2014)) have shown the contrary: the former is not really more profitable than the latter. This leads us to think that the current predictive maintenance frameworks (used e.g., in the above references) are not suitable to efficiently exploit the prognostic information for maintenance decision-making. Faced to this issue, the present paper attempts at developing a new predictive maintenance decision framework which can overcome the drawbacks of the more classical ones.

More precisely, the study is based on a single-unit deteriorating system subject to maintenance actions such as inspection and replacement. The system degradation evolution is described by a *homogeneous Gamma process*, and the system fails when the degradation level exceeds a fixed failure threshold. Such a degradation and failure model allows us to compute and analyze some prognostic condition indices characterizing the future health state of the system. The *standard deviation* and the *mean value* of the system *Residual Useful Life* (RUL) are the prognostic condition indices of interest, and we investigate how these indices can be exploited to make pertinent maintenance decisions. In fact, adopting the well-known parametric structure-based decision rule (Bérenguer 2008), the former is used to decide whether or not carrying out an intervention (i.e., inspection or replacement) on the system, while the latter is used to determine proper replacement times. A framework with such maintenance decisions is known as parametric predictive maintenance decision framework considering improving health prognosis accuracy. To quantify the performance of the new maintenance framework, we develop and optimize its mathematical cost model on the basis of the *long-run expected maintenance rate* and the *semi-regenerative theory*. The comparisons with a more classical benchmark framework under various configurations of maintenance costs and system characteristics allow us to emphasize the performance of the proposed framework, and justify the interest of introducing the system health prognostic information in maintenance decision-making.

The remainder of this paper is organized as follows. Section 2 is devoted to modeling the system and to computing associated condition indices. Section 3 deals with the detailed description and theoretical analyses of the considered predictive maintenance decision frameworks including the maintenance assumptions, the most well-known maintenance framework, and the new framework considering the system health prognosis accuracy. In Section 4, the maintenance cost models of these frameworks are developed and optimized. The assessment and discussions on the performance of the new maintenance framework are carried out in Section 5. Finally, the paper is end with some conclusions and perspectives.

2 SYSTEM MODELING, CONDITION INDICES

2.1 *System modeling*

Consider a single-unit deteriorating system consisting of 1 component or 1 group of associated components (from the maintenance viewpoint). The system suffers an underlying degradation process which can cause random failures. Such a process may be a physical deterioration process such as cumulative wear, crack growth, erosion, corrosion, fatigue, etc. (Grall et al. 2002); or it may be an artificial process describing the phenomenon that the system health state or its performance worsen with usage and age (Xu et al. 2008). For such system, it is recommended by Singpurwalla (1995) to base the degradation modeling on time-dependent stochastic processes. The notion of process helps us to describe more finely the behavior of the system, and hence allowing, e.g., a more accurate prediction of its RUL (Si et al. 2011). By this way, let X_t be a scalar random variable representing the accumulated degradation of the system at time $t \geq 0$; without any maintenance operation, $\{X_t\}_{t\geq 0}$ is an increasing stochastic process with $X_0 = 0$ (i.e., the system is initially new). Moreover, assuming the degradation increment between 2 times t and s ($t \leq s$), $X_s - X_t$, is s-independent of degradation levels before t, one can apply any monotone stochastic process from the Lévy family (Abdel-Hameed 2014) to model the evolution of the system degradation. In the present paper, the well-known homogeneous Gamma process with shape parameter α and scale parameter β is used. The choice of such a process for degradation modeling has been justified by diverse practical applications (e.g., corrosion damage mechanism (Kallen and van Noortwijk 2005), carbon-film resistors degradation (Wang 2009), SiC MOSFET threshold voltage degradation (Santini et al. 2014), fatigue crack growth (Bousquet et al. 2015), actuator performance loss (Langeron et al. 2015), etc.), and it is considered appropriate by experts (Blain et al. 2007). Moreover, using Gamma process can make the mathematical formulation feasible. As such, for $t \leq s$, the degradation increment $X_s - X_t$ follows a Gamma law with *probability density function* (pdf)

$$f_{\alpha\cdot(s-t),\beta}(x) = \frac{\beta^{\alpha\cdot(s-t)} x^{\alpha\cdot(s-t)-1} e^{-\beta x}}{\Gamma(\alpha\cdot(s-t))} \cdot 1_{\{x\geq 0\}}, \tag{1}$$

and *cumulative distribution function* (cdf)

$$F_{\alpha\cdot(s-t),\beta}(x) = \frac{\Gamma(\alpha\cdot(s-t),\beta x)}{\Gamma(\alpha\cdot(s-t))}, \tag{2}$$

where $\Gamma(\alpha) = \int_{\mathbb{R}_+} z^{\alpha-1} e^{-z} dz$ and $\Gamma(\alpha, x) = \int_0^x z^{\alpha-1} e^{-z} dz$ denote the complete and *lower* incomplete Gamma functions respectively, and $1_{\{\cdot\}}$ denotes the indicator function which equals 1 if the argument is true and 0 otherwise. The couple of parameters (α, β) allows to model various degradation behaviors from almost-deterministic to very-chaotic, and its average degradation rate and the associated variance are $m = \alpha / \beta$ and $\sigma^2 = \alpha / \beta^2$ respectively. When degradation data are available, these parameters can be estimated by classical statistical methods such as maximum likelihood estimation, moments estimation, etc. (Van Noortwijk 2009).

Associated with the degradation process, we use a threshold-type model to define the system failure. For economic (e.g., poor products quality, high consumption of raw material) or safety reasons (e.g., high risk of hazardous breakdowns), a system is usually declared as failed when it is no longer able to fulfill its mission in an acceptable condition even if it is still functioning. A high system degradation level is thus unacceptable. According to this view, we consider that the system fails as soon as its degradation level exceeds a critical prefixed threshold L. The system failure time τ_L is thus expressed as

$$\tau_L = \inf\left\{t \in \mathbb{R}_+ \mid X_t \geq L\right\}. \tag{3}$$

2.2 *Condition indices*

Condition indices are indices characterizing the health state of a system, based on which one can make a maintenance decision (Huynh et al. 2014). Such indices may be the result of the real-time diagnosis of impending failures (i.e., *diagnostic condition indices*) or of the prognosis of future system health (i.e., *prognostic condition indices*). For our considered system, the degradation level returned by an inspection at time τ_i, X_{τ_i}, is a diagnostic condition index,because it can define the system health state at the current time τ_i. Note that the diagnosis in reality is not a simple task and may require sophisticated techniques (Travé-Massuy`es 2014). However, since the diagnosis is beyond the scope of the paper, we simply assume that the diagnosis is attached to inspection operations which can perfectly reveal the system degradation level. Given the diagnostic information and the degradation and failure model, one can predict prognostic condition indices. In the literature, RUL, defined as the length from the current time to the end of the system useful life, is a well-known prognostic index because it can provide an idea about how long a system at a particular age will still survive. The concept of RUL has been widely investigated by many works in the *Prognostics and Health Management* research area (Liao and Kottig 2014). In this paper, a so-called *conditional* RUL is considered, and its mathematical expression at time τ_i given the degradation level X_{τ_i} is defined as (Banjevic 2009)

$$\rho\left(\tau_i \mid X_{\tau_i}\right) = \left(\tau_L - \tau_i \mid X_{\tau_i}\right) \cdot 1_{\left\{X_{\tau_i} < L\right\}}, \tag{4}$$

where τ_L is the system failure time given from (3). Obviously, $\rho(\tau_i|X_{\tau_i})$ is a random variable, it can be then characterized by the *mean value* and the *standard deviation*. Indeed, the former is usually used to locate the distribution of $\rho(\tau_i|X_{\tau_i})$, while the latter is adopted to describe the variability existing in $\rho(\tau_i|X_{\tau_i})$. The mean value of $\rho(\tau_i|X_{\tau_i})$ is known as conditional *mean residual lifetime* (MRL) of the system (Huynh et al. 2014). At time τ_i, given $X_{\tau_i} = x$, the conditional MRL and the associated standard deviation are computed by

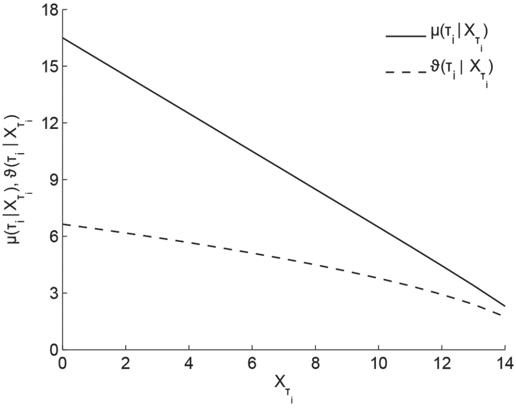


Figure 1. $\mu(\tau_i|x)$ and $\vartheta(\tau_i|x)$ with respect to x.

$$\mu\left(\tau_i \mid x\right) = \int_{\mathbb{R}_+} F_{\alpha u,\beta}\left(L - x\right) du \tag{5}$$

$$\vartheta\left(\tau_i \mid x\right) = \sqrt{\int_{\mathbb{R}_+} 2u F_{\alpha u,\beta}\left(L - x\right) du} - \mu\left(\tau_i \mid x\right) \tag{6}$$

where $F_{\alpha u \beta'}$ (·) is derived from (2). Applying (5) and (6) to the system characterized by $L = 15$ and $\alpha = \beta = 1/3$, we obtain the shape of $\mu\left(\tau_i \mid x\right)$ and $\vartheta\left(\tau_i \mid x\right)$ as in Figure 1. It is easy to prove that $\mu\left(\tau_i \mid x\right)$ and $\vartheta\left(\tau_i \mid x\right)$ are non-increasing functions in x. The RUL prognosis becomes then more precise for higher value of x. Moreover, $\mu\left(\tau_i \mid x\right)$ and $\vartheta\left(\tau_i \mid x\right)$ depend only on x. Thus, given the degradation and failure model as in Section 2.1, the considered diagnostic and prognostic condition indices are equivalent. Even so, each of them has its own meaning in maintenance decision-making support. A proper maintenance framework should take care to this point.

3 MAINTENANCE DECISION FRAMEWORKS

We propose in this section a new predictive maintenance decision framework considering improving health prognosis accuracy. The framework is relied on the parametric structure-based decision rule described in (Bérenguer 2008). To better understand the originality of the proposed framework, we introduce at first assumptions on the maintained system, then we analyze the maintenance decision framework most well-known in the literature through a representative strategy. Our maintenance decision framework is introduced next. To illustrate how to use this new framework, a predictive maintenance strategy is also derived. The proposed illustrations in this section are based on the system defined by $\alpha = \beta = 1/3$ and $L = 15$, and on the optimal configurationof the considered strategies when the set of maintenance costs $C_i = 5$, $C_p = 50$, $C_c = 100$ and $C_d = 25$ is applied.

3.1 *Maintenance assumptions*

Consider the system presented in Section 2.1, we assume that its degradation level is hidden, and its failure state is non-self-announcing. Inspection activities are then necessary to reveal the system state. The notion of inspection here is not simply the data collection, but also the feature extraction from the collected data, the construction of degradation indicators, and perhaps more. In other words, this activity includes all the tasks before the *Maintenance Decision Making* task in a predictive maintenance program. Such an inspection operation is itself costly, and takes time; but, compared to the life cycle of a system, the time for an inspection is negligible. Thus, we assume that each inspection operation is instantaneous, perfect, non-destructive, and incurs a cost $C_i > 0$.

Two maintenance actions are available: a *Preventive Replacement* (PR) with cost $C_p > C_i$, and a *Corrective Replacement* (CR) with cost C_c. Since maintenance actions are the true physical replacement such that the system is as-good-as-new rather than repairs, they take negligible times and incur fixed costs irrespective of the degradation level of the system. Even though both the PR and CR operations put the system back in the as-good-as-new state, they are not necessarily identical in practice because the CR is unplanned and performed on a more deteriorated system, moreover the cost C_c can comprise different costs associated to failure like damage to the environment. It is thus likely to be more complex and more expensive (i.e. $C_c > C_p$). Futhermore, a replacement, whether preventive or corrective, can only be instantaneously performed at predetermined times (i.e., inspection time or scheduled replacement time). Therefore, there exists a system downtime after failure, and an additional cost is incurred from the failure time until the next replacement time at a cost rate C_d.

3.2 *Classical framework and representative strategy*

The classical predictive maintenance decision framework considers that a replacement is always attached to an inspection and can only be carried out at a certain inspection time τ_i only. Most maintenance strategies in the literature belong to thisframework. For example, in (Huynh et al. 2011), a periodic inspection schedule is implemented and a replacement is performed whenever the system degradation level at an inspection time reaches a threshold. In (Ponchet et al. 2010), the same degradation index is used for a replacement decision, but the inspection schedule is non-periodic. In (Huynh et al. 2014) and (Huynh et al. 2015), the replacement decisions are made according to the system conditional MRL and the system conditional reliability respectively. Within this classical framework, when the system is multi-unit or subject to multi-failure mode (e.g., shock and degradation), using prognostic condition indices (e.g., the system MRL, the system reliability, the system RUL standard deviation, etc.) for maintenance decision-making is still more profitable than using diagnostic ones (e.g., the system degradation level) thanks to their "overarching" property (Huynh et al. 2015). But, when the system is single-unit and its failure is due to the degradation solely, they always lead to the same maintenance performance (Huynh et al. 2014). This means that the classical predictive maintenance decision framework does not allow efficiently exploiting the prognostic condition indices. In the following, we will learn about the reasons through a representative maintenance strategy of this classical framework.

To facilitate the comprehension, a periodic inspection schedule is assumed and the replacement decisions are made according to the detected system degradation level at an inspection time. Let define a *renewal cycle* as the time interval between two successive replacement operations, the periodic inspection and degradation-based replacement strategy over a renewal cycle is stated as follows. The system is regularly inspected with period $\delta > 0$. At a scheduled inspection date $\tau_i = i \cdot \delta$, $i = 1,2,\ldots$, a CR of the system is carried out if it fails (i.e., $X_{\tau_i} \geq L$). But, if the system is still running, a decision based on the degradation level X_{τ_i} is made. If $\zeta \leq X_{\tau_i} < L$, the running system is considered too degraded, and a PR should be carried out at τ_i. Otherwise, nothing is done at τ_i, and the maintenance decision is postponed until the next inspection time at $\tau_{i+1} = \tau_i + \delta$. The inspection period δ and the PR threshold ζ are the 2 decision variables of this strategy, so we call it (δ, ζ) *strategy*. Applying the (δ, ζ) strategy to the system defined at the beginning of Section 3, one obtain the behavior the maintained system as in Figure 2. The optimal decision variables are $\delta_{opt} = 4.6$ and $\zeta_{opt} = 9.1478$ (the cost model to derive these optimal variables will be dealt with in Section 4). In Figure 2, the degradation evolution of the maintained system and the associated inspection and replacement operations are shown on the top, and the evolution of the conditional RUL standard deviation at inspection times τ_i andreplacement times τ_r are represented in the bottom.

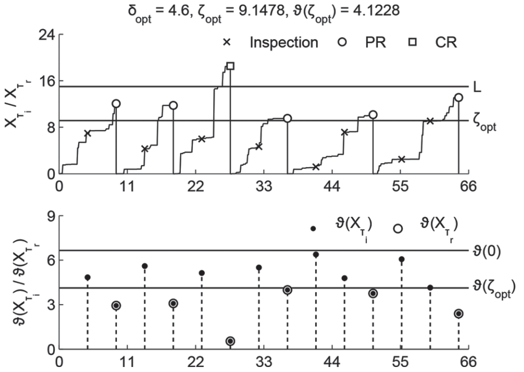


Figure 2. Illustration of the (δ,ζ) strategy.

Under such a maintenance strategy, the condition index (e.g., the threshold ζ of the (δ,ζ) strategy) has a double role. On one hand, it, together with the inspection period δ, decides the number of inspections carried out over a renewal cycle, and on the other hand, it is adopted to decide whether or not to trigger a PR. This is the biggest weakness of the classical maintenance decision framework, because a single condition index may not have all the required properties to make efficient decision in order to, at the same time, avoid inopportune inspection operations and properly prolong the system useful lifetime. Indeed, to avoid a large number of inspection operations, the (δ,ζ) strategy lowers the value of ζ; however a lower ζ will shorten the system useful lifetime because of early PR operations. On the contrary, a high value of ζ can lengthen the system useful lifetime, but more cost could be paid for redundant inspection operations. More reasonable maintenance decisions should handle this weakness.

3.3 *Proposed framework and representative strategy*

To avoid the drawback of the above classical maintenance decision framework, we propose not to use the same condition index for both scheduling the inspections and deciding of a replacement. Two condition indices instead of a single one should be used: *one to control the inspection decision*, and *the other to control the replacement decision*. Accordingly, a replacement is not necessarily always performed at an inspection time τ_i, but at a predetermined time τ_r (normally $\tau_r \neq \tau_i$). The decision of performing an inspection could be based on the property that *the system health prognosis accuracy improves with age*. As such, since a certain age τ_i of the system, we can know quite precisely its failure time. Additional inspection operations are thus no longer necessary, and it might be enough just to wait ψ time units, $\psi \geq 0$, from the last inspection to do a system replacement at $\tau_r = \tau_i + \psi$. Obviously, because of the use of two different health indexes for the inspection scheduling and replacement decision, this new predictive maintenance decision framework is more flexible than the classical one. Furthermore, when the waiting time $\psi = 0$, the new framework returns to the classical one; so it is more general and more profitable in most cases. In the following, a typical and representative maintenance strategy is presented in order to illustrate this new framework.

Usually, the accuracy of health prognosis can be measured through the standard deviation of the conditional system RUL, and the waiting time ψ can be determined from the system conditional MRL. For the system model considered in Section 2.1, the degradation level detected at an inspection time returns the same information as the standard deviation of the conditional system RUL: *the higher the degradation, the more the system RUL prognosis is accurate*, so we merely use the system degradation level to control inspection decisions. This choice, on one hand, simplifies the computation, and on the other hand, allows a maintenance decision rule consistent with the above representative of the classical framework. Of course, when the system is more complex (e.g., multi-unit or multi-failure mode), the more "overarching" standard deviation of the conditional system RUL should be used instead of the system degradation level. As a result, an exemplary maintenance strategy can be stated as follows. Over a certain renewal cycle, the system is regularly inspected with period δ. At a scheduled inspection date $\tau_i = i \cdot \delta$, $i = 1, 2, \ldots$, a CR of the system is carried out if it fails (i.e., $X_{\tau_i} \geq L$). But, if the system is still running, a decision based on the accuracy of the system RUL prognos is given at τ_i is adopted. If $\xi \leq X_{\tau_i} < L$, where ξ is a degradation threshold indicating the accuracy of RUL prognosis, no additional inspection is needed for the *current* renewal cycle because the system failure time can be already predicted with an acceptable precision, and a system replacement is planned ψ time units later (i.e., at time $\tau_r = \tau_i + \psi$). The waiting time i is defined from the system conditional MRL as follows

$$\psi(y) \triangleq \left(\mu(\tau_i \mid y) - \eta\right) \cdot 1_{\{0 \leq \eta \leq \mu(\tau_i \mid y)\}}, \qquad (7)$$

where $\mu(\tau_i \mid y)$ is the system conditional MRL given from (5), η is known as safety time interval. The replacement at τ_r may be either preventive or corrective depending on the working or failure state of the system at this time. After the replacement, a new renewal cycle begins, and the next

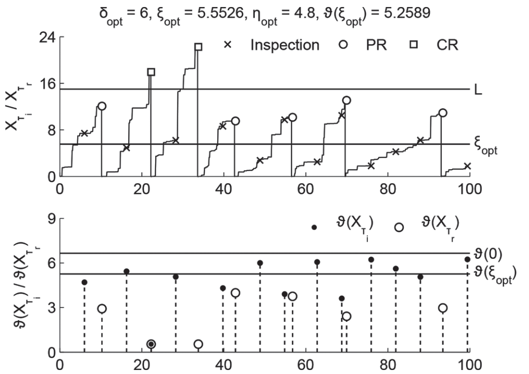


Figure 3. Illustration of the (δ, ξ, η) strategy.

inspection time will be carried out on this cycle at $\tau_{i+1} = \tau_r + \delta$. Otherwise (i.e., $X_{\tau_i} < \xi$), we cannot predict precisely the failure time, so one or more inspections are needed to gather additional information about the system, and the maintenance decisions are postponed until the next inspection time $\tau_{i+1} = \tau_i + \delta$. The maintenance strategy admits the inspection period δ, the degradation threshold to control prognosis accuracy ξ and the safety time interval η as decision variables, so we called it (δ,ξ,η). With the same system and maintenance costs as in the illustration of Section 3.2, the (δ,ξ,η) strategy reaches its optimal configuration at $\delta_{opt} = 6$, $\xi_{opt} = 5.5526$ and $\eta_{opt} = 4.8$. The evolution of the maintained system under the optimal configuration is shown in Figure 3. We can see that ξ_{opt} of the (δ,ξ,η) strategy is much smaller than ζ_{opt} of the (δ,ζ) strategy; this means that the (δ,ξ,η) strategy does not need a RUL prediction accuracy as high as the (δ,ζ) strategy. In other words, the information about the health prognosis accuracy has been taken into account to improve the maintenance decision-making.

4 COST MODEL AND OPTIMIZATION

The long-run expected maintenance cost rate is used here as a cost criterion to assess the performance of the considered maintenance frameworks

$$C_\infty = \lim_{t\to\infty} \frac{E\left[C(t)\right]}{t}, \tag{8}$$

$C(t)$ denotes the total maintenance cost including the downtime cost up to time t: $C(t) = C_i \cdot N_i(t) + C_p \cdot N_p(t) + C_c \cdot N_c(t) + C_d \cdot W(t)$, where $N_i(t)$, $N_p(t)$ and $N_c(t)$ are respectively the number of inspections, PR and CR operations in [0, t], and $W(t)$ is the system downtime interval in [0, t]. In the literature, the cost rate C_∞ is usually evaluated analytically by the *renewal-reward theorem* (Tijms 2003). This classical method is normally useful for *static* maintenance decision rules (Huynh et al. 2014). When the decision rules are *dynamic* as in the present paper, it is more interesting to take advantage of the *semi-regenerative theory* (Cocozza-Thivent 1997). Consequently, (8) can be rewritten as (Bérenguer 2008)

$$\begin{aligned} C_\infty = C_i \cdot \frac{E_\pi\left[N_i(\Delta\tau)\right]}{E_\pi[\Delta\tau]} + C_p \cdot \frac{E_\pi\left[N_p(\Delta\tau)\right]}{E_\pi[\Delta\tau]} \\ + C_c \cdot \frac{E_\pi\left[N_c(\Delta\tau)\right]}{E_\pi[\Delta\tau]} + C_d \cdot \frac{E_\pi\left[W(\Delta\tau)\right]}{E_\pi[\Delta\tau]}, \end{aligned} \tag{9}$$

where $\Delta\tau = \tau_i^- - \tau_{i-1}^-$, $i = 1,2,\ldots$, denotes the length of a single *Markov renewal cycle* which is the time interval between 2 successive inspections, E_π denotes the expectation with respect to π. In the following, we focus on formulating the stationary law π and the expectation quantities in (9).

4.1 *Stationary law of the maintained system state*

The behavior of maintained system at inspection times can be characterized by the stationary law π of the Markov chain continuous state space $\mathbb{R}_+$, $\{Y_i\}_{i\in N}$. Let consider the Markov renewal cycle $[\tau_{i-1}^-, \tau_i^-]$, $i = 1, 2, \ldots$, y and x are respectively the degradation levels of maintained system at the beginning and the end of the cycle (i.e., $X_{\tau_{i-1}^-} = y$ and $X_{\tau_i^-} = x$), the stationary law π is the solution of the following invariant equation

$$\pi(x) = \int_{\mathbb{R}_+} F(x \mid y)\pi(y)dy, \tag{10}$$

where $F(x \mid y)$ is the degradation transition law from y to x. This transition can be obtained by an exhaustive analysis of all the possible evolution and maintenance scenarios on the Markov renewal cycle $\left[\tau_{i-1}^-, \tau_i^-\right]$. As a result, we obtain

$$\begin{aligned} \pi(x) = \int_0^\xi f_{\alpha\delta,\beta}(x-y)\pi(y)dy + f_{\alpha\delta,\beta}(x) \\ \times \int_\xi^L \left(\int_y^\infty f_{\alpha\psi(y),\beta}(z-y)dz\right)\pi(y)dy \\ + f_{\alpha\delta,\beta}(x)\int_L^\infty \pi(y)dy, \end{aligned} \tag{11}$$

where $f_{\alpha\psi(y),\beta}(\cdot)$ and $f_{\alpha\delta,\beta}(\cdot)$ are derived from (1). Given (11), we can adapt the fixed-point iteration algorithm to numerically evaluate $\pi(x)$. Many numerical tests were carried out, and they have shown that the algorithm converges very quickly to the true stationary law.

4.2 *Expected quantities*

On a Markov renewal cycle, $E_{\pi}\left[N_i\left(\Delta\tau\right)\right]=1$. The other expected quantities are all computed in a similar way by integration with respect to the stationary law π determined in Section 4.1.

$$E_{\pi}[\Delta\tau]=\delta+\int_{\xi}^{L}\psi(y)\pi(y)dy. \tag{12}$$

$$E_{\pi}\left[N_p\left(\Delta\tau\right)\right]=\int_{\xi}^{L}F_{\alpha\psi(y),\beta}(L-y)\pi(y)dy. \tag{13}$$

$$\begin{aligned}E_{\pi}\left[N_c\left(\Delta\tau\right)\right]&=\int_{\xi}^{L}\bar{F}_{\alpha\psi(y),\beta}(L-y)\pi(y)dy\\&+\bar{F}_{\alpha\delta,\beta}(L)\int_{\xi}^{L}\pi(y)dy+\int_{0}^{\xi}\bar{F}_{\alpha\delta,\beta}(L-y)\pi(y)dy.\end{aligned} \tag{14}$$

$$\begin{aligned}E_{\pi}\left[W\left(\Delta\tau\right)\right]=&\int_{\xi}^{L}\pi(y)dy\int_{0}^{\delta}\bar{F}_{\alpha u,\beta}(L)du\\&+\int_{\xi}^{L}\left(\int_{0}^{\psi(y)}\bar{F}_{\alpha u,\beta}(L-y)du\right)\pi(y)dy\\&+\int_{0}^{\xi}\left(\int_{0}^{\delta}\bar{F}_{\alpha u,\beta}(L-y)du\right)\pi(y)dy.\end{aligned} \tag{15}$$

In the above expressions, $\bar{F}_{\alpha\cdot(\cdot),\beta}(\cdot)=1-F_{\alpha\cdot(\cdot),\beta}(\cdot)$ and $F_{\alpha\cdot(\cdot),\beta}(\cdot)$ is given from (2).

4.3 *Maintenance optimization*

Using (11), and introducing (12), (13), (14) and (15) into (9), we obtain the full mathematical cost model of the proposed predictive maintenance decision framework. The classical framework is a particular case of the new one, its cost model is also derived from (9) by taking $\psi=0$. Given the cost model, optimizing the strategies (δ,ζ) and (δ,ξ,η) returns to find the set of decision variables of each strategy that minimizes the associated long-run expected cost maintenance rate

$$C_{\infty}\left(\delta_{opt},\zeta_{opt}\right)=\min_{(\delta,\zeta)}\left\{C_{\infty}\left(\delta,\zeta\right)\right\},$$
$$C_{\infty}\left(\delta_{opt},\xi_{opt},\eta_{opt}\right)=\min_{(\delta,\xi,\eta)}\left\{C_{\infty}\left(\delta,\xi,\eta\right)\right\},$$

where $\delta\geq 0$, $0\leq\zeta,\xi\leq L$ and $\eta\geq 0$. The *generalized pattern search algorithm* presented in (Audet et al. 2002) can be resorted to find the optimal maintenance cost rate and the associated decision variables. Applying this algorithm to the system and the set of maintenance costs presented at the beginning of Section 3, we obtain optimal quantities as in Table 1. Based on the optimal values of cost rates, the (δ,ξ,η) strategy is more profitable than the (δ,ζ) strategy. However, this is just the conclusion for the present special case. More general conclusions on the performance these strategies are given in Section 5.

Table 1. Optimal configuration of (δ, ζ) and (δ, ξ, η) strategies.

Optimal decision variables	Optimal cost rate
$\delta_{opt}=4.6$, $\zeta_{opt}=9.1478$	$C_{\infty}(\delta_{opt}, \zeta_{opt})=6.3422$
$\delta_{opt}=6$, $\xi_{opt}=5.5526$ $\eta_{opt}=4.8$	$C_{\infty}(\delta_{opt}, \xi_{opt}, \eta_{opt})=5.9746$

5 PERFORMANCE ASSESSMENT

This section aims at seeking a general conclusion on the effectiveness of the new predictive maintenance decision framework. To this end, we compare the performance of the (δ, ξ, η) strategy to the (δ, ζ) strategy under various configurations of maintenance operations costs and system characteristics. The so-called *relative gain in the optimal long-run expected maintenance cost rate* (Huynh et al. 2015) is resorted for this purpose

$$\kappa_C\left(\%\right)=\frac{C_{\infty}\left(\delta_{opt},\zeta_{opt}\right)-C_{\infty}\left(\delta_{opt},\xi_{opt},\eta_{opt}\right)}{C_{\infty}\left(\delta_{opt},\zeta_{opt}\right)}\cdot 100\%.$$

If $\kappa_C\left(\%\right)>0$, the (δ,ξ,η) strategy is more profitable than the (δ,ζ) strategy; if $\kappa_C\left(\%\right)=0$, they have the same profit; otherwise, the (δ, ξ, η) strategy is less profitable than the (δ, ζ) strategy.

At first, we are interested in the impact of the maintenance costs on the performance of the new predictive maintenance decision framework. This is why we fix the system characteristic at $L=15$ and $\alpha=\beta=0.2$ (i.e., $m=1$ and $\sigma^2=5$), the practical constraint $C_i<C_p<C_c$ also leads us to take $C_c=100$ and consider the three cases

- *varied inspection cost*: C_i varies from 1 to 49 with step equals 1, $C_p=50$, and $C_d=25$,
- *varied PR cost*: $C_i=5$, C_p varies from 6 to 99 with step equals 3, and $C_d=25$,
- *varied downtime cost rate*: $C_i=5$, $C_p=50$, and C_d varies from 10 to 190 with step equals 5.

For each of above cases, we sketch the relative gain $\kappa_C\left(\%\right)$, and the results are obtained as in Figure 4. Not surprisingly, the (δ,ξ,η) strategy is always more profitable than the (δ,ζ) strategy (i.e., $\kappa_C\left(\%\right)>0$). Thus, there is no risk when using the proposed maintenance decision framework (i.e., it returns to the classical one in the worst case). This is a significant advantage of this new framework. Moreover, it is especially profitable

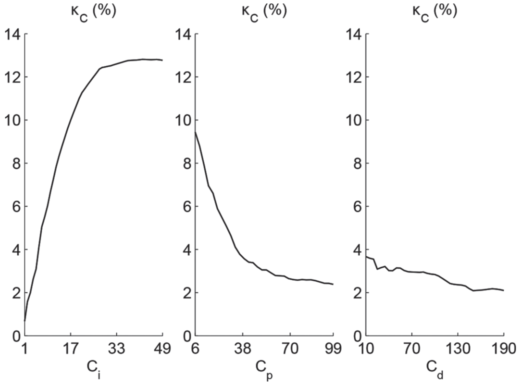


Figure 4. $\kappa_C(\%)$ with respect to maintenance costs.

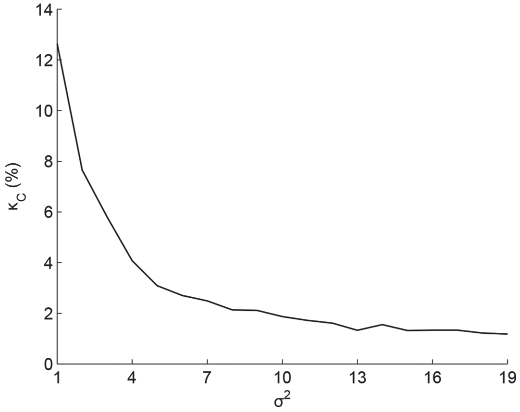


Figure 5. $\kappa_C(\%)$ with respect to the degradation variance.

when the inspection is expensive (see the left side of Figure 4) as it can avoid inopportune inspection operations. When the PR cost C_p or the downtime cost rate C_d increases, the inspection cost C_i becomes relatively smaller, and hence the relative gain $\kappa_C(\%)$ is weaker (see the middle and the left side of Figure 4). Consequently, unlike the inspection cost, the PR cost or the downtime cost rate do not much affect the performance of the proposed framework.

To investigate the impact of the degradation variance σ^2 on the performance of the proposed maintenance decision framework, we take $L = 15$ and $m = 1$, vary σ^2 from 1 to 19 with step 1, and we study the evolution of $\kappa_C(\%)$ when the set of maintenance costs is fixed at $C_i = 5$, $C_p = 50$, $C_c = 100$ and $C_d = 25$. The result is as in Figure 5. Once again, the (δ,ξ,η) strategy is always economically better than the (δ,ζ) strategy by the same reason as above. Under the new maintenance decision framework, the sooner the accuracy level of RUL prognosis is reached, the less inopportune inspections are performed. The lower degradation variance allows a RUL prognosis with higher precision, hence it is not surprising that the (δ,ξ,η) strategy is most profitable at the small values of σ^2, and its profit decreases when the system becomes more chaotic (see Figure 5).

6 CONCLUSIONS & PERSPECTIVES

We have proposed in this paper a new parametric predictive maintenance decision framework considering improving health prognosis accuracy for stochastically deteriorating single-unit systems. Many numerical experiments show the advantage of the proposed maintenance decision framework compared to the most well-known one in the literature. In fact, the proposed framework is more general and more flexible than the classical one, so there is no risk when using the new framework. Furthermore, this new framework is especially suitable for the systems with small degradation variance and incurred high inspection costs. The results in the present paper also confirm the interest of the system health prognosis information for maintenance decision-making when it is properly used. This encourages us to continue investing in prognostics and health management technologies, and building new predictive maintenance strategies.

For our future works, we continue studying the advantage of the proposed predictive maintenance decision framework for multi-unit systems (e.g., deteriorating systems with k-out-of-n structure). We also believe that the framework will be particularly suitable for systems with limited number of inspection and repair facilities (e.g., offshore systems such as submarine power cables, offshore wind turbines, subsea blowout preventer system, etc.).

REFERENCES

Abdel-Hameed, M. (2014). *Lévy Processes and Their Applications in Reliability and Storage*. SpringerBriefs in Statistics. Springer.

Audet, C., J. Dennis, & E. John (2002). Analysis of generalized pattern searches. *SIAM Journal on Optimization 13*(3), 889–903.

Banjevic, D. (2009). Remaining useful life in theory and practice. *Metrika 69*(2–3), 337–349.

Bérenguer, C. (2008). On the mathematical condition-based maintenance modelling for continuously deteriorating systems. *International Journal of Materials and Structural Reliability 6*(2), 133–151.

Blain, C., A. Barros, A. Grall, & Y. Lefebvre (2007). Modelling of stress corrosion cracking with stochastic processes–application to steam generators. In *Proc.*

of the European Safety and Reliability Conference - ESREL 2007, pp. 2395–2400.

Bousquet, N., M. Fouladirad, A. Grall, & C. Paroissin (2015). Bayesian gamma processes for optimizing condition-based maintenance under uncertainty. *Applied Stochastic Models in Business and Industry 31*(3), 360–379.

Cocozza-Thivent, C. (1997). *Processus stochastiques et fiabilité des systèmes*, Volume 28 of *Mathématiques & Applications*. Springer. In French.

Grall, A., C. Bérenguer, & L. Dieulle (2002). A condition-based maintenance policy for stochastically deteriorating systems. *Reliability Engineering & System Safety 76*(2), 167–180.

Huynh, K.T., A. Barros, & C. Bérenguer (2015). Multi-level decision-making for the predictive maintenance of k-out-of-n:f deteriorating systems. *IEEE Transactions on Reliability 64*(1), 94–117.

Huynh, K.T., A. Barros, C. Bérenguer, & I.T. Castro (2011). A periodic inspection and replacement policy for systems subject to competing failure modes due to degradation and traumatic events. *Reliability Engineering & System Safety 96*(04), 497–508.

Huynh, K.T., I.T. Castro, A. Barros, & C. Bérenguer (2014). On the use of mean residual life as a condition index for condition-based maintenance decision-making. *IEEE Transactions on Systems, Man, and Cybernetics: Systems 44*(7), 877–893.

Jardine, A.K.S., D. Lin, & D. Banjevic (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing 20*(7), 1483–1510.

Kallen, M.J. & J.M. van Noortwijk (2005). Optimal maintenance decisions under imperfect inspection. *Reliability Engineering & System Safety 90*(2–3), 177–185.

Khoury, E., E. Deloux, A. Grall, & C. Bérenguer (2013). On the use of time-limited information for maintenance decision support: A predictive approach under maintenance constraints. *Mathematical Problems in Engineering* 2013. Article ID 983595, 11 pages, doi:10.1155/2013/983595.

Langeron, Y., A. Grall, & A. Barros (2015). A modeling framework for deteriorating control system and predictive maintenance of actuators. *Reliability Engineering & System Safety 140*, 22–36.

Liao, L. & F. Kottig (2014). Review of hybrid prognostics approaches for remaining useful life prediction of engineered systems, and an application to battery life prediction. *IEEE Transactions on Reliability 63*(1), 191–207.

Ponchet, A., M. Fouladirad, & A. Grall (2010). Assessment of a maintenance model for a multi-deteriorating mode system. *Reliability Engineering & System Safety 95*(11), 1244–1254.

Santini, T., S. Morand, M. Fouladirad, L.V. Phung, F. Miller, B. Foucher, A. Grall, & B. Allard (2014). Accelerated degradation data of sic mosfets for lifetime and remaining useful life assessment. *Microelectronics Reliability 54*(9), 1718–1723.

Shin, J.H. & H.B. Jun (2015). On condition based maintenance policy. *Journal of Computational Design and Engineering 2*(2), 119–127.

Si, X.S., W. Wang, C.H. Hu, & D.H. Zhou (2011). Remaining useful life estimation—a review on the statistical data driven approaches. *European Journal of Operational Research 213*(1), 1–14.

Singpurwalla, N.D. (1995). Survival in dynamic environments. *Statistical Science 10*(1), 86–103.

Tijms, H. (2003). *A first course in stochastic models*. Wiley, New York.

Travé-Massuyès, L. (2014). Bridging control and artificial intelligence theories for diagnosis: A survey. *Engineering Applications of Artificial Intelligence 27*, 1–16.

Van Noortwijk, J.M. (2009). A survey of the application of gamma processes in maintenance. *Reliability Engineering & System Safety 94*(1), 2–21.

Wang, X. (2009). Nonparametric estimation of the shape function in a gamma process for degradation data. *Canadian Journal of Statistics 37*(1), 102–118.

Xu, Z., Y. Ji, & D.H. Zhou (2008). Real-time reliability prediction for a dynamic system based on the hidden degradation process identification. *IEEE Transactions on Reliability 57*(2), 230–242.

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Condition-based maintenance by minimax criteria

O. Abramov & D. Nazarov
Laboratory of Complex Systems Reliability Control, Institute of Automation and Control Processes FEB RAS, Vladivostok, Russia

ABSTRACT: An approach to design the condition-based maintenance of engineering systems is considered in the paper. This approach is based on prediction of system state. The algorithms which provide perfect predicting of system state and designing of optimal preventive maintenance strategies in the case when inspection data are incomplete and insignifican are discussed.

1 INTRODUCTION

The excessive servicing and increasing requirements in the efficient operation of equipment determine the necessity to state a maintenance problem in a new way to solve it for every particular unit of equipment on the individual basis. An information base for preventive maintenance is formed by predicting and estimating of an engineering system state. There prove to be many difficulties in solving the matter, These difficulties are caused by the lack on storage of apriority statistic information on field variation of system parameters. In this case the application of classical methods of mathematical statistic to the solution of problems of predicting the state and predictive maintenance scheduling may cause serious errors. One is then confronted with problems of designing a predictor and an optimal strategy of predictive maintenance in the event of incomplete model knowledge.

Both these problems are referred to minimax problems. With that the first problem or the problem of minimax predicting the state was considered early (Abramov et al. 2000), (Abramov and Rozenbaum 2004). But it should not be considered that this problem is settled completely. The second problem, referred to as minimax control, has attracted attention to handle predictive maintenance. In this paper, we present a new minimax algorithm for predicting the state and solve a problem of picking an optimal minimax strategy of individual maintenance without using information about stochastic properties of measurement noises and errors of the model chosen to describe the random processes of parameter variation. The predicting algorithm is constructed by a minimax squaring criterion. An optimal minimax strategy of individual maintenance is found by methods of dynamic programming. Our approach requires only the possible variation limits for unknown noises and model errors, which may be given in a rather coarse way. This paper is organized as follows. In Section II, we formulate the problem of predicting the state of an engineering system in the case when inspection data are incomplete and insignificant. In Section III, we formulate the problem of optimal minimax (maximin) predictive maintenance. In Section IV, we construct the predictor by the minimax squaring criterion. In Section V, we solve the problem of picking an optimal minimax strategy of predictive maintenance. Some conclusions complete the paper.

2 FORMULATION OF THE PROBLEM OF MINIMAX STATE PREDICTING

Assume that variations in system state parameters can be approximated in a rather coarse way as follows:

$$Y(t) = \mathbf{A}^T \mathbf{u}(t),\, t \in T, \tag{1}$$

where $\mathbf{A} = \{a_j\}_{j=0}^{n}$ is a set of random coefficients, $\mathbf{u}(t) = \{u_j(t)\}_{j=0}^{n}$ are continuous deterministic functions of time, and T is the operating period. The model errors are present here but they are not determined.

The representation (1) can be discussed as an expansion $y(t)$ by some function system. Such expansion allows to approximate theoretically any real process $y(t)$ and it is variance with the facts.

The system state $y(t)$ is monitored on the interval time $T_p \subset T$ with additive error $e(t)$. Measurements form a sequence $\mathbf{Z} = \{z(t_k)\}_{k=1}^{p}, t \in T_p \subset T$. Probability properties $e(t)$ are not determined, it is only known that

$$|\, e(t) | \le c(t), t \in T_p \subset T, \tag{2}$$

where $c(t)$ is a given function.

The problem consists in determining of estimations $\tilde{y}(t), t \in T \setminus T_p$.

The model of system state variations (1), the constrains (2) on disturbances and measurements $\mathbf{Z}$ form an initial data base for a solution of the problem. The deficiency and uncertainty of data base, in particular, the default of full probability properties of the disturbances and the presence of unknown errors in the model (1) make in difficult to obtain the sought estimate $\tilde{y}(t), t \in T \setminus T_p$ by using well-known statistical techniques, such as least-squares method, least-magnitudes method, and so on. It seems preferential to construct the estimate $\tilde{y}(t), t \in T \setminus T_p$ proceeding from the worst cases, i. e, on basic of minimax concept (Abramov, Rozenbaum, & Suponya 2000);

3 OPTIMAL MINIMAX (MAXIMIN) PREDICTIVE MAINTENANCE

Let the state of the equipment unit during service is described by a parameter y. Under influence of destabilizing factors, y varies in a random manner. These variations can be approximated by the expansion (1). The stochastic properties of $y(t)$ are unknown.

The variations of $y(t)$ as a matter of fact can be lead to failures or to a deterioration of functioning of engineering systems. To prevent such undesirable occurrences we must control of $y(t)$. The control is realized by an predictive maintenance of engineering systems.

It is necessary to give a performance criterion to optimize an predictive maintenance. Different technical-and-economic indexes: reliability, capacity, efficiency and so on can be such criteria. A performance index generally represents a functional s. It is evident that any change of equipment state $y \in Y$ involves a change in the performance index $G(Y,T)$. Then the predictive maintenance consists in tracking $y(t)$ and forcedly changing it in some instants of time $t \in T$. It represents field inspection and adjustment of equipment parameters, replacement of units, assemblies and components whose parameters reach their critical values. An inspection result $z(t) = y(t) + e(t)$, where $e(t)$ is the random error whose stochastic properties are undeterminable but whose range of value E is known. Adjustment consists in changing $y(t)$ by nonrandom value $r \in R$.

We shall consider, as a control strategy, a function $s(t)$, where $s \in S$ (S is the set of preventive actions). The problem of constructing $s(t)$ in minimax (maximin) statement may be written as follows:

$$\begin{aligned} g^* &= \min_{s(t) \in S \times T} \max_{y(t) \in Y \times T} G(y,s,t), \\ g^{**} &= \max_{s(t) \in S \times T} \min_{y(t) \in Y \times T} G^*(y,s,t). \end{aligned} \tag{3}$$

It is evident that a function which delivers minimum of maximum (or maximum of minimum) of index $G(y,s,t)$ (or $G^*(y,s,t)$) is a sought minimax (or maximin) predictive maintenance strategy. In other words, the representation (3) are problem formulation of optimal minimax (maximin) predictive maintenance.

4 MINIMAX PREDICTING ALGORITHM

In accordance with the statement making in section 2 a problem of predicting the state consist in definition of estimations $\tilde{y}(t), t \in T \setminus T_p$, such estimations must ensure the fulfilling $\min \| y(t) - \tilde{y}(t) \|, \forall t \in T \setminus T_p$. In assumption of presents of model errors e(t) in the relationship (1) for $t \in T$ the considering problem can be formulated as follows:

$$G = \min_{A} \max_{|\mathbf{e}| \le \mathbf{c}} \| \mathbf{z} - \mathbf{e} - \mathbf{A}^T \cdot U \|, \tag{4}$$

where $\mathbf{z} = \{z(t_k))\}_{k=1}^{p}, U = \| u_j(t_k) \|_{k=1, j=0}^{p,m}, \mathbf{e} = \{e(t_k)\}_{k=1}^{p}$, $\mathbf{c} = \{c(t_k)\}_{k=1}^{p}$.

A norm of misclosure, i. e. $\| \mathbf{z} - \mathbf{e} - \mathbf{A}^T \cdot U \|$, serves as the optimality criterion in (4).

This norm can be given as $\| \mathbf{z} - \mathbf{e} - \mathbf{A}^T \cdot U \| = (\mathbf{z} - \mathbf{e} - \mathbf{A}^T \cdot U)^T (\mathbf{z} - \mathbf{e} - \mathbf{A}^T \cdot U)$, i. e as squaring. With that we obtain:

$$G = \max_{|\mathbf{e}| \le \mathbf{c}} \min_{A} (\mathbf{z} - \mathbf{e} - \mathbf{A}^T \cdot U)^T (\mathbf{z} - \mathbf{e} - \mathbf{A}^T \cdot U). \tag{5}$$

By defining minimum from $\mathbf{A}$ in (5) we find that

$$\mathbf{A}^{opt} = (U^T \cdot U) \cdot U^T \cdot (\mathbf{z} - \mathbf{e}). \tag{6}$$

The substitution (6) into (5) gives the following:

$$G = \max_{|e| \le c} (\mathbf{z} - \mathbf{e})^T \cdot L \cdot (\mathbf{z} - \mathbf{e}), \tag{7}$$

where L is the symmetric matrix an arrangement of p elements taken $(n+1)$ at a time $L = (I - U \cdot (U^T \cdot U)^{-1} \cdot U^T)^T \cdot (I - U \cdot (U^T \cdot U)^{-1} \cdot U^T)$, I is unit matrix.

Using (6) and (7) we can obtain numerical values $\{a_j^{opt}\}_{j=0}^{n}$, and then define a sought estimation $g(t), t \in T \setminus T_p$ using (1). A solution (6) can be found by the methods of nonlinear programming here. It should noted that $\gamma(t)$ maximum error of determination $\tilde{y}(t), t \in T \setminus T_p$ we can estimate as the following:

$$\gamma(t) \leq \max|c(t^*)| \cdot \max|\mathbf{u}(t)|, \tag{8}$$

where t^* is a fixing instant out of $T_p \subset T$.

The algorithm under discussion meet general requirement to any predicting procedure. Estimates found are unique, optimal, and unbiased.

5 PREDICTIVE MAINTENANCE STRATEGY

To obtain a concrete solution of the task (3) we must give a concrete optimality criterion. In principle, such criterion should be chosen based on the requirements to the performance of the particular equipment unit. The requirements a specified on the basis of a certain index system. Within this system economic indexes are most general. The indexes, in particular, include a guaranteed level of total losses when using the engineering system on set T. The index can be written as follows:

$$W_T = \sup_{y(t) \in Y \times T} \int_T H(y(t))dt + V_T, \tag{9}$$

where $H(y(t))$ is the loss function which describes losses when equipment state differs from normal one; V_T is the operating expenses.

We may obtain a globally optimal strategy $S(t)$ by stepwisely minimizing criterion W_T based on Bellman's optimality principle. Algorithms applied are simple enough and can be implemented in recurrent form. We should build a state space to form recurrent relations for solving problem on the basis of the optimality principle. In other words, we should find a coordinate set which contains all information about the engineering system in a given time interval regard less of its past behavior. When $y(t)$ is described by (1) a sought set can be represented as a collection $(\mathbf{A},t,t_c)$ where $\mathbf{A}$ is a vector whose members define ranges of coefficients $a_j, j=0,n$ in model (1), $t,t_c \in T, t \leq t_c$.

Let the function $W(\mathbf{A},t,t_c)$ describes limiting losses associated with the optimal servicing of the engineering system in state $(\mathbf{A},t,t_c)$. Predictive maintenance consists in inspecting and adjusting $y(t)$ (it is not difficult to show that replacing units, assemblies and components of the engineering system is equivalent to the adjustment of $y(t)$).

The limiting losses when using the engineering system in state $(\mathbf{A},t,t_c)$ without any maintenance in interval $[t,t_c]$ can be represented in the form:

$$W_1(\mathbf{A},t,t_c) = \max_{\mathbf{A}^*} \int_t^{t_c} H(y(\tau))d\tau. \tag{10}$$

If at an instant $\tilde{t}, t \leq \tilde{t} \leq t_c$ we take a reading of $y(t)$ associated with expanses α and we obtain a value $z(t) = y(t) + e(t)$ ($e(t)$ is a random measurement error whose stochastic properties are undeterminable but only its range E is known), then the information state of the engineering system will be $(\mathbf{A}^*,t,t_c)$, where $\mathbf{A}^*$ is the vector of coefficients a_j obtained from measurement result. The the limiting losses are $W_2(\mathbf{A},t,t_c) = W_1(\mathbf{A},t,\tilde{(t)}) + \alpha + W(\mathbf{A}^*,\tilde{t},t_c)$.

If at instant $\tilde{t}, t \leq \tilde{t} \leq t_c$ we adjust $y(t)$ (a change of $y(t)$ by $r \in \mathbb{R}$) and associated expanses are β, then limiting losses for state $(\mathbf{A},t,t_c)$ can be described as follows:

$$W_3(\mathbf{A},t,t_c) = W_1(\mathbf{A},t,\tilde{t}) + \beta + W(\mathbf{A}_*,\tilde{t},t_c), \tag{11}$$

where $\mathbf{A}_*$ is the vector of coefficients a_j with allowance for a change in state of the engineering system after the adjustment.

Based on (10), we can form recurrent relations for finding an optimal strategy $S(t)$:

$$W(t,t_c,\mathbf{A}) = \min_{i=1,2,3} W_i, \tag{12}$$

$$W_1 = W_1(t,t_c,\mathbf{A}), \tag{13}$$

$$W_2 = \min_{t \leq \tilde{t} \leq t_c} W_1(t,\tilde{t},\mathbf{A}) + \alpha + \min_{\mathbf{A}^*} W(\mathbf{A}^*,\tilde{t},t_c), \tag{14}$$

$$W_3 = \min_{t \leq \tilde{t} \leq t_c} W_1(t,\tilde{t},\mathbf{A}) + \beta + \min_{r \in \mathbb{R}} W(\mathbf{A}_*,\tilde{t},t_c). \tag{15}$$

The value of i for which minimum in (12) is achieved and values $\tilde{t}$, r for which minimum of minima in (14) and (15) are achieved are functions of $(\mathbf{A},t,t_c)$ and describe a sought optimal predictive maintenance strategy $S(t)$.

As a matter of fact, solving equations (12), (13), (14), (15) is a problem of dynamic programming. To form $S(t)$ we can make use of space approximation technique. With that for finding $\mathbf{A}^*$ it is necessary to use minimax algorithms of predicting the state, in particular, it can be the algorithm from Section 4.

6 CONCLUSIONS

The problem of designing of predictive maintenance of engineering systems is solved for the case when inspection data are incomplete and insignificant. It is usually the case in actual practice. With that we find a global optimal (in minimax sense) strategy of predictive maintenance here. This strategy guarantees the efficient functioning of operated engineering systems on the operating interval with minimal expanses. There is a minimax predicting algorithm in this paper. The algorithm allows to obtain strict estimations for own unknown quantities. The estimation are more useful and reliable than statistical estimations of the task of predicting the state where the single realization $y(t)$ is observed. The considered algorithm intends for accompaniment of solution of tasks of predictive maintenance. The proposed approach has been realized on practice as the theoretical basis for designing of controlling and measurement systems (Abramov and Rozenbaum 2004), (Abramov 2010).

ACKNOWLEDGMENT

This work is partially supported with the grant 14-08-00149 of the Russian Foundation for Basic Research.

REFERENCES

Abramov, O. & A. Rozenbaum (2004). Passive control of the operation of measuring instruments. *Measurement Techniques. 47 No.3*, 233–239.

Abramov, O. (2010). Parallel algorithms for computing and optimizing reliability with respect gradual failures. *Automation and Remote Control. 71, No.7*, 1394–1402.

Abramov, O., A. Rozenbaum, & A. Suponya (2000). Failure prevention based on parameters estimation and prediction. In *Preprints of 4th IFAC Symposium "SAFEPROCESS 2000".*, Budapest, Hungary, pp. 584–586.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Impact of cost uncertainty on periodic replacement policy with budget constraint: Application to water pipe renovation

Khanh T.P. Nguyen
University of Technology of Troyes, LM2S, Troyes, France

Do T. Tran
Inria Lille—Nord Europe, Villeneuve d'Ascq, France University of Lille, CRIStAL (CNRS UMR 9189), Lille, France

Bang Minh Nguyen
Non-Revenue Water Reduction Department
Nha Be Water Supply Joint-Stock Company, HCMC, Vietnam

ABSTRACT: Due to market flexibility, repair and replacement costs in reality are often uncertain and can be described by an interval that includes the most preferable values. Such a portrayal of uncertain costs naturally calls for the use of fuzzy sets. In this paper, we therefore propose using fuzzy numbers to characterize uncertainty in repair and replacement costs. The impact of fuzzy costs on the optimal decision is then investigated in the context of an industrial problem: optimizing water pipe renovation strategies. Here we examine specifically the risk of violating a budget constraint imposed on the annual cost associated with pipe failure repairs. This risk is evaluated using the mean chance of the random fuzzy events that represent random fuzzy costs exceeding a given budget. The benefit of taking account of cost uncertainty is then validated through various numerical examples.

1 INTRODUCTION

The role of maintenance policies in industry is increasingly highlighted as they can reduce production costs, extend the useful life of industrial equipment, and also alter the strategy for new investments in equipment. Among maintenance policies, periodic maintenance is a traditional policy that is commonly used in reality thanks to its simplicity of implementation and ability to easily integrate related constraints into decision processes. The main objective of periodic maintenance is to determine the best replacement time that maximizes system reliability or availability and safety, and minimizes maintenance cost. The basic policy proposed by Barlow & Hunter (1960) recommends that equipment be replaced after $k.\ T$ hours, where $k = 1, 2, \ldots$, and any failure occurring between two successive replacements be restored with a minimal repair. This minimal-repair model assumes that the cost of minimal repairs is lower than the cost of a preventive replacement.

Various extensions and variations of this basic model have been proposed in the literature over the time. Many of them have been surveyed in (Wang 2002, Ahmad & Kamaruddin 2012). Those extensions can generally be classified into three groups. The first group focuses on improving system failure models and takes into account the effects of different failure modes or "shock" condition. Sheu & Griffith (2002), Sheu et al. (2010) proposed a periodic replacement policy for systems subjected to shocks. Lai & Chen (2006) considered a two-unit system with failure rate interaction between the units. In (Sheu et al. 2012), the authors considered non-homogeneous pure birth shocks for the block replacement policy. The second group aims to solve the questions of large size and complex systems. Wang & Pham (2006) studied the correlated failures of multiple components in a serial system and aimed to optimize system availability and/or maintenance costs. Scarf & Cavalcante (2010) proposed hybrid block replacement and inspection policies for a multi-component system in serial structures. The third group extends maintenance policies by considering numerous maintenance decisions and evaluate the performance of these activities. The model proposed by Sheu (1992) considered the possibility of installing a new alternative or performing a minimal repair or doing nothing when a failure occurs. Jamali et al. (2005) proposed a joint optimal periodic and conditional maintenance strategy to improve the policy's efficiency. Lai (2007) optimized a periodic replacement model based on the information of cumulative repair cost limit.

All of the above studies are based on the assumption of constant costs. This assumption may not reflect market flexibility in practice. In (Sheu et al. 1995), the authors extended a replacement problem with two types of failures considering random repair costs that depend on the system age. However, the assumption about random costs is only valid when there are a lot of samples to perform statistical evaluations on probability measures. In many cases, there is not sufficient information and it is not easy to propose an adequate probability distribution for random parameters. Especially in the case of water pipe breakages, damage costs, such as water losses, social damage consequences, are uncertain. There are not enough samples to estimate a probability distribution. Then, consulting with the expert, the repair/replacement cost can be predicted by somewhere between a and c with a preference for b, where $a < b < c$. In this context, it is natural to extend the representation to using fuzzy quantities (Zimmermann 2010).

To the best of our knowledge, no previous work addresses the problem of fuzzy costs in block replacement with budget constraint and studies their impacts on the optimal policy. This paper is therefore aimed at filling this gap of the literature. On the other hand, our study is especially dedicated to a real case study of the replacement problem in water distribution networks. The paper is structured as follows. Section 2 presents the statement and formulation of the water pipe renovation problem. Section 3 is dedicated to identifying the optimal renovation time when considering fuzzy costs and budget constraints. In Section 4, we perform an experimental analysis to examine the impact of fuzzy costs on the optimal decision. Finally, Section 5 concludes the paper with prospects for future work.

2 PROBLEM DESCRIPTION AND FORMULATION

In this paper, we consider a block replacement policy in which the system is correctively repaired at failures and preventively replaced at periodic times, T years. We assume that the failure rate is not disturbed by each repair. This policy is applied to a District Metering Area (DMA), i.e., a small water distribution network in which pipes are subject to the same shocks and the same hydraulic pressure. In fact, the block replacement policy is mainly developed in the literature of the water resource domain (Shamir & Howard 1979, Kleiner & Rajani 1999, Kanakoudis & Tolikas 2001, Kleiner & Rajani 2010, Nguyen 2014), and is usually used by the Nha Be Water Supply Company[1]—a major water supply company for Districts 4, 7, and Nha Be of Ho Chi Minh City, Vietnam.

Let C_b be the cost associated with a pipe repair event and C_r be preventive renovation cost. The objective of block replacement is to determine the value of T in order to minimize the expected life cycle cost per unit time, $C(T)$:

$$C(T) = \frac{C_b \bar{N}(T) + C_r}{T} \tag{1}$$

where $\bar{N}(T)$ is the expected number of pipe-break repair event during the time period (0, T]. The details of $\bar{N}(T)$ will be described in the next section.

2.1 *Pipe break prediction model*

As pipe break prediction is one of the most important aspects of water network management, numerous studies have addressed this issue in the literature, see (Kleiner & Rajani 2001), for example, for an overview of statistical breakage models. Most pipe break prediction models are deduced from the maintenance records of pipe repairs data, thus they could be considered as pipe-break repair event models.

Shamir & Howard (1979), Kleiner & Rajani (1999) considered deterministic exponential and linear models for the annual breakages. Kanakoudis & Tolikas (2001) used an exponential expression of breakage numbers to perform economic analysis of the pipe replacement policy. In (Dandy & Engelhardt 2001), the authors performed a linear regression of breakage data to predict pipe breaks for the scheduling of pipe replacement with genetic algorithms. However, it is not easy to collect enough data required to establish a deterministic model. Therefore, probabilistic models have been proposed to model pipe breakage in water distribution networks. Among them, the Weibull-exponential distribution has been widely used in statistical models to describe the interval time between pipe installation and the first pipe break or between two successive breaks (Eisenbeis 1994, Mailhot et al. 2000). Le Gat & Eisenbeis (2000) used the Weibull proportional hazard model to characterize the distribution of times to failure and showed that short maintenance records (5–10 years) could give as good results as long maintenance records. Kleiner & Rajani (2010) developed a non-homogeneous Poisson model that allows us to take

[1] According to the Company's breakage repair reports.

into account pipe-dependent, time-dependent, and pipe-and time-dependent breakages, to represent the probability of breakage in individual water pipelines. Renaud et al. (2012) presented a break prediction tool, the “Casses” freeware, which is based on a counting process that relies not only on the pipe age and previous breakages, but also on the pipe’s characteristics and the environment.

Focusing on the impact of cost uncertainty on the periodic replacement optimization, we only consider in this paper a small DMA in which the environment has the same characteristics. Hence, we propose using a non-homogeneous counting process to model the pipe break repair in time. In detail, let the non-homogeneous Poisson process $\{N(t), t \geq 0\}$ characterize the number of pipe break repairs during the interval (0, t], the expected value $\bar{N}(T)$ is given by:

$$\bar{N}(t) = \int_0^t w(x)dx \tag{2}$$

where $w(x)$ is called the Rate Of Occurrence Of Failures (ROOF). Consider the increasing ROOF, $w(x)$ is assumed to follow:

Case 1: exponential expression,

$$w(x) = \beta \exp(\alpha x) \; 0 < \alpha, \beta < \infty; x \geq 0 \tag{3}$$

Case 2: Weibull expression,

$$w(x) = \alpha \, \beta x^{\beta - 1} \; 0 < \alpha, \beta < \infty; x \geq 0 \tag{4}$$

2.2 *Handling cost uncertainty with fuzzy numbers*

Depending on the available knowledge, the cost associated with a pipe breakage, C_b, and the preventive renovation cost, C_r, can be modeled by precise values or probability distributions or the most typical values. In reality, when only little knowledge is available, the costs can be predicted by somewhere between a and c, with a preference for b, of course $a < b < c$. In this context, it is preferred to extend the representation to using a Triangular Fuzzy Number (TFN), $\tilde{C}_b$ or $\tilde{C}_r$, (Zimmermann 2010).

Definition 1. *Fuzzy number*
Let X be a universal set, then a fuzzy number $\tilde{X}$ is a convex normalized fuzzy set $\tilde{X}$, defined by its membership function: $\mu_{\tilde{X}} : X \rightarrow [0,1]$, called the grade of membership of x in $\tilde{X}$. This membership function assigns a real number $\mu_{\tilde{X}}(x)$ in the interval [0,1] to each element x ∈ X.

The triangular fuzzy number is also noted by $\tilde{X} = (x_1, x_2, x_3)$ where $x_1, x_2, x_3 \in \mathbb{R}$ and $x_1 < x_2 < x_3$, the term x_2 is the most probable value of $\tilde{X}$ with $\mu_{\tilde{X}}(x_2) = 1$, the terms x_1 and x_3 are the lower and upper bounds of the possible area.

Definition 2. *Fuzzy measure*
Let r be a real number, the possibility, the necessity and the credibility of event ($\tilde{X} \leq k$) are respectively given by, (Liu & Liu 2002):

$$Pos(\tilde{X} \leq k) = \sup_{x \leq k} \mu_{\tilde{X}}(x),$$
$$Nec(\tilde{X} \leq k) = 1 - \sup_{x > k} \mu_{\tilde{X}}(x),$$
$$Cre(\tilde{X} \leq k) = \frac{1}{2}(Pos(\tilde{X} \leq k) + Nec(\tilde{X} \leq k))$$

Considering an example of a triangular fuzzy number, $\tilde{X}$ (x_1, x_2, x_3), its credibility is given by:

$$Cre(\tilde{X} \leq k) = \begin{cases} 0, & k < x_1 \\ \frac{1}{2}\left(\frac{x - x_1}{x_2 - x_1}\right), & x_1 \leq x \leq x_2 \\ \frac{x_3 - x}{x_3 - x_2}, & x_2 < x \leq x_3 \\ 0, & x > x_3 \end{cases} \tag{5}$$

Definition 3. *Fuzzy arithmetic*
Let $\tilde{Z} = f(\tilde{X}, \tilde{Y})$ denote the system characteristic of interest (e.g. steady state availability), evaluated by a function of fuzzy numbers $\tilde{X}, \tilde{Y}$ then $\tilde{Z}$ is also a fuzzy number. Following Zadeh’s extension principle (Zadeh 1965), the membership function of $\tilde{Z}$ is defined as:

$$\mu_{\tilde{Z}}(z) = \sup_{\tilde{X}, \tilde{Y}} \min\{\mu_{\tilde{X}}(x), \mu_{\tilde{Y}}(y) \mid z = f(x,y)\} \tag{6}$$

In practice, the a-cut method is developed to evaluate the membership function of $\tilde{Z} = f(\tilde{X}, \tilde{Y})$.

Definition 4. *α-cut set*
Given a fuzzy set $\tilde{X}$ in X and any real number α ∈ [0,1], *then the α-cut set of $\tilde{X}$, denoted by $\tilde{X}_\alpha$, is the crisp set: $\tilde{X}_\alpha = \{x \in X, \mu_{\tilde{X}}(x) \geq \alpha\}$.*

Definition 5. *Fuzzy arithmetic with α-cut set*
Let $\left[x_\alpha^L, x_\alpha^R\right]$ and $\left[y_\alpha^L, y_\alpha^R\right]$ are respectively the α*-cut interval of $\tilde{X}$ and $\tilde{Y}$ with the corresponding* α *value, then the α-cut interval of $\tilde{Z} = f(\tilde{X}, \tilde{Y})$ is defined as $\left[z_\alpha^L, z_\alpha^R\right]$ where:*

$$\begin{cases} z_\alpha^L = \min\{x \in [x_\alpha^L, x_\alpha^R], y \in [y_\alpha^L, y_\alpha^R]\} f(x,y) \\ z_\alpha^R = \min\{x \in [x_\alpha^L, x_\alpha^R], y \in [y_\alpha^L, y_\alpha^R]\} f(x,y) \end{cases} \tag{7}$$

Therefore, the membership function $\mu_{\tilde{Z}}(z)$ can be deduced by considering the lower bound and upper bound of the a-cuts of $\tilde{Z}$.

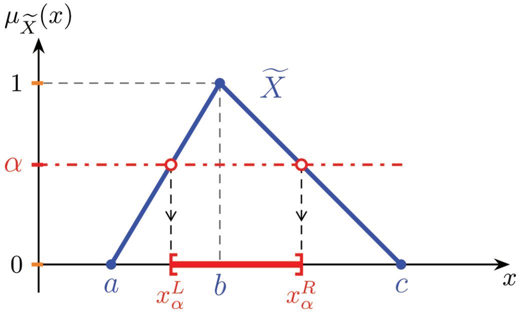


Figure 1. Illustration of the α-cut set of a triangular fuzzy number.

From Eq. (7), the following α-cuts of functions of positive fuzzy numbers can be easily derived:

$$\tilde{Z} = \tilde{X} + \tilde{Y}; \qquad \tilde{Z}_\alpha : \left[x_\alpha^L + y_\alpha^L, x_\alpha^R + y_\alpha^R\right] \tag{8}$$

$$\tilde{Z} = A\tilde{X} + B; \qquad \tilde{Z}_\alpha : \left[Ax_\alpha^L + B, Ax_\alpha^L + B\right] \tag{9}$$

where A and B are constant and A, $B > 0$.

2.3 *Handling annual budget constraint for fuzzy random costs*

As the number of pipe breaks $N(t)$ occurring during a certain time (0, t] is a random variable and the cost associated with every pipe break, $\tilde{C}_b$, is a fuzzy number, the accumulated pipe-break repair cost during a certain time, $N(t)\tilde{C}_b$, is a fuzzy discrete random variable.

Definition 6. *A fuzzy random variable is a random variable taking fuzzy values, (Liu 2001).*

Let (Ω, A, $\mathbb{P}$) *be a probability space of a discrete random variable N, and F be a collection of fuzzy variables* $\tilde{X}$. *A fuzzy discrete random variable, noted by* $\tilde{X}_N$ *is defined by a function from* Ω *to F, such that:*

$$\mu_{\tilde{X}_N}(x) = \begin{cases} \mu_{\tilde{X}_{n_1}}(x) & \textit{with probability } \mathbb{P}(N = n_1) \\ \mu_{\tilde{X}_{n2}}(x) & \textit{with probability } \mathbb{P}(N = n_2) \\ \dots & \\ \mu_{\tilde{X}_{n_m}}(x) & \textit{with probability } \mathbb{P}(N = n_m) \end{cases}$$

Definition 7. *Let* $\tilde{X}_N$ *be a fuzzy discrete random variable. Then the mean chance that fuzzy random event* $(\tilde{X}_N \geq B)$ *occurs is given by:*

$$Ch(\tilde{X}_N \geq B) = \sum_{n \in \Omega} Cre(\tilde{X}_N \geq B)\mathbb{P}(N = n) \tag{10}$$

In the case of a limited annual repair resource, a budget B is allocated to the pipe-break repair cost during a given strategic time unit (i.e., each year). Hence, the manager wants to handle the risk that the annual pipe-break repair cost exceeds the given budget B. That risk is a fuzzy random event, whose occurrence's mean chance should be lower than a real number r, i.e., $Ch(\tilde{X}_N \geq B) \leq r$. This is equivalent to:

$$Ch(\tilde{X}_N \leq B) \geq 1 - r \tag{11}$$

Note that, when:

- the pipe-break repair cost is a constant and the annual pipe break number is a random variable, the constraint in Eq. (11) is measured by:

  $$\mathbb{P}(X_N \leq B) \geq 1 - r,$$

  where X_N is a random variable representing the annual pipe-break repair cost.
- the pipe-break repair cost is a fuzzy variable and the pipe break number is characterized by the mean value $\bar{N}$, the constraint in Eq. (11) is measured by:

  $$Cre(\tilde{X}_{\bar{N}} \leq B) \geq 1 - r,$$

  where $\tilde{X}_{\bar{N}}$ is a fuzzy variable representing the fuzzy cost associated with the expected annual pipe break number.

3 IDENTIFYING THE OPTIMAL RENOVATION TIME

3.1 *Without budget constraint*

Let T_l (year) be the life time of the water pipe network, the objective of the problem is to determine the optimal renovation time T^* (year), where $T^* \in$ (0, T_l], that is:

$$\begin{aligned} T^* &= \arg_{T \in (0, T_l]} \min \tilde{C}(T) \\ &= \arg_{T \in (0, T_l]} \min \left(\frac{\tilde{C}_b \bar{N}(T) + \tilde{C}_r}{T} \right) \end{aligned} \tag{12}$$

where $\bar{N}(T)$ is evaluated corresponding to the two cases of the ROOF w(x) as follows:

$$\bar{N}(T) = \frac{\beta}{\alpha}(\exp(\alpha T) - 1) \tag{13}$$

$$\bar{N}(T) = \alpha T^{\beta} \tag{14}$$

As the Eq. (12) cannot be directly solved by the analytical approach, in this paper we propose using the grid search approach to find $T^* \in (0, T_l]$. We evaluate $\tilde{C}(T)$ for every $T \in (0:\frac{1}{2}:T_l]$, i.e. one month for each search step, and then compare them to find the minimum value. As the repair and/or renovation costs are characterized by TFNs, $\tilde{C}(T)$ is also a TFN compare TFNs, we propose using the *expected value* method. For a given TFN $\tilde{X} = (x_1, x_2, x_3)$, a typical model (Liu & Liu 2002) for defining its expected value E[X̃] is given by:

$$E[\tilde{X}] = \frac{1}{4}(x_1 + 2x_2 + x_3). \tag{15}$$

This expected value coincides with the neutral scalar substitute of a fuzzy interval (Yager 1981). The neutral scalar substitute is among the most natural defuzzification procedures proposed in the literature (Bortolan & Degani 1985). We have: $\tilde{X} \le \tilde{Y} \Leftrightarrow E[\tilde{X}] \le E[\tilde{Y}]$.

3.2 *With budget constraint*

As the ROOF is increasing with time, more pipe breaks occur as time goes by. Recall that T is the preventive renovation year and $[T]$ be the nearest integer less than or equal to T, then the worst situation may appear during:

- the previous year of the renovation time, ($[T]$ – 1, $[T]$).
- or the renovation year, ($[T]$, T).

Therefore, $N([T]$ – 1, $[T])$ and $N([T], T)$. spectively the pipe break numbers that occur during the previous year of the renovation time and the recent year of the renovation time. In order to handle the budget constraint, when performing the grid search, we evaluate the following mean chances corresponding to every T:

$$\begin{cases} Ch\left(N\left(\lfloor T \rfloor\right) - 1, \lfloor T \rfloor \tilde{C}_b \le B\right) \\ Ch\left(N\left(\lfloor T \rfloor T\right), \tilde{C}_b \le B\right) \end{cases} \tag{16}$$

If the above mean chance is higher than 1 – r, we then evaluate the corresponding life cycle cost per year, $\tilde{C}(T)$. Otherwise, this value of T is eliminated from the set of possible solutions of the grid search.

Let $\tilde{C}_b = (c_{b1}, c_{b2}, c_{b3})$ be a TFN, the mean chance Ch(-) is evaluated by Eq. (10), in which,

- the probability that the number of pipe breaks in the time interval (t, $t + v$] is n, denoted by $\mathbb{P}_{(t,t+v)}(n)$, is given by:

$$\begin{aligned} \mathbb{P}_{(t,t+v)}(n) &= \mathbb{P}(N + (t+v) - N(t) = n) \\ &= \left(\frac{\tilde{N}(t+v) - \tilde{N}(t)}{n!} \right)^{n} \exp\left(\bar{N}(t+v)\right) \end{aligned} \tag{17}$$

- corresponding to each integer n, from Eq. (5), the credibility, which is the repair cost associated with n pipe breaks, being lower than the budget constraint B, is given by:

$$Cre = \begin{cases} 1; & n \le \left\lfloor \dfrac{B}{c_{b_3}} \right\rfloor \\ 1 - \dfrac{1}{2}\left(\dfrac{nc_{b_3} - B}{nc_{b_3} - nc_{b_2}} \right); & \left\lfloor \dfrac{B}{c_{b_3}} \right\rfloor < n \le \left\lfloor \dfrac{B}{c_{b_2}} \right\rfloor \\ \dfrac{1}{2}\left(\dfrac{B - nc_{b_1}}{nc_{b_2} - nc_{b_1}} \right); & \left\lfloor \dfrac{B}{c_{b_2}} \right\rfloor < n \le \left\lfloor \dfrac{B}{c_{b_1}} \right\rfloor \\ 0, & n > \left\lfloor \dfrac{B}{c_{b_1}} \right\rfloor \end{cases} \tag{18}$$

4 EXPERIMENTAL ANALYSIS

4.1 *Parameter estimation for pipe repair event models*

We have processed the data obtained from the report on daily pipe break repair activities of the Nha Be Water Supply Company in the period from January 2008 to September 2015. From this data set, we specifically select two DMAs, namely Tan My Street and Cu xa Ngan Hang, to apply the proposed models. These DMAs have homogeneous pipes across the areas and, on the other hand, present sufficient data for the estimation. Their information is provided in Table 1. We see that most of the breakages occurred on branch pipes: 92.6% for DMA I and 98.7% for DMA II. The parameters α, β of the repair event models in Eqs. (3) and (4) for branch pipes are estimated as follows.

Let s_i and n be the occurrence time of the i-th pipe-break repair event and the number of repair events during the observed period T_o. Then the maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$ of α and β, respectively, in Eq. (3) are obtained by solving the following equations (Rausand & Høyland 2004):

Table 1. Data of the selected district metering areas (DMAs).

	DMA I Tan My Street		DMA II Cu xa Ngan Hang	
Name Pipe type	Main	Branch	Main	Branch
Material	uPVC	PE	uPVC	PE
Installation year	2000	2000	2003	2003
Total length (m)	6355	4610	4712	3025
Number of breaks	6	75	1	67

$$\sum_{i=1}^{n} s_i + \frac{n}{\hat{\alpha}} - \frac{nT_o}{1-\exp(-\hat{\alpha}T_o)} = 0 \tag{19}$$

$$\hat{\beta} = \frac{n\hat{\alpha}}{\exp(\hat{\alpha}T_o)-1} \tag{20}$$

Similarly, the maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$ of α and β, respectively, in Eq. (4) are given by:

$$\hat{\beta} = \frac{n}{n\ln T_o - \sum_{i=1}^{n} \ln s_i} \tag{21}$$

$$\hat{\alpha} = \frac{n}{T_o^{\hat{\beta}}} \tag{22}$$

Considering Figure 2, we observe that both models (the exponential and Weibull ROOF) are appropriate for the repair data of both DMAs from January 2008 to September 2016. Among them, the coefficient of determination R^2 of the Weibull model is higher than that of the exponential model. Hereafter, the Weibull model is chosen to characterize the counting process of pipe-repair events for both DMAs.

4.2 *Impact of fuzzy costs on the optimal renovation time without budget constraint*

The pipe repair or replacement cost depends on the the pipe material/diameter and especially on the road types such as alley or road/route, asphalt or dirt road. In addition, the variation of the pipe-break detection time and maintenance time can lead to different damage cost, including water loss, disruption in service, and so on. Therefore, it is difficult to expect a precise value of the renovation cost for the overall DMA or of the cost associated with only pipe repair activities. Classical approaches normally use the most probable value in the calculation and optimization. In this paper, we employ TFNs, $\tilde{C}_r = (c_{r_1}, c_{r_2}, c_{r_3})$ and $\tilde{C}_b = (c_{b_1}, c_{b_2}, c_{b_3})$ to solve the problem.

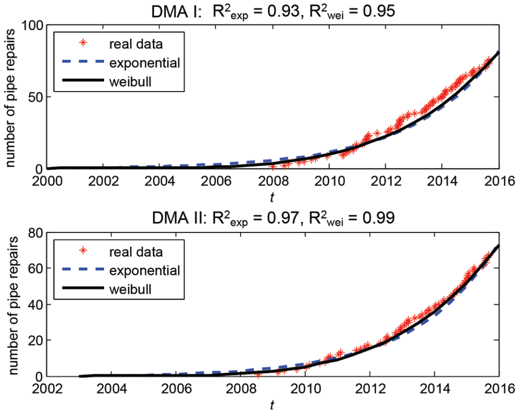


Figure 2. Accumulated number of pipe repair events from January 2008 to September 2016.

Table 2. Values of fuzzy costs.

$\tilde{C}_b$($)	$\tilde{C}_r$($)
(232 525 734)	DMA I: (300000, 375000, 450000) DMA II: (300000, 320000, 450000)

On the other hand, the risk that the annual cost associated with repair events exceeds $20,000 is recommended to be lower than 10%. In this section, we will study how fuzzy costs impact on the optimal decision in the case without this budget constraint and in the case with this budget constraint.

4.2.1 *Optimal renovation time*

The detailed values of the fuzzy costs are presented in Table 2. As it is generally recommended that the life time of an uPVC main pipe should not exceed 25 years, we run the grid search over the interval (0, 25) years with the step of one month to find the optimal renovation time for both DMAs. Four cases will be examined:

- Case A: not considering fuzzy costs and budget constraint
- Case B: considering fuzzy costs but not taking into account budget constraint
- Case C: not considering fuzzy costs but taking into account budget constraint
- Case D: considering both fuzzy costs and budget constraint

Table 3 presents the optimal renovation time corresponding to the above cases for both DMAs. In detail, if we do not take into account the budget constraint, the optimal renovation time in Case B

Table 3. Optimal renovation time (years after installation).

DMA	Case A	Case B	Case C	Case D
I	19	19.5	18.83	17.83
II	16.83	17	14.92	13.92

is longer than that of Case A. Indeed, with fuzzy cost, the renovation time of DMA I is postponed six months from the renovation time of 19 years when using precise values.

However, if we consider the budget constraint, the renovation time when using fuzzy numbers is earlier than that with only the most probable value used. For instance, the optimal decisions obtained say that DMA I needs to be renovated after 14 years and 11 months from its installation. This renovation time is one year earlier if we evaluate with fuzzy numbers.

In the next part, we will focus on DMA I to deeply study the impact of fuzzy costs on the optimal decision. We also assess whether the manager can obtain the real benefit of taking fuzzy costs into account.

4.2.2 *Improvement factor when using fuzzy numbers*

In this section, we will present a parameter, called improvement factor in order to consider if using fuzzy numbers brings back more benefit or not.

Firstly, we find the renovation time for the case with or without budget constraint, using the most probable value or fuzzy number $(T_i^*, i \in A, B, C, D)$.

Then, we sample 1000 values of the cost, which is predicted in the interval [*a*, *c*] with the most probable value *b*, and evaluate the optimal cycle life cost per year $C(T_i^*)$, with *i* ∈ *A*, *B*, *C*, *D*, that is corresponding to every sample of the cost.

Finally the improvement factor $f_{(i,j)}$, which is corresponding to every cost sample, is evaluated by following equation:

$$f_{(i,j)} = \frac{C(T_i^*) - C(T_j^*)}{C(T_i^*)}$$

where $C(T)$ is evaluated by Eq. (1), T_i^* is the optimal renovation time corresponding to case *i*, *i*, *j* ∈ {*A*, *B*, *C*, *D*}.

Impact of the most probable value of $\tilde{C}_b$:
The renovation cost is assumed to be a precise value. The cost associated with a pipe break event is characterized by a TFN $(c_{b_1}, c_{b2}, c_{b3})$ where c_{b_1} and c_{b_3} are fixed while c_{b2} varies from c_{b_1} to c_{b_3}. Figure 3 presents the impact of the position of the most probable value in the possible area of $\tilde{C}_b$ on the optimal renovation time. As $\tilde{C}_{b_2}$ increases when c_{b2} goes up, the optimal renovation time is therefore accelerated in all four Cases A, B, C, and D.

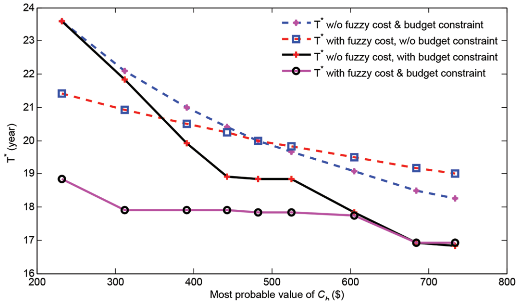
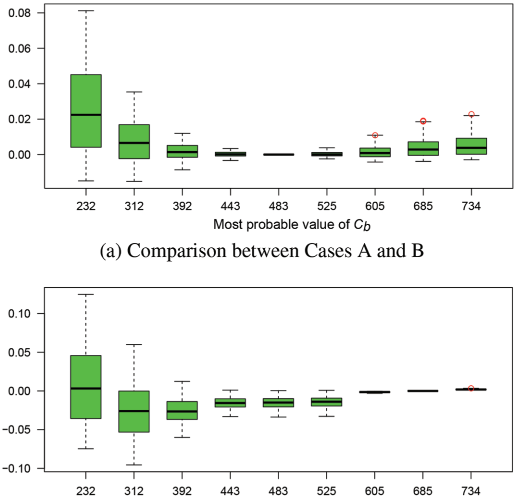


Figure 3. Impact of C_{b_2} on the optimal renovation time (T^*).



(a) Comparison between Cases A and B

(b) Comparison between Cases C and D

Figure 4. Distribution of improvement factor according to C_{b_2}.

When the budget constraint is not considered, we find that the renovation time of Case B (fuzzy cost), is sooner than the one of Case A (precise cost) if $c_{b_2} < (c_{b_1} + c_{b_3})/2$. On the contrary, the renovation time is postponed when considering the fuzzy cost if $c_{b_2} > (c_{b_1} + c_{b_3})/2$. They are coincident if $c_{b_2} = (c_{b_2} + c_{b_2})/2$.

These T^* adjustments help the manager reduce the annual cycle life cost. Indeed, considering Figure 4a, the distribution of the improvement factor $f_{(A,B)}$ shows that using fuzzy numbers helps us save 0 to 2.5% or 0 to 10% (on average) of the

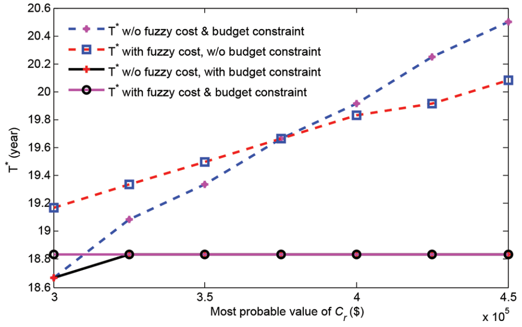


Figure 5: Impact of C_{r_2} on the optimal renovation time (T^*).

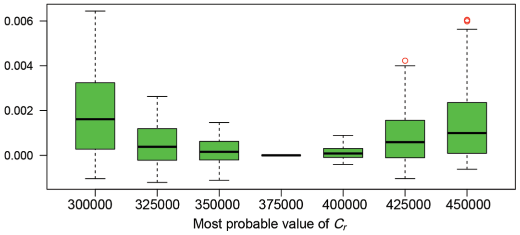


Figure 6. Distribution of improvement factor according to C_{r_2} when do not consider budget constraint.

expected annual life cycle cost when the most probable value moves from the middle point to the left end point or the right end point of the possible area.

When considering the budget constraint, we find that the renovation time of Case D (fuzzy cost), is sooner than the one of Case C (precise cost) if c_{b_2} < 600 $. The gap is larger when the most probable value c_{b_2} approaches to the left end of the possible value interval. When c_{b_2} > 600, the optimal renovation time in Cases C and D are almost same. If the right end of the possible value interval is also the most probable value, the renovation time when considering fuzzy cost is slightly postponed compared to the one of the case without fuzzy cost.

These T^* adjustments help the manager reduce the risk that the annual cost associated with pipe repair events exceeds the budget B. However, the annual life cycle cost slightly increases when using fuzzy number in this case (Figure 4b). It is necessary to balance the satisfaction degree of the budget constraint and the reduction in the annual life cycle cost.

Impact of the most probable value of $\tilde{C}_r$:
Considering a precise value of the cost associated with a pipe break event, the renovation cost is characterized by a TFN $(c_{r_1}, c_{r_2}, c_{r_3})$ where c_{r_1} and c_{r_3} are fixed while c_{r_2} varies. Figure 5 presents the impact of the position of the most probable value in the possible area of $\tilde{C}_r$ on the optimal renovation time. When c_{r_2} goes up, the optimal renovation time is postponed in all of the four cases A, B, C, and D because the corresponding $\tilde{C}_{r_2}$ is increasing.

When the budget constraint is not considered, we find that the renovation time of Case B (fuzzy number), is later than the one of Case A (precise cost) if $c_{r_2} < (c_{r_1} + c_{r_3})/2$. On the contrary, the renovation time is accelerated when considering the fuzzy cost where $c_{b_2} > (c_{b_1} + c_{b_3})/2$. They are coincident if $c_{b_2} = (c_{b_1} + c_{b_3})/2$. These adjustments of the renovation time help the manager reduce the annual cycle life cost. Indeed, considering the fuzzy number $\tilde{C}_r$, the manager can save 0 to 0.2%, or 0 to 0.15% (on average) of the expected annual life cycle cost when the most probable value changes from the middle point to the left end point and the right end point of the possible area (Figure 6).

Considering the budget constraint, as only the cost associated with a pipe break, C_b, affects on the risk of violating the budget constraint, we find that the renovation time of Case D (fuzzy cost) is equal to the one of Case C (precise cost) in most cases. Only for the case where $c_{r_2} = c_{r_1} = \$300{,}000$, the renovation time is sooner. However, it is not caused by the impact of fuzzy numbers but by the description of the optimal decision in the case without a budget constraint.

5 CONCLUSIONS

We have examined an optimal periodic replacement policy in which repair and replacement costs are not precise. The model was used to optimize the strategy for water pipe renovation. First, a pipe-break repair event model was constructed from real maintenance records of the Nha Be Water Supply Company. Based on this model, we considered the impact of fuzzy costs on the optimal decision and also highlighted the real benefit of using fuzzy numbers. It was shown that, without budget constraints, the use of fuzzy numbers helps reduce the life cycle cost per year. When budget constraints are taken into account, it is necessary to weigh the degree of satisfying budget constraints against the augmentation of annual life cycle costs.

In future work, hydraulic constraints may be considered in the optimization of maintenance policies for water pipe networks. Moreover, the impact of the spread of fuzzy numbers, i.e. degree of cost uncertainty, on the optimal decision will be studied. In that case, a distance-based ranking method, such as the Bertoluzza measure, can be used to compare fuzzy life cycle costs in order to find the optimal renovation time.

REFERENCES

Ahmad, R. & Kamaruddin, S. (2012). An overview of time-based and condition-based maintenance in industrial application. *Computers & Industrial Engineering* 63(1), 135–149.

Barlow, R. & Hunter, L. (1960). Optimum preventive maintenance policies. *Operations Research* 8(1), 90–100.

Bortolan, G. & Degani, R. (1985). A review of some methods for ranking fuzzy subsets. *Fuzzy Sets and Systems* 15(1), 1–19.

Dandy, G.C. & Engelhardt, M. (2001). Optimal scheduling of water pipe replacement using genetic algorithms. *Journal of Water Resources Planning and Management* 127(4), 214–223.

Eisenbeis, P. (1994). *Modelisation statistique de la prevision des defaillances sur les conduites d'eau potable*. Ph.D. thesis, Louis-Pasteur, Strasbourg, France.

Jamali, M., Ait-Kadi, D., Cléroux, R., & Artiba, A. (2005). Joint optimal periodic and conditional maintenance strategy. *Journal of Quality in Maintenance Engineering* 11(2), 107–114.

Kanakoudis, V.K. & Tolikas, D.K. (2001). The role of leaks and breaks in water networks: technical and economical solutions. *Journal of Water Supply: Research and Technology* 50(5), 301–311.

Kleiner, Y. & Rajani, B. (1999). Using limited data to assess future needs. *Journal - American Water Works Association* 91(7), 47–61.

Kleiner, Y. & Rajani, B. (2001). Comprehensive review of structural deterioration of water mains: statistical models. *Urban Water* 3(3), 131–150.

Kleiner, Y. & Rajani, B. (2010). I-WARP: Individual water main renewal planner. *Drinking Water Engineering and Science* 3(1), 71–77.

Lai, M.-T. & Chen, Y.C. (2006). Optimal periodic replacement policy for a two-unit system with failure rate interaction. *The International Journal of Advanced Manufacturing Technology* 29(3–4), 367–371.

Lai, M.-T. (2007). A periodical replacement model based on cumulative repair-cost limit. *Applied Stochastic Models in Business and Industry* 23(6), 455–464.

Le Gat, Y. & Eisenbeis, P. (2000). Using maintenance records to forecast failures in water networks. *Urban Water* 2(3), 173–181.

Liu, B. & Liu, Y.-K. (2002). Expected value of fuzzy variable and fuzzy expected value models. *IEEE Transactions on Fuzzy Systems* 10(4), 445–450.

Liu, B. (2001). Fuzzy random chance-constrained programming. *IEEE Transactions on Fuzzy Systems* 9(5), 713–720.

Mailhot, A., Pelletier, G., Noel, J.F., & Villeneuve, J.P. (2000). Modeling the evolution of the structural state of water pipe networks with brief recorded pipe break histories: Methodology and application. *Water Resources Research* 10(36), 3053–3062.

Nguyen, B.M. (2014). Optimal pipe replacement strategy based on the cost function in water distribution system using MAT- LAB programming. In *The 10th International Conference on Sustainable Water Environment*, Seoul, South Korea.

Rausand, M. & Hϕyland, A. (2004). *System Reliability Theory: Models, Statistical Methods, and Applications* (2nded.). Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley.

Renaud, E., Gat, Y.L., & Poulton, M. (2012). Using a break prediction model for drinking water networks asset management: From research to practice. *Water Science and Technology: Water Supply* 12(5), 674–682.

Scarf, P.A. & Cavalcante, C.A.V. (2010). Hybrid block replacement and inspection policies for a multi-component system with heterogeneous component lives. *European Journal of Operational Research* 206(2), 384–394.

Shamir, U. & Howard, C. (1979). An analytical approach to scheduling pipe replacement. *American Water Works Association* 71(5), 248–258.

Sheu, S.H., Chen, Y.L., Chang, C.C., & Zhang, Z.G. (2012). A block replacement policy for systems subject to non- homogeneous pure birth shocks. *IEEE Transactions on Reliability* 61(3), 741–748.

Sheu, S., Griffith, W.S., & Nakagawa, T. (1995). Extended optimal replacement model with random minimal repair costs. *European Journal of Operation Research* 85(3), 636–649.

Sheu, S.H. & Griffith, W.S. (2002). Extended block replacement policy with shock models and used items. *European Journal of Operational Research* 140(1), 50–60.

Sheu, S.H. (1992). Optimal block replacement policies with multiple choice at failure. *Journal of Applied Probability* 29(1), 129–141.

Sheu, S.H., Chang, C.C., Chen, Y.L., & Zhang, Z. (2010). A periodic replacement model based on cumulative repair-cost limit for a system subjected to shocks. *IEEE Transactions on Reliability* 59(2), 374–382.

Wang, H. & Pham, H. (2006). Availability and maintenance of series systems subject to imperfect repair and correlated failure and repair. *European Journal of Operational Research* 174(3), 1706–1722.

Wang, H. (2002). A survey of maintenance policies of deteriorating systems. *European Journal of Operational Research* 139(3), 469–489.

Yager, R.R. (1981). A procedure for ordering fuzzy subsets of the unit interval. *Information Sciences* 24(2), 143–161.

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control* 8(3), 338–353.

Zimmermann, H.-J. (2010). Fuzzy set theory. *Wiley Interdisciplinary Reviews: Computational Statistics* 2(3), 317–332.

This page intentionally left blank

Monte Carlo methods for parallel computing of reliability and risk

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Acceleration of multi-factor Merton model Monte Carlo simulation via Importance Sampling and GPU parallelization

M. Béreš & R. Briš
VŠB—Technical University of Ostrava, Ostrava, Czech Republic

ABSTRACT: Credit risk refers to the risk of losses due to unexpected credit events, as a default of a counterparty. The modelling and controlling of credit risk is a very important topic within banks. Very popular and frequently used tools for modelling credit risk are multi-factor Merton models. Practical implementation of these models requires time-consuming Monte Carlo (MC) simulations, which significantly limits their usability in daily credit risk calculation. In this paper we present acceleration techniques of Merton model Monte Carlo simulations, concretely parallel GPU implementation and Importance Sampling (IS) employment. As the importance sampling distribution we choose the Gaussian mixture model and for calculating the IS shifted probability distribution we use the Cross-Entropy (CE) method. The speed-up results are demonstrated using portfolio Value at Risk (VaR) and Expected Shortfall (ES) calculation.

1 INTRODUCTION

In this paper we present a new approach to the Importance Sampling (IS) in the multi-factor Merton model. In the standard IS approach the normal distribution is used as a family of the IS distributions. This approach results in a decent variance reduction but a certain level of degeneracy of probability can be observed. The observed degeneracy of probability is caused by a relatively high difference between the IS distribution chosen from the normal distribution family and the optimal IS distribution and it also limits the achievable variance reduction. As a correction to this problem we use the Gaussian mixture model for the IS family of distributions. This new approach limits the level of the observed degeneracy of probability as well as increases the variance reduction.

The other significant part of this paper is the implementation of discussed models and IS procedures via CUDA on the the GPU devices. The GPU implementation of the model enables very fast calculation of the observed parameters (VaR or ES) with or without the use of the IS.

First we present a short recapitulation of the multi-factor Merton model and the terminology used, then we state a detailed specification of the tested model. For a deeper understanding of the Merton model see (Lütkebohmert 2008).

1.1 *Briefly about multi-factor Merton model*

Let assume we have a portfolio of N risky loans (*exposures*) indexed by $n = 1, \ldots, N$. We are interested in the possible defaults, which can occur in the fixed time interval $[0,T]$. Let D_n denote the default indicator of an exposure *n*, which can be represented as a Bernoulli random variable taking the values

$$D_n = \begin{cases} 1, & \text{if the exposure } n \text{ is in the default} \\ 0, & \text{otherwise} \end{cases}. \tag{1}$$

We assume that the probabilities $PD_n = \mathbb{P}(D_n = 1)$ are given as a portfolio parameter.

The portion of the exposure *n* which can be lost in the time of default is called the exposure at default denoted by EAD_n. For simplicity we assume EAD_n is constant in the whole time interval $[0,T]$ and it is given as the portfolio parameter.

The portion of EAD_n representing the real loss in the case of default, is given by a random variable *loss given at default* $LGD_n \in [0,1]$. The distribution, the expectation $ELGD_n$ and the standard deviation $VLGD_n$ of LGD_n are given as the portfolio parameters. The portfolio loss L_N is than defined as a random variable

$$L_N = \sum_{n=1}^{N} EAD_n \cdot LGD_n \cdot D_n. \tag{2}$$

Now we can define the value at risk (*VaR*) as *p* quantile (or *confidence level*) of L_N

$$\begin{aligned} VaR_p(L_N) &= \inf\{x \in \mathbb{R} : \mathbb{P}(L_N > x) \le 1-p\} \\ &= \inf\{x \in \mathbb{R} : F_{L_N}(x) \ge p\}, \end{aligned} \tag{3}$$

where $F_{L_N}(x)$ is the cumulative distribution function of L_N. And the Expected Shortfall (*ES*) as a conditional tail expectation with the condition $x \ge VaR_p(L_N)$

$$\begin{aligned} ES_p(L_N) &= \frac{1}{1-p}\int_{VaR_p(L_N)}^{\infty} x\,\mathbb{P}(L_N = x)dx \\ &= \frac{1}{1-p}\int_p^1 VaR_u(L_N)du. \end{aligned} \tag{4}$$

1.1.1 *Exposure correlation factors*

In the reasonable portfolio the single exposure's defaults are correlated, let us outline, how the correlation is handled in the Merton model. We assume that every exposure has a unique owner (obligor). Let $V_n(t)$ denote *n*-th obligor's assets, $S_n(t)$ obligor *n* equity and $B_n(t)$ obligor *n* bond, so

$$V_n(t) = S_n(t) + B_n(t), 0 \le t \le T. \tag{5}$$

In the Merton model a default can occur only at the maturity *T*, which leads into two possibilities

1. $V_n(T) > B_n(T)$: obligor has sufficient asset to fulfil debt, $D_n = 0$.
2. $V_n(T) \le B_n(T)$: obligor cannot fulfil debt and defaults, $D_n = 1$.

Let r_n denote the *n*-th obligor's asset-value log-return $r_n = \log(V_n(T)/V_n(0))$. The multi-factor Merton model assumptions to resolve correlations between exposure defaults are:

1. r_n depends linearly on K standard normally distributed risk (*systemic*) factors $X = (X_1, \ldots, X_K)$.
2. r_n depends linearly on the standard normally distributed idiosyncratic term ε_n, which is independent of the systemic factors X_k.
3. single idiosyncratic factors ε_n are uncorrelated.
4. asset-value log-return random variable can be represented as $r_n = \beta_n \cdot Y_n + \sqrt{1-\beta_n^2} \cdot \varepsilon_n$, where $Y_n = \sum_{k=1}^{K} \alpha_{n,k} X_k$ represents exposure composite factor, β_n represents exposure sensitivity to systemic risk and weights $\alpha_{n,k}$ represents dependence on single factors X_k.
5. r_n has standard normal distribution if condition $\sum_{k=1}^{K} \alpha_{n,k}^2 = 1$ is satisfied.

Variables $\alpha_{n,k}$ and β_n are assumed as a given portfolio parameters.

When PD_n is given and r_n has the standard normal distribution, one can calculate threshold $c_n = \Phi^{-1}(1-PD_n)$ so default indicator can be represented as

$$D_n = r_n > c_n. \tag{6}$$

1.1.2 *Monte Carlo simulation of multi-factor Merton model*

With previous knowledge and full portfolio specification we can now approximate the portfolio *VaR* and *ES* via the Monte Carlo simulations. Single exposure defaults can be directly calculated from the systemic and the idiosyncratic shocks $X_k^{(i)}$ and $\varepsilon_n^{(i)}$ drawn from the standard normal distribution $N(0,1)$, upper index (*i*) indicate index of the Monte Carlo sample. With the generated random $LGD_n^{(i)}$ we can calculate the total random scenario loss

$$L_N^{(i)} = \sum_{n=1}^{N} EAD_n \cdot LGD_n^{(i)} \cdot D_n^{(i)}. \tag{7}$$

The Monte Carlo simulation consisting of *M* trials approximate portfolio *VaR* as

$$\begin{aligned} \overline{VaR_p}(L_N) &= \min\left\{L_N^{(i)} : \psi(L_N^{(i)}) \le (1-p)\cdot M\right\} \\ &= L_N^{[\lceil M\cdot p\rceil]}, \end{aligned} \tag{8}$$

where $\psi\left(L_N^{(i)}\right) = \sum_{j=1}^{M}(L_N^{(j)} > L_N^{(i)})$, $L_N^{[j]}$ is the *j*-th loss in the ascendant sorted loss sequence $L_N^{(i)}$, and *ES* as

$$\overline{ES_p}(L_N) = \frac{1}{M - \lceil M\cdot p\rceil} \cdot \sum_{j=\lceil M\cdot p\rceil}^{M} L_N^{[j]}. \tag{9}$$

1.2 *Tested portfolio structure specification*

The most important part of the multi-factor Merton model is the structure of the portfolio (exposure dependence on the risk factors). To obtain a portfolio with a realistic behaviour we use a natural risk factor construction considering the region-industry (*sector*) and the direct (*hierarchy*) links between exposures.

Hierarchy links are represented by Hierarchy Systemic Factors (HSF), which can be interpreted as direct links between the exposures (for example two subsidiary companies with a common parent company), each of these systemic factors usually has impact only on a small fraction of the portfolio exposures. Sector links are represented by

Sector Systemicfactors (SSF), which can be interpreted as industrial and regional factors, each of these systemic factors usually impacts majority of the portfolio exposures. Therefore every exposure's asset-value log-return random variable r_n depends on two composite factors H_n (hierarchy composite factor) and S_n (sector composite factor) according to following formula:

$$\overline{r_n} = g_n \cdot H_n + \sqrt{1-g_n^2} \cdot \varepsilon_n, \tag{10}$$

$$r_n = \sqrt{1-\omega_n^2} \cdot S_n + \omega_n \cdot \overline{r_n}, \tag{11}$$

where H_n is composite factor of hierarchy correlation risk factors (HSF), $g_n \in (0,1)$ is group correlation coefficient with composite HSF, S_n is composite factor of sector correlation risk factors (SSF), $\omega_n \in (0,1)$ is idiosyncratic weight towards composite SSF and ε_n is exposure idiosyncratic factor.

Let K_S denote the number of SSF and K_H denote the number of HSF. We assume that, there are corresponding K_S sector composite factors and K_H hierarchy composite factors. Links (correlation) between single composite factors are represented differently for HSF and SSF.

In the case of HSF we assume links between systemic factors take form of a dependence tree structure. Let $H_{(1)},\ldots,H_{(K_H)}$ denote the unique composite factors of HSF corresponding to K_H. Composite factors are ordered according to a given tree structure and their calculation is given recursively, where every node $H_{(k)}$ has at most one parent $H_{(l)}$ and specified correlation coefficient g_k^H, see formula (12).

$$H_{(k)} = \begin{cases} g_k^H H_{(l)} + \sqrt{1-(g_k^H)^2}\varepsilon_k^H, & p(k) = l \\ \varepsilon_k^H & p(k) = \varnothing \end{cases}, \tag{12}$$

where ε_k^H denotes idiosyncratic term for HSF k and $p(k)$ is parent mapping function. Example of calculating HSF composite factors can be seen in Figure 1.

In the case of SSF we assume links between systemic factors take form of the full correlation matrix. Let $S_{(1)},\ldots,S_{(K_S)}$ denote unique composite factors of SSF. Single composite factors $S_{(1)},\ldots,S_{(K_S)}$ are defined by a given correlation matrix Σ and are calculated as

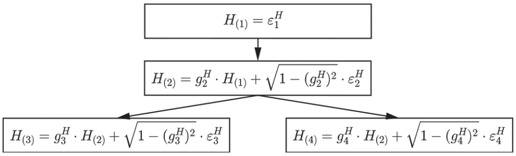

$H_{(1)} = \varepsilon_1^H$
$H_{(2)} = g_2^H \cdot H_{(1)} + \sqrt{1-(g_2^H)^2} \cdot \varepsilon_2^H$
$H_{(3)} = g_3^H \cdot H_{(2)} + \sqrt{1-(g_3^H)^2} \cdot \varepsilon_3^H$
$H_{(4)} = g_4^H \cdot H_{(2)} + \sqrt{1-(g_4^H)^2} \cdot \varepsilon_4^H$

Figure 1. Example of group correlation tree.

$$\begin{bmatrix} S_{(1)} \\ \vdots \\ S_{(K_S)} \end{bmatrix} = \sqrt{\Sigma} \cdot \begin{bmatrix} \varepsilon_1^S \\ \vdots \\ \varepsilon_{K_S}^S \end{bmatrix}, \tag{13}$$

where $\varepsilon_k^S s$ denote idiosyncratic term for SSF k.

All of the aforementioned parameters g_n, ω_n, g_k^H, HSF tree structure and correlation matrix Σ are given as portfolio parameters and can be interpreted in the standard form of $\alpha_{n,k}$ and β_n parameters, where $\sum_{k=1}^{K} \alpha_{n,k}^2 = 1$ is satisfied.

For tested model the LGDs are considered from the Beta distribution with mean and standard deviation given by portfolio parameters.

For a better illustration in the further text we will use normalized EAD_n:

$$\sum_{k=1}^{N} EAD_k = 1, \tag{14}$$

which express EAD_n as a portion of the total portfolio exposure.

2 EMPLOYING IMPORTANCE SAMPLING

As mentioned before, we are interested in the VaR and the ES of the observed portfolio loss random variable L_N. The Monte Carlo approximation of these values is highly sensitive to the stated *confidence level p*, which is usually very close to 1. In our study we use the confidence levels of 0.99995, 0.9995 and 0.995. For example when the confidence level is 0.99995 the MC simulation of 10^6 samples provides only 50 samples with the information about VaR/ES.

One of the straightforward ways to increase the number of samples in the region of VaR/ES calculation is to change the distribution of the portfolio loss random variable so called the Importance Sampling (IS) method. The principle of the IS can be easily demonstrated on the ES calculation. The ES can be represented as the conditional mean or mean of the specific function

$$H_p(x) = \begin{cases} 0, & x < VaR_p(L_N) \\ \frac{x}{1-p}, & x \geq VaR_p(L_N) \end{cases} \tag{15}$$

$$\begin{aligned} ES_p(L_N) &= \mathbb{E}_f(H_p(L_N)) \\ &= \int_{\Omega} H_p(L_N^*(\overline{y})) \cdot f(\overline{y}) d\overline{y}, \end{aligned} \tag{16}$$

where $\overline{y}$ are values of the random vector $\overline{Y}$ of all random variables contributing to L_N (idiosyncratic terms, LGDs), Ω is the set of the all possible values of $\overline{y}, f(\overline{y})$ is the joint probability density function of $\overline{Y}$, $L_N^*(\overline{y}) : L_N^*(\overline{Y}) = L_N$ is the function mapping $\overline{y}$ to corresponding value of L_N and $\mathbb{E}_f$ is mean under the pdf $f(\overline{y})$. If we use the IS with the new probability distribution of L_N given by pdf $g(\overline{y})$ we can calculate original ES as

$$\begin{aligned}&\mathbb{E}_g\left(H_p\left(L_N^*(\overline{Y})\right)\cdot\frac{f(\overline{Y})}{g(\overline{Y})}\right)\\&=\int_\Omega H_p\left(L_N^*(\overline{y})\right)\cdot\frac{f(\overline{y})}{g(\overline{y})}\cdot g(\overline{y})d\overline{y}\\&=\mathbb{E}_f(H_p(L_N))=ES_p(L_N).\end{aligned}\tag{17}$$

The ratio of probability density functions $w(\overline{y}) := \frac{f(\overline{y})}{g(\overline{y})}$ is called the *the likelihood ratio* (LR). From formula (17) we can see the natural requirement on $g(\overline{y}) : H_p\left(L_N^*(\overline{y})\right)\cdot f(\overline{y}) > 0 \Rightarrow g(\overline{y}) > 0$. Formula (17) also provide the MC estimation of the ES when using the IS

$$\begin{aligned}\overline{ES_p^g}(L_N)&=\frac{1}{N}\sum_{i=1}^{M}H_p\left(L_N^*\left(\overline{Y_i}\right)\right)\cdot w(\overline{Y_i})\\&=\frac{\sum_{i=1}^{M}L_N^*(\overline{Y_i})\left(L_N^*(\overline{Y_i})\geq\overline{VaR_p^g}(L_N)\right)w(\overline{Y_i})}{M\cdot(1-p)},\end{aligned}\tag{18}$$

where $\overline{Y_i}$ is i-th sample of $\overline{Y} \sim g(\overline{y})$ and M is the number of random samples. It remains to define $\overline{VaR_p^g}(L_N)$ as

$$\overline{VaR_p^g}(L_N)=\min\left\{L_N^*(\overline{Y_i}) : \psi\left(\overline{Y_i}\right)\leq(1-p)\cdot M\right\},\tag{19}$$

where $\psi\left(\overline{Y_i}\right) := \sum_{j=1}^{M}\left(L_N(\overline{Y_i}) > L_N^*(\overline{Y_j})\right)\cdot w(\overline{Y_j})$.

2.1 *Cross-Entropy method*

We already know the principles of the IS and have the IS estimators of VaR and ES, but a new IS pdf $g(\overline{y})$ is still unknown. The most straightforward method for the estimation of $g(\overline{y})$ is to minimize the variance of the ES IS estimator:

$$g(\overline{y})=\underset{v(\overline{y})\in X}{\arg\min}\left\{S_v^2\left(H\left(L_N^*\left(\overline{Y}\right)\right)\frac{f\left(\overline{Y}\right)}{v\left(\overline{Y}\right)}\right)\right\},\tag{20}$$

where $S_v^2(X)$ denote variance according to pdf $v(\overline{y})$ and X is an arbitrary system of the pdfs fulfilling the condition $v(\overline{y}) : H_p(L_N^*(\overline{y}))\cdot f(\overline{y}) > 0 \Rightarrow v(\overline{y}) > 0$. This approach is called *variance minimization* (VM) method. Usually the VM method leads to very difficult problems, which have to be solved numerically.

Another approach to obtain the IS pdf $g(\overline{y})$ is the Cross-Entropy (CE) method. The CE method similarly to the VM method solve a minimization problem, but instead of minimizing the variance it minimize the Kullback-Leibler (KL) divergence $D(g^*, v)$ with the optimal (zero variance) IS distribution

$$g^*(\overline{y})=\frac{\left|H_p\left(L_N^*(\overline{y})\right)\right|\cdot f(\overline{y})}{\mathbb{E}_f\left(\left|H_p(L_N)\right|\right)}\tag{21}$$

$$\begin{aligned}g(\overline{y})&:=\underset{v(\overline{y})\in X}{\arg\min}\left\{D(g^*(\overline{y}),v(\overline{y}))\right\}\\&=\underset{v(\overline{y})\in X}{\arg\min}\left\{\int_\Omega g^*(\overline{y})\ln\frac{g^*(\overline{y})}{v(\overline{y})}d\overline{y}\right\}\\&=\underset{v(\overline{y})\in X}{\arg\min}\left\{\int_\Omega\left|H_p(L_N^*(\overline{y}))\right|f(\overline{y})\ln v(\overline{y})d\overline{y}\right\}.\end{aligned}\tag{22}$$

To obtain a solvable problem, we need to add some constrain to the system of pdfs X. Usual choice is a parametrized family of pdfs:

$$X := \left\{v\left(\overline{x};\theta\right)\forall\theta\in\Theta\right\},\tag{23}$$

where $v\left(\overline{x};\theta\right)$ is pdf taking vector of parameters θ and $\Theta := \{\theta : H_p(L_N^*(\overline{y}))\cdot f(\overline{y}) > 0 \Rightarrow v(\overline{y};\theta) > 0\}$. Obtained minimization problem is usually concave, therefore we can replace the optimization problem with the following equation

$$\theta : \int_\Omega\left|H_p(L_N^*(\overline{y}))\right|f(\overline{y})\nabla_\theta\ln v(\overline{y};\theta)d\overline{y}=0.\tag{24}$$

To solve the problem (24) we use the Monte Carlo simulation:

$$\theta : \sum_{i=1}^{M}\left|H_p(L_N^*(\overline{Y_i}))\right|\nabla_\theta\ln v(\overline{Y_i};\theta)=0,\tag{25}$$

this is called the *Stochastic Counterpart* (SC) of the problem (24). Note that (25) is usually a system of non-linear equations, but for some pdfs results into an explicit solution.

In this paper we focus mainly on the IS of idiosyncratic terms of the systemic factors (HSF and SSF). Therefore to simplify the notation of the random vector $\overline{Y}$ of all random variables contributing to L_N will be in further text understood as a vector of $K_S + K_H$ independent standard normal random variables. LGDs or other random variables will be still part of $\overline{Y}$, but the IS won't affect them.

Now if we consider X as a system of $K_S + K_H$ independent normally distributed random variables parametrized by mean and variance, we will get the following solution of problem (25):

$$\widetilde{\mu_j} = \frac{\sum_{i=1}^{M} \left| H_p(L_N^*(\overline{Y_i})) \right| \left(\overline{Y_i} \right)_j}{\sum_{i=1}^{M} \left| H_p(L_N^*(\overline{Y_i})) \right|}, \forall j, \tag{26}$$

$$\widetilde{\sigma_j}^2 = \frac{\sum_{i=1}^{M} \left| H_p(L_N^*(\overline{Y_i})) \right| \left(\left(\overline{Y_i} \right)_j - \widetilde{\mu_j} \right)^2}{\sum_{i=1}^{M} \left| H_p(L_N^*(\overline{Y_i})) \right|}, \forall j, \tag{27}$$

where $\widetilde{\mu_j}, \widetilde{\sigma_j}^2$ is the SC approximation of mean, variance of j-th component of $\overline{Y}$ and $\left(\overline{Y_i} \right)_j$ is j-th component of i-th MC sample.

2.2 *Gaussian mixture model*

In the end of previous part we presented formulas for calculating the "optimal" IS distribution in the family of normal distributions. This approach is commonly used for the IS in the multi-factor Merton model, see for example (Glasserman & Li 2005). The choice of the IS family of distributions as normal distributions is not always optimal and can improved by more complex IS family of distributions.

The IS family of distributions examined in this paper is the family of the Gaussian mixture distributions, the same approach in different application can be found in (Kurtz & Song 2013). The Gaussian mixture random variable is defined as a weighted sum of different normal random variables. The pdf of the Gaussian mixture random variable can be expressed as

$$g(x; \boldsymbol{p}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{i=1}^{n} p_i \cdot f_N(x; \mu_i, \sigma_i), \tag{28}$$

where $f_N(x; \mu_i, \sigma_i)$ is the pdf of the normal distribution with the mean μ_i and the variance σ_i^2 and $\|\mathbf{p}\|_1 = \sum_{i=1}^{n} p_i = 1$. New IS Gaussian mixture joint pdf of $\overline{Y}$ will be

$$g_{\overline{Y}}(\overline{x}; \overline{\boldsymbol{p}}, \overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\sigma}}) = \prod_{j=1}^{K_S + K_H} g(x_j; \boldsymbol{p}_j, \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j), \tag{29}$$

where $\overline{\boldsymbol{p}}, \overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\sigma}}$ are matrices of $K_S + K_H$ columns of parameters $\boldsymbol{p}_j, \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j$. Therefore the system of pdfs for the IS is

$$X := \left\{ g_{\overline{Y}}\left(\overline{x}; \overline{\boldsymbol{p}}, \overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\sigma}} \right) : \left\| \boldsymbol{p}_j \right\|_1 = 1, \sigma_{j,i} > 0 \right\}. \tag{30}$$

Because the support of the pdf of the normal distribution is $\mathbb{R}$, the condition $f(\overline{x}) > 0 \Rightarrow g_{\overline{Y}}(\overline{x}; \overline{\boldsymbol{p}}, \overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\sigma}}) > 0$ is fulfilled. Since the components of $g_{\overline{Y}}(\overline{x}; \overline{\boldsymbol{p}}, \overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\sigma}})$ are independent, the problem (24) reduces into $K_S + K_H$ systems of non-linear equations. Therefore together with the condition $\left\| \boldsymbol{p}_j \right\|_1 = 1$ we will receive $\forall j = 1, \ldots, K_S + K_H, \forall i = 1, \ldots, n$:

$$\mu_{j,i} = \frac{\sum_{k=1}^{M} \left| H_p\left(L_N^*\left(\overline{Y_k} \right) \right) \right| \gamma_{k,j,i} \left(\overline{Y_k} \right)_j}{\sum_{k=1}^{M} \left| H_p\left(L_N^*\left(\overline{Y_k} \right) \right) \right| \gamma_{k,j,i}},$$

$$\sigma_{j,i}^2 = \frac{\sum_{k=1}^{M} \left| H_p\left(L_N^*\left(\overline{Y_k} \right) \right) \right| \gamma_{k,j,i} \left(\left(\overline{Y_k} \right)_j - \mu_{j,i} \right)^2}{\sum_{k=1}^{M} \left| H_p\left(L_N^*\left(\overline{Y_k} \right) \right) \right| \gamma_{k,j,i}}, \tag{31}$$

$$p_{j,i} = \frac{\sum_{k=1}^{M} \left| H_p\left(L_N^*\left(\overline{Y_k} \right) \right) \right| \gamma_{k,j,i}}{\sum_{k=1}^{M} \left| H_p\left(L_N^*\left(\overline{Y_k} \right) \right) \right|},$$

where

$$\gamma_{k,j,i} := \frac{p_{j,i} \cdot f_N\left(\left(\overline{Y_k} \right)_j ; \mu_{j,i}, \sigma_{j,i} \right)}{\sum_{i=1}^{n} p_{j,i} \cdot f_N\left(\left(\overline{Y_k} \right)_j ; \mu_{j,i}, \sigma_{j,i} \right)}. \tag{32}$$

We obtain $K_S + K_H$ systems, each representing a problem of approximation of the Gaussian mixture from data sample. This sub-problems can be solved for example by EM or K-means algorithm see (Bishop 2006, Redner & Walker 1984).

But the computation effort of the system (31) will be significantly smaller if we have an information from which component of $g\left(x_j; \boldsymbol{p}_j, \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j \right)$ was $\left(\overline{Y_k} \right)_j$ generated. Let $\overline{z}_{k,j}$ denote Bernoulli vector of identificators, such as

$$\left(\overline{z}_{k,j} \right)_i = \begin{cases} 1, & \left(\overline{Y_k} \right)_j \sim f_N\left(x; \mu_{j,i}, \sigma_{j,i} \right) \\ 0, & \text{otherwise} \end{cases}. \tag{33}$$

One can show that if we know the values of $\bar{z}_{k,j}$, then $\gamma_{k,j,i} = \left(\bar{z}_{k,j}\right)_i$ Therefore the system (31) results in explicit solution of the problem (24).

2.3 *Objective function for component identification*

In the previous part we constructed formulas for the calculation of the IS Gaussian mixture distribution. These formulas depend on the knowledge of the sample's source component $\bar{z}_{k,j}$, but this is not easily obtainable information. In this part we propose a numerical approximation of $\bar{z}_{k,j}$ based on model behaviour.

First let's consider a set of $K_S + K_H$ functions

$$\psi_j\left(\bar{y}\right) := \frac{\sum_{i=1}^{N} EAD_i \cdot D_i\left(\bar{y}\right) \cdot \beta_i \cdot \alpha_{i,j}}{\max_{i=1,\dots,N} \left\{\beta_i \cdot \alpha_{i,j}\right\} \cdot \sum_{i=1}^{N} EAD_i \cdot D_i\left(\bar{y}\right)}, \qquad (34)$$

where $\beta_i, \alpha_{i,j}, EAD_i$ are portfolio parameters of exposure *i* and $D_i\left(\bar{y}\right)$ is the default indicator of exposure *i* under the vector of all idiosyncratic shocks $\bar{y}$. In the case of no defaulting exposure the function $\psi_j\left(\bar{y}\right)$ yields 0. It can be easily shown that $0 \le \psi_j\left(\bar{y}\right) \le 1$.

To demonstrate a link between $\bar{z}_{k,j}$ and $\psi_j\left(\overline{Y_k}\right)$ let's consider portfolio containing a component *j* with huge impact on L_N. In Figure 2 we show dependence between component idiosyncratic shock X_j and $\psi_j\left(\bar{y}\right)$ under the condition $L_N \ge VaR_p(L_N)$. From the study of the aforementioned figure we can conclude, that:

- X_j distribution under the condition $L_N \ge VaR_p(L_N)$ consist of multiple components,
- $\psi_j(\bar{y})$ separate these components by it's value, in other words we can assume

$$\left(\psi_j\left(\overline{Y_k}\right) \in \left(a_i, a_{i+1}\right)\right) \Rightarrow \left(\left(\bar{z}_{k,j}\right)_i = 1\right), \qquad (35)$$

where $0 = a_1 \le \& \le a_{n+1} = 1$ (*n* denote number of the Gaussian mixture components) are some known values.

Numerical justification of the assumption (35) can be seen in Figure 3, where we can see the histogram of the simulation of X_j distribution under the condition $L_N \ge VaR_p(L_N)$ and it's approximation by the 3 component Gaussian mixture in comparison with approximation by the normal distribution. Approximation by the Gaussian mixture was obtained by using the objective function $\psi_j\left(\bar{y}\right)$ and the pre-calculated bounds $a_1 = 0, a_2 = 0.2, a_3 = 0.8, a_4 = 1$. Other fact beside very good approximation obtained from the proposed procedure is that the approximation obtained by the normal distribution differ significantly from the approximated distribution. Note that X_j distribution under the condition $L_N \ge VaR_p(L_N)$ is an optimal distribution found by the CE method for $H_p(x) = \left(x \ge VaR_p(L_N)\right)$.

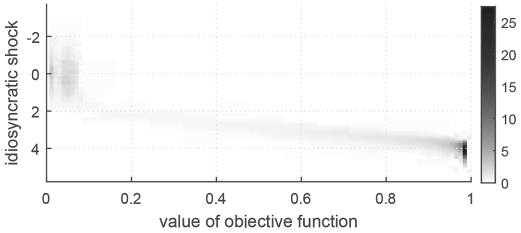


Figure 2. Approximation of dependence between $\psi_j\left(\bar{y}\right)$ and X_j.

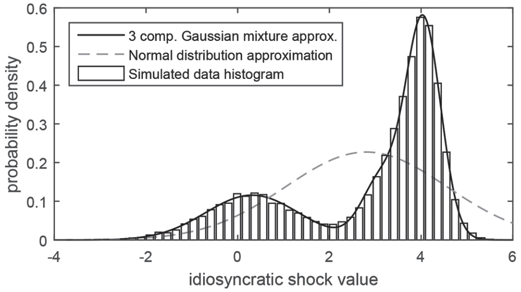


Figure 3. Approximation of X_j distribution under the condition $L_N \ge VaR_p(L_N)$ by 3 component Gaussian mixture.

Since we want to calculate both VaR and ES, the CE problem formulation based on $H_p(x)$ given by (15) does not have to be optimal. The VaR approximation can suffer if the CE method favours samples with very high value of loss and disfavours those close to $VaR_p(L_N)$ bound. Therefore we will use

$$H_p(x) = \left(x \ge VaR_p(L_N)\right), \qquad (36)$$

which give all samples with $L_N \ge VaR_p(L_N)$ same weight.

Till now we haven't dealt with bounds a_i calculation. Generally it can be a difficult problem, but $\psi_j\left(\bar{y}\right)$ component recognition is not sensitive to small changes of a_i, therefore rough approximation is sufficient. Such computationally feasible sufficient approximation can be obtained by minimizing (by e.q. line-search methods) difference between MC sample of X_j distribution under the condition $L_N \geq VaR_p(L_N)$ and the Gaussian mixture obtained using $\psi_j(\overline{y})$ component recognition.

2.4 *Adaptive CE method for IS calculation*

So far we have constructed formulas for calculating the Gaussian mixture IS, stated the optimal form of the function $H_p(x)$ in (36) and constructed an instrument for the Gaussian mixture *j*-th component identification using objective function $\psi_j(\overline{y})$. But the single calculation from M MC samples would result in the poor approximation, if the M was not high enough. The sufficient number of the MC samples for stable and precise approximation of the CE problem is comparable with the number of MC samples for sufficient approximation of VaR/ES. This would make the whole IS principle useless, because it won't bring savings in the computational time/effort. Solution to this inconvenience is iterative process, slowly shifting the IS distribution to the CE method optimal one.

The formulas for the CE method SC (31) can be modified by using the IS during the SC process:

$$\widetilde{\mu}_{j,i,t} = \frac{\sum_{k=1}^{M} \left|H_p\left(\overline{Y_k}\right)\right| w\left(\overline{Y_k}\right)\left(\overline{z}_{k,j}\right)_i \left(\overline{Y_k}\right)_j}{\sum_{k=1}^{M} \left|H_p(\overline{Y_k})\right| w\left(\overline{Y_k}\right)\left(\overline{z}_{k,j}\right)_i},$$

$$\widetilde{\sigma}^2_{j,i,t} = \frac{\sum_{k=1}^{M} \left|H_p\left(\overline{Y_k}\right)\right| w\left(\overline{Y_k}\right)\left(\overline{z}_{k,j}\right)_i \left(\left(\overline{Y_k}\right)_j - \widetilde{\mu}_{j,i,t}\right)^2}{\sum_{k=1}^{M} \left|H_p\left(\overline{Y_k}\right)\right| w\left(\overline{Y_k}\right)\left(\overline{z}_{k,j}\right)_i},$$

$$\widetilde{p}_{j,i,t} = \frac{\sum_{k=1}^{M} \left|H_p\left(\overline{Y_k}\right)\right| w\left(\overline{Y_k}\right)\left(\overline{z}_{k,j}\right)_i}{\sum_{k=1}^{M} \left|H_p\left(\overline{Y_k}\right)\right| w\left(\overline{Y_k}\right)}, \qquad (37)$$

where t denotes iteration, $(\overline{z}_{k,j})_i$ denote if the *i*-th component of *j*-th systemic factor's Gaussian mixture was the source of the sample *k*, $H_p(\overline{Y_k}) := (L_N^*(\overline{Y_k}) \geq VaR_p(L_N))$ and

$$w\left(\overline{Y_k}\right) = \frac{f\left(\overline{Y_k}\right)}{g_{\overline{Y}}\left(\overline{Y_k}; \overline{\boldsymbol{p}}_{t-1}, \overline{\boldsymbol{\mu}}_{t-1}, \overline{\boldsymbol{\sigma}}_{t-1}\right)}, \qquad (38)$$

where $f\left(\overline{Y_k}\right)$ is the pdf of nominal distribution (joint distribution of the independent normal distributions) and $g_{\overline{Y}}\left(\overline{Y_k}; \overline{\boldsymbol{p}}_{t-1}, \overline{\boldsymbol{\mu}}_{t-1}, \overline{\boldsymbol{\sigma}}_{t-1}\right)$ is the pdf of IS Gaussian mixture distribution given by parameters approximated in the iteration $t-1$.

In the definition of $H_p\left(\overline{Y_k}\right)$ is still present the unknown value of $VaR_p(L_N)$, which can be replaced by it's approximation $VaR_p^{g_{\overline{Y}}}(L_N)$ from *t*-th iteration. The last obstacle is that the $H_p\left(\overline{Y_k}\right)$ will be for most samples zero and the iteration process will crash at the beginning. The solution to this is the replacement of the confidence level *p* by a sequence of p_i which is at first few iteration significantly lower than *p* and at the and of iterative process equals *p*.

All of the previous observations lead to algorithm 1. Obtained algorithm can be further enhanced for example by the Screening method or by the adaptive smoothing parameter sequence see (Kroese, Taimre, & Botev 2013, Rubinstein & Kroese 2013, Rubinstein & Kroese 2011).

3 IMPLEMENTATION AND GPU PARALLELIZATION

The serial Matlab implementation is a straightforward interpretation of the multi-factor Merton model with the Matlab built-in functions. The whole simulation (all of the MC samples) can be calculated at once without the use of loops. Most computationally expensive parts of the simula-

Algorithm 1 Adaptive iterative calculation of the CE problem

```
Inputs: $\overline{\boldsymbol{p}}_0, \overline{\boldsymbol{\mu}}_0, \overline{\boldsymbol{\sigma}}_0$, for every systemic factor $j$ sequence of bounds $a_1 \leq \ldots \leq a_{n+1}$, smoothing parameter $\alpha \in (0,1)$, sequence of $p_1, p_2, \ldots, p_i, p, p, \ldots$, sequence of sample sizes $M_t$, set $t = 1$

1. Simulate $M_t$ samples $\overline{Y_1}, \ldots, \overline{Y_{M_t}}$ from the Gaussian mixture distribution given by parameters $\overline{\boldsymbol{p}}_{t-1}, \overline{\boldsymbol{\mu}}_{t-1}, \overline{\boldsymbol{\sigma}}_{t-1}$, calculate $\overline{VaR_{pt}^{g_{\overline{Y}}}}(L_N)$, $H_{p_t}\left(\overline{Y_k}\right), w\left(\overline{Y_k}\right), \overline{z}_{k,j} \forall j$.
2. Calculate $\widetilde{\boldsymbol{p}}_{t-1}, \widetilde{\boldsymbol{\mu}}_{t-1}, \widetilde{\boldsymbol{\sigma}}_{t-1}$ using formula (37).
3. Update parameters:
   $\overline{\boldsymbol{p}}_t = \alpha \cdot \widetilde{\boldsymbol{p}}_{t-1} + (1-\alpha) \cdot \overline{\boldsymbol{p}}_{t-1}$,
   $\overline{\boldsymbol{\mu}}_t = \alpha \cdot \widetilde{\boldsymbol{\mu}}_{t-1} + (1-\alpha) \cdot \overline{\boldsymbol{\mu}}_{t-1}$,
   $\overline{\boldsymbol{\sigma}}_t = \alpha \cdot \widetilde{\boldsymbol{\sigma}}_{t-1} + (1-\alpha) \cdot \overline{\boldsymbol{\sigma}}_{t-1}$
4. If some stopping condition is fulfilled (e.g. $\overline{\boldsymbol{p}}_t, \overline{\boldsymbol{\mu}}_t, \overline{\boldsymbol{\sigma}}_t \approx \overline{\boldsymbol{p}}_{t-1}, \overline{\boldsymbol{\mu}}_{t-1}, \overline{\boldsymbol{\sigma}}_{t-1}$) return the approximation of optimal parameters $\overline{\boldsymbol{p}}_t, \overline{\boldsymbol{\mu}}_t, \overline{\boldsymbol{\sigma}}_t$, if not set $t = t + 1$ and go back to step 1.
```

Note: sequences p_t and M_t should be calculated inside the iterative process with respect to current sample $\overline{Y_1}, \ldots, \overline{Y_{M_t}}$ (e.g. from the position of sample representing $\overline{VaR_{pt}^{g_{\overline{Y}}}}(L_N)$ in sorted sequence $L_N^*\left(\overline{Y_k}\right)$)

tion can be calculated by very well optimized Matlab matrix functions and therefore this implementation can serve as a good comparison tool of the performance efficiency for further GPU implementations.

3.1 *GPU parallelization*

As was already mentioned the simulation of the multi-factor Merton model consists of many MC samples, that are mutually independent. This is suitable for a massively parallel computation hardware such as the GPU device.

3.1.1 *Shortly about GPUs*

Let us very shortly outline main parameters of GPUs, which are crucial for model implementation:

- GPUs consist of many (in current devices in order of thousands) computation cores, grouped into *streaming multiprocessors* (SM), communication between single cores is strictly restricted to groups belonging to one SM unit. Execution of CUDA kernel (parallel GPU implementation) must mirror this structure and we must specify block size (how many threads per SM will run) and grid size (how many blocks will be executed).
- There are four basic types of memory on the GPUs:
  - global memory: main storage memory, large, high latency (thread waits long time before get the data), must be accessed in pattern (*i*-th core access *i*-th element) to obtain reasonable utilization of bandwidth
  - shared memory: small, shared between cores in one SM, low latency
  - constant memory: small, can broadcast content of array among all cores
  - registers: cannot be directly accessed, separated for every core, very fast, buffer some small local variables

For software implementation on GPU we use the NVIDIA CUDA technology. For further informations see (NVIDIA 2015).

3.1.2 *GPU implementations overview*

When implementing multi-factor Merton model we decided to create multiple implementations, which can benefit from different type of portfolios:

- "base" GPU implementation: straightforward interpretation of the model, single threads perform single MC samples in the same way as the serial implementation,
- "sparse" GPU implementation: similar to "base" implementation, but the matrix of $\alpha_{i,j}$ coefficients is handled in sparse format (only column/row index and value of non-zero elements is stored)
- "specialized" GPU implementation: is applicable only on specialized type of portfolios which use systemic factor grouping into SSF and HSF, implementation fits the mathematical description in subsection 1.2 (correlation matrix of SSF is stored in constant memory).

Finally some remarks shared by all GPU implementations:

- usage of shared memory buffering - as all cores need the same portfolio data, we can (by selected cores) copy the data from global to shared memory (which is much faster than global),
- generating random numbers from normal or uniform distribution is done by cuRAND library,
- compiled with *-use_fast_math* tag, which decreases precision of math functions in favour of speed
- Beta random number generator is not present in the cuRAND library, therefore we implemented own procedure based on rejection-sampling method see (Dubi 2000, Kroese, Taimre, & Botev 2013).

4 NUMERICAL RESULTS

In this section we test all of the aforementioned procedures and implementations. First we examine the behaviour of the GPU implementations and then we look at the variance reduction achievable by the proposed Gaussian mixture IS.

4.1 *GPU acceleration*

As was mentioned before we implemented three different approaches to simulate the multi-factor Merton model. Now we test their behaviour in comparison with the Matlab serial implementation on three different scenarios.

1. increasing number of the systemic factors which impacts majority of exposures (SSF), majority of corresponding $\alpha_{i,j}$ are non-zero
2. increasing number of systemic factors which impacts a small fraction of exposures (HSF), majority of corresponding $\alpha_{i,j}$ are zero
3. increasing number of exposures

All tests were performed on Intel Sandy Bridge E5-2470 processor (294.4 Gflops, 38.4 GB/s) and NVIDIA Kepler K20 accelerator (3520 Gflops, 208 GB/s), the serial Matlab implementation uses double precision and the GPU implementations use single precision. The theoretical performance benefit of GPU implementations is 192 × (single core + double precision vs. all GPU cores + single precision) and the theoretical memory bandwidth benefit of the GPU implementations is 11 × (double vs. single precision).

4.1.1 *Increasing number of SSF*

This test is designed to test implementation's behaviour when the number of systemic factors increases while matrix of $\alpha_{i,j}$ coefficients becomes more dense. We use the sequence of portfolios with 1000 exposures, 100 HSF and the sequence of (16, 25, 36, 49, 64, 81, 100) SSF. The density of matrix of $\alpha_{i,j}$ coefficients rises from 16% up to 51%. The scaling results can be seen in Figure 4.

From results we can observe following

- "specialized" GPU implementation's speed-up drops from factor 515 × (for 16 SSF) to factor 209 × (for 100 SSF),
- "sparse" GPU implementation suffers the most, the speed-up drops from factor 77 × (for 16 SSF.) to factor 16 × (for 100 SSF), this could be expected because size of sparse interpretation equals 3 × number of non-zero elements.
- "base" GPU implementation speed-up drops from factor 35 × (for 16 SSF) to factor 19 × (for 100 SSF).

The drop in performance of all the GPU implementations is caused by the increasing memory complexity, which bounds the computation utilization.

4.1.2 *Increasing number of HSF*

The second test is designed as the counter example to the first one. Now we test the sequence of portfolios with 1000 exposures, 25 SSF and sequence of (100, 200, 400, 800, 1600) HSF. The density of matrix of $\alpha_{i,j}$ coefficients decreases from 22% down to 1.7%. The results can be seen in Figure 5.

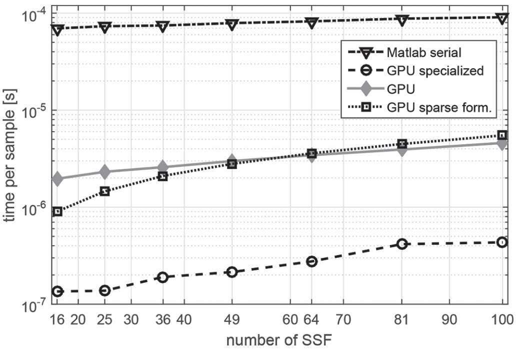


Figure 4. Implementations scaling based on rising number of high impact systemic factors.

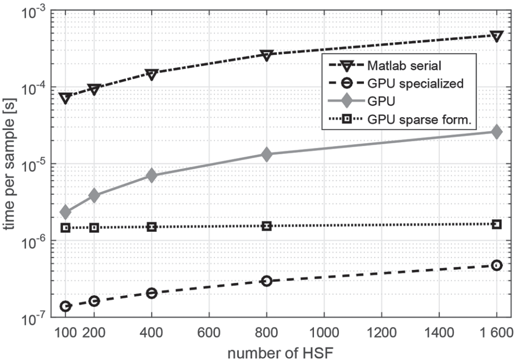


Figure 5. Implementations scaling based on rising number of low impact systemic factors.

From results we can observe following

- "specialized" GPU implementation speed-up rise from factor 537 × (for 100 HSF) to factor 1001 × (for 1600 HSF),
- "sparse" GPU implementation benefits the most, speed-up rise from factor 51 × (for 100 HSF) to factor 287 × (for 1600 HSF), this could be again expected because number of non-zero elements of matrix of $\alpha_{i,j}$ coefficients does not increase much.
- "base" GPU implementation speed-up drops from factor 32 × (for 100 HSF) to factor 18 × (for 1600 HSF).

The drop in performance of "base" GPU implementation is caused again by the increasing memory complexity, because it does not take in account the sparsity of matrix of $\alpha_{i,j}$ coefficients.

4.1.3 *Increasing number of exposures*

The last test serves as insight of the implementations behaviour when applied on the very large portfolios. We test the sequence of portfolios with 25 SSF, 100 HSF and sequence of (1000, 2000, 4000, 8000, 16000, 32000) exposures. The results can be seen in Figure 6.

From results we can observe following

- "specialized" GPU implementation speed-up rise from factor 537 × (for 100 exposures) to factor 784 × (for 3200 exposures),
- "sparse" GPU implementation speed-up is approximately 50 × for all tested portfolios,
- "base" GPU implementation speed-up is approximately 30 × for all tested portfolios.

All of the GPU implementations exhibit good scaling when the number of exposures rises, even more the "specialized" GPU implementation benefits from the large portfolios.

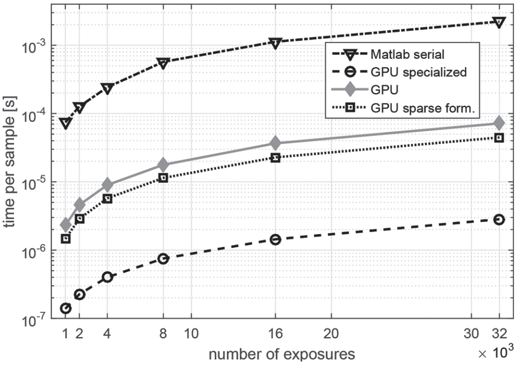


Figure 6. Implementations scaling based on rising number of exposures.

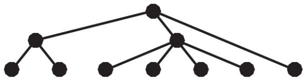


Figure 7. Template structure of HSF correlation tree.

4.2 *IS variance reduction*

In this part we examine the variance reduction achievable by the IS. We compare the standard IS approach using the family of normal distributions and the IS with the Gaussian mixture family of distributions.

4.2.1 *Portfolio parameters specification*

For numerical tests we constructed four different portfolios according to the structure mentioned in section 1.2. Each of the constructed portfolios consists of $N = 10^4$ exposures, $K_S = 25$ SSF and $K_H = 600$ HSF. Properties which are shared by all of the constructed portfolios are

- $EAD_i = i^2 / \sum_{j=1}^{N} j^2$,
- $PD_i = 0.001 + 0.001 \cdot (1 - \frac{i}{N})$,
- the distribution of LGDs is Beta distribution with mean $ELGD_n = 0.5$ and standard deviation $VLGD_n = 0.25$ for all exposures,
- the structure of HSF correlation is defined by the tree template shown in Figure 7. duplicated 60 times, correlation coefficients $g_k^H = 0.9, \forall k = 1, \ldots, K_H$.
- the SSF correlation matrix is defined by 5 region and 5 industry factors, each SSF represent unique combination of the region and the industry. Correlation between two SSF is 0.2 if they share same region, 0.15 if they share same industry and 0.03 otherwise.
- exposures are assigned to a composite SSF/HSF randomly by defined probabilities $p_k^S = \mathbb{P}(S_n = S_{(k)})$ and $p_k^H = \mathbb{P}(H_n = H_{(k)})$.

Single portfolios differs in exposure assignation to SSF, HSF and coefficients g_n, ω_n.

Portfolio 1. $p_k^S \sim \ln N(0, 0.5)$ and normalized, $p_k^H \sim$ In N (0, 10) and normalized, $g_n = 0.9, \omega_n = 0.5$.

Portfolio 2. $p_k^S = \frac{1}{K_S}$, $p_k^H = \frac{1}{K_H}$, $g_n = 0.9, \omega_n = 0.5$

Portfolio 3. $p_k^S \sim \ln N(0, 0.5)$ and normalized, $p_k^H \sim$ In N (0, 10) and normalized, $g_n = 0.5, \omega_n = 0.9$

Portfolio 4. $p_k^S = \frac{1}{K_S}$, $p_k^H = \frac{1}{K_H}$, $g_n = 0.5, \omega_n = 0.9$

Portfolio 1. represents a portfolio with clustered exposures (large groups of exposures with the same HSF/SSF composite factor) with high dependence on the systemic factors.

Portfolio 2. has the same level of exposure dependence on the systemic factors as portfolio 1., but exposures are equally distributed among the HSF/SSF composite factors.

Portfolio 3. has exposures clustered as in portfolio 1., but the level of exposure dependence is as low as in portfolio 2.

Portfolio 4. has exposures evenly distributed as portfolio 2. and low level of exposure dependence as in portfolio 3.

4.2.2 *Variance reduction in comparison with the standard approach*

Beside different portfolios we also test different levels of *confidence level* $p \in \{0.99995, 0.9995, 0.995\}$. First lets examine VaR and ES of selected portfolios and *confidence levels*, VaR/ES calculated by MC using 10^7 samples are listed in Table 1.

Measured levels of VaR, ES shows that the lower level of exposure dependence and even distribution of exposures leads to the lower value of VaR,ES. This can suggest, that the IS for portfolio 3. and 4. could be less effective. The impact of *confidence level* is predictable, the IS effectiveness will be lower for lower *confidence levels.* This is caused by reducing rarity of samples providing information about VaR, ES and therefore no large change of the distribution is needed.

Table 1. Tested portfolios VaR and ES.

		Confidence level p		
Characteristic	Portf. idx.	0.99995	0.9995	0.995
VaR	1	0.0371	0.0251	0.0129
	2	0.0291	0.0203	0.0123
	3	0.0057	0.0041	0.0027
	4	0.0051	0.0038	0.0026
ES	1	0.0417	0.0304	0.0181
	2	0.0332	0.024	0.016
	3	0.0065	0.0048	0.0033
	4	0.0058	0.0044	0.0032

Table 2. Measured variance of Crude MC, IS normal dist. and IS 3 comp. Gaussian mixture (10^6 samples, 1000 simmulations).

		Crude Monte Carlo			IS normal distribution			IS Gaussian mixture		
		Confidence level p			Confidence level p			Confidence level p		
Char.	Portf. idx.	0.99995	0.9995	0.995	0.99995	0.9995	0.995	0.99995	0.9995	0.995
VaR	1	4.95e-07	6.21e-08	4.88e-09	6.40e-09	2.42e-09	6.96e-10	6.71e-10	4.93e-10	1.90e-10
	2	3.41e-07	2.10e-08	2.95e-09	4.98e-09	1.01e-09	6.39e-10	5.47e-10	1.59e-10	1.50e-10
	3	1.14e-08	7.63e-10	5.73e-11	8.62e-11	2.33e-11	6.14e-12	2.81e-11	1.21e-11	4.79e-12
	4	7.34e-09	6.02e-10	4.64e-11	8.31e-11	2.23e-11	5.77e-12	2.31e-11	9.36e-12	4.25e-12
ES	1	8.64e-07	1.02e-07	1.05e-08	4.24e-09	1.17e-09	4.65e-10	5.06e-10	3.69e-10	2.02e-10
	2	6.99e-07	6.03e-08	5.25e-09	2.78e-09	9.70e-10	3.37e-10	3.66e-10	1.53e-10	8.00e-11
	3	2.81e-08	1.89e-09	1.42e-10	6.88e-11	1.90e-11	5.17e-12	2.05e-11	1.00e-11	4.01e-12
	4	1.61e-08	1.37e-09	1.13e-10	7.12e-11	1.99e-11	4.52e-12	1.73e-11	7.84e-12	2.96e-12

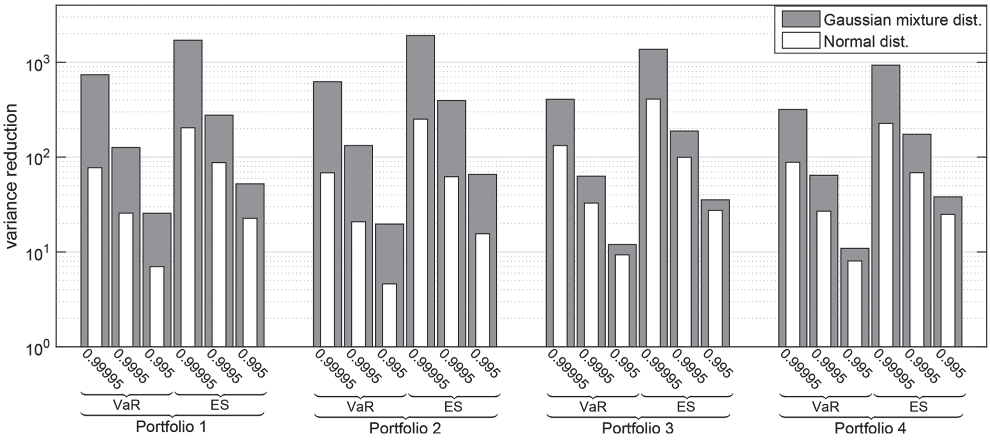


Figure 8. Variance reduction achieved by IS: Gaussian mixture and normal distribution.

Let's proceed to the testing of the variance reduction. In Table 2 we can see the variance of all combinations of tested confidence levels and portfolios for the plain (crude) MC simulation, the IS using the normal distribution and the IS using the Gaussian mixture. The variance is calculated as an empirical value of 1000 simulations consisted of 10^6 samples.

For more illustrative view of achieved variance reduction see Figure 8. Figure shows a comparison of the variance reduction between the standard and the Gaussian mixture approach for all confidence levels and portfolios combinations. Clearly the IS using the Gaussian mixture achieve better variance reduction in every test, this was evident because the normal distributions family is a subset of the Gaussian mixture distributions family.

For exact comparison of the two IS approaches, see Table 3. Table shows ratios of the variance reduction between the IS using the normal distribution and the IS using the Gaussian mixture.

Table 3. Variance reduction ratio Gaussian mix./normal dist.

		Confidence level p		
Characteristic	Portf. idx.	0.99995	0.9995	0.995
VaR	1	9.54	4.90	3.65
	2	9.10	6.35	4.26
	3	3.06	1.91	1.28
	4	3.58	2.38	1.35
ES	1	8.37	3.16	2.30
	2	7.59	6.34	4.20
	3	3.35	1.88	1.28
	4	4.11	2.54	1.52

The improvement of the IS by using the Gaussian mixture is given by the presence of systemic factor with very high impact on loss L_N. These components can be found mostly in the portfolio 1. and 2., therefore in these portfolios we obtain the best improvements in the variance reduction. Sample of such component was presented in Figure 3.

5 CONCLUSION

The objective of this paper was to speed-up the multi-factor Merton model MC simulation. This was fully accomplished by the GPU implementation and the IS application.

We presented three different GPU implementations, each better for different purpose. Two of the GPU implementations solve the general multi-factor Merton model with speed-up against serial model in range of 19 × to 287 × depending on structure of portfolio, see section 4.1. Third GPU implementation was specialized, taking input in form of structure described in section 1.2. This implementation achieves speed-up in range of 209 × to 1001 × depending on the portfolio structure.

For the IS we proposed a new approach using the Gaussian mixture distribution. Using this approach we achieved a significant variance reduction improvement for the certain portfolio structures, see section 4.2.2. In comparison to the standard IS approach we got from 9.5 × to 1.3 × better results. The total achieved variance reduction was up to 1911 × for the ES calculation and up to 737 × for the VaR calculation.

The combination of the IS and the GPU implementation can bring a speed-up of the standard serial MC simulation in orders of hundreds of thousands for portfolios with high dependence on systemic factors.

ACKNOWLEDGEMENT

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project “IT4Innovations excellence in science - LQ1602”.

REFERENCES

Bishop, C.M. (2006). *Pattern recognition and machine learning*. Springer.

Dubi, A. (2000). *Monte Carlo applications in systems engineering*. Wiley.

Glasserman, P. & J. Li (2005). Importance sampling for portfolio credit risk. *Management science 51*(11), 1643–1656.

Kroese, D.P., T. Taimre, & Z.I. Botev (2013). *Handbook of Monte Carlo Methods*. John Wiley & Sons.

Kurtz, N. & J. Song (2013). Cross-entropy-based adaptive importance sampling using gaussian mixture. *Structural Safety 42*, 35–44.

Lütkebohmert, E. (2008). *Concentration risk in credit portfolios*. Springer Science & Business Media.

NVIDIA (2015). Cuda c best practices guide. http://docs.nvidia.com/cuda/cuda-c-best-practicesguide/. Version 7.5.

Redner, R.A. & H.F. Walker (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review 26*(2), 195–239.

Rubinstein, R.Y. & D.P. Kroese (2011). *Simulation and the Monte Carlo method*. John Wiley & Sons.

Rubinstein, R.Y. & D.P. Kroese (2013). *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Highly reliable systems simulation accelerated using CPU and GPU parallel computing

S. Domesová & R. Briš
VŠB—Technical University of Ostrava, Ostrava-Poruba, Czech Republic

ABSTRACT: Highly reliable systems simulation is a complex task that leads to a problem of rare event probability quantification. The basic Monte Carlo method is not a sufficiently powerful technique for solving this type of problems, therefore it is necessary to apply more advanced simulation methods. This paper offers an approach based on the importance sampling method with distribution parameters estimation via the cross-entropy method in combination with the screening algorithm. This approach is compared to another one based on the Permutation Monte Carlo method particularly in terms of the achieved variance reduction. The paper also explains, how to apply these simulation methods to systems with independent components, that can be represented by the use of the adjacency matrix. A new generalized algorithm for the system function evaluation, which takes into account an assymetric adjacency matrix, is designed. The proposed simulation method is further parallelized in two ways, on GPU using the CUDA technology and on CPU using the OpenMP library. Both types of implementation are run from the MATLAB environment, the MEX interface is used for calling the C++ subroutines.

1 INTRODUCTION

Highly reliable systems can be found in many branches of engineering, typical examples include communication networks, production lines, storage systems, etc. The computer simulation helps to ensure almost flawless operation of these systems.

This paper offers an innovative approach to the simulation methodology usable for reliability quantification of highly reliable systems with independent components. This approach uses a static model of the simulated system, however it can also be utilized while simulating the system operation over a long period of time. Generally the approach can cover the following situations:

1. **A static system simulation.** The values of probability, that the individual components are operational, are given. Based on these values, the system reliability (i.e. the probability, that the whole system is operational) is estimated.
2. **A dynamic system simulation.** When simulating the system operation over a period of time, it is assumed, that the time dependent components reliability follows a proper probability distribution. In specific time points within the studied period, components reliabilities are computed to be than used for the system reliability quantification, the problem can simply be converted to the situation 1. For example when the operation of a component is modelled using two random variables time to failure and time to repair, both following the exponential distribution, the well-known formula

$$p(t) = \frac{\mu}{\mu+\lambda} + \frac{\lambda}{\mu+\lambda} e^{-(\mu+\lambda)t} \tag{1}$$

(see e.g. (Dubi 2000)) can be used to calculate the probability, that the component is operational in a given time t, λ stands for the failure rate, μ stands for the repair rate and it is assumed that $p(0) = 1$. The reference (Briš 2007) offers other component models, for example for the case, when the component subjects to more than one type of failure, or the case, when the time to failure follows the log-normal distribution.
3. **A steady state simulation.** If the steady state of the system is studied, the input values of the problem are the values of probability, that the components are operational in the infinite time. Therefore the problem is also convertible to the situation 1. In the aforementioned case of the random variables time to failure and time to repair following the exponential distribution, this probability is given by

$$p(\infty) = \lim_{t\to\infty} p(t) = \frac{\mu}{\mu+\lambda}. \tag{2}$$

In the case of highly reliable systems, the probability of the system failure is very low, therefore the system simulation leads to the problem of rare event probability quantification.

1.1 *System specification*

Consider a system of n $(n \in \mathbb{N})$ components. Each component remains in one of the two states, operational or failed. The state of the system is described by the state vector $\boldsymbol{b} = (b_1, \ldots, b_n)$, its elements represent states of the components. If the ith component is operational, $b_i = 0$, and if it is failed, $b_i = 1$. Furthermore it is necessary to define the system function H. This function returns 0, if the system is operational for a specific state vector $\boldsymbol{b}$, and otherwise $H(\boldsymbol{b}) = 1$.

The stochastic properties of the system are described by the random vector $\boldsymbol{B} = (B_1, \ldots, B_n)$, where the random variable B_i is assigned to the ith component. The probability distribution of B_i is Bernoulli with the parameter p_i, i.e. $\mathbb{P}(B_i = 1) = p_i$. Event $B_i = 1$ indicates the failure of the ith component, whereas $B_i = 0$ indicates the operational state. Actually the vector $\boldsymbol{p} = (p_1, \ldots, p_n)$ is a vector of the unreliabilities of the components.

System availability is defined as probability, that the system is operational. However when dealing with rare event probabilities it is more suitable to formulate the problem as calculating the unavailability of the system, i.e. the probability ℓ that the system is not operational. Obviously $\ell = \mathbb{E}(H(\boldsymbol{B}))$ holds.

2 SIMULATION METHODS

A system of n components, a system function H and a vector $\boldsymbol{p}$ of components unreliabilities are given. The aim is to estimate the probability ℓ .

2.1 *Monte Carlo*

When a basic Monte Carlo (MC) approach is used, we first generate N ($N \in \mathbb{N}$) samples of the random vector $\boldsymbol{B}$ and therefore we obtain random samples $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_N$. The value of ℓ is then estimated as

$$\hat{\ell}_{MC} = \frac{1}{N} \sum_{k=1}^{N} H(\mathbf{B}_k). \tag{3}$$

It's an unbiased estimator of ℓ , $\mathbb{E}(\hat{\ell}_{MC}) = \ell$, however this approach is not suitable for highly reliable systems (Kleijnen, Ridder, & Rubinstein 2010). For the variance of $\hat{\ell}_{MC}$ we can easily obtain $\operatorname{Var}(\hat{\ell}_{MC}) = \sigma^2 / N$, where $\sigma^2 = \operatorname{Var}(H(\boldsymbol{B}))$. Using the central limit theorem we can determine a $1 - \alpha$ confidence interval for ℓ as

$$\left(\hat{\ell}_{MC} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}}, \hat{\ell}_{MC} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}} \right), \tag{4}$$

where α usually equals 0.05 or less in practical applications and $z_{1-\frac{\alpha}{2}}$ denotes the $\left(1 - \frac{\alpha}{2}\right)$-quantile of the standard normal distribution. Number of samples required to achieve the predetermined accuracy ε with probability $1 - \alpha$ is

$$N_{\varepsilon} \approx \left(\frac{\sigma}{\varepsilon} z_{1-\frac{\alpha}{2}} \right)^2. \tag{5}$$

Example 1. Consider an example of rare event probability estimation. Determine the number of samples required to estimate the value of ℓ with accuracy $\varepsilon = 0.1 \cdot \ell$ and with probability $1 - \alpha = 0.95$. In this case $\sigma^2 = \ell - \ell^2$, therefore $N_{\varepsilon} = \left(\ell - \ell^2\right)\left(\frac{1}{\varepsilon} z_{1-\frac{\alpha}{2}}\right)^2 \doteq 384 \cdot \left(\frac{1}{\ell} - 1\right)$. For $\ell = 10^{-m}$ it is necessary to perform more than $3.84 \cdot 10^{m+2}$ samples, which would be excessively time consuming especially for complicated systems and high values of m. (The value of m equals 4 and more for rare event probabilities.)

The previous example shows the need of using variance reduction techniques, that allow achieving the same accuracy when performing a significantly lower number of samples.

2.2 *Importance sampling*

For variance reduction the Importance Sampling (IS) technique can be used. Random samples $\mathbf{B}_1, \ldots, \mathbf{B}_N$ are generated from a different distribution and the value of ℓ is then estimated as

$$\hat{\ell}_{IS} = \frac{1}{N} \sum_{k=1}^{N} H(\boldsymbol{B}_k) \frac{f(\boldsymbol{B}_k)}{g(\boldsymbol{B}_k)}, \tag{6}$$

where f is the original probability density function of random vector $\boldsymbol{B}$ (called nominal pdf) and g is the probability density function, from which the samples were generated (called IS pdf). The ratio $f(\cdot)/g(\cdot) = W(\cdot)$ is often called the likelihood ratio. The IS pdf must satisfy the condition

$$g(\boldsymbol{x}) = 0 \Rightarrow H(\boldsymbol{x}) f(\boldsymbol{x}) = 0, \tag{7}$$

see for example (Rubinstein & Kroese 2011, Kroese, Taimre, & Botev 2013). The principle of this technique is simple, but it can be difficult to find an appropriate IS pdf, which will lead to a massive variance reduction.

It is not a rule, but it is usual to select an IS pdf from the same family of distributions as the nominal pdf comes from. In this case the nominal pdf is a joint probability density function of a multivariate Bernoulli distribution with independent variables, therefore it is a product of the probability density functions of random variables $B_1,\ldots,B_n$. As the IS pdf we will also use a product of Bernoulli probability density functions, so the key issue is to find appropriate parameters $q_1,\ldots q_n$ of the new Bernoulli distributions. This can be done using the cross-entropy method (Rubinstein & Kroese 2013).

2.3 *The Cross-Entropy method*

The Cross-Entropy (CE) method offers an effective way to find an IS pdf g for which the variance of the estimator (6) is small. This method is based on the minimization of the Kullback-Leibler divergence between the unknown IS pdf and theoretical optimal IS pdf, which is explained thoroughly by (Rubinstein & Kroese 2013). For this purpose it is sufficient just to briefly outline the method.

It is assumed that the nominal pdf f has the form $f(\cdot;\boldsymbol{p})$ and the IS pdf belongs to the same parametric family, i.e. $g = f(\cdot;\boldsymbol{q})$. It can be shown that the aforementioned minimization is equivalent to the maximization of $\mathbb{E}_{\mathbf{q}}\left(H(\boldsymbol{B})\ln f(\boldsymbol{B};\boldsymbol{q})\right)$ according to vector $\boldsymbol{q}$. The requested vector

$$\underset{\mathbf{q}\in\Theta}{\operatorname{argmax}}\,\mathbb{E}_{\mathbf{q}}\left(H(\boldsymbol{B})\ln f(\boldsymbol{B};\boldsymbol{q})\right) \tag{8}$$

can naturally be estimated using the MC method as

$$\hat{\mathbf{q}} = \underset{\mathbf{q}}{\operatorname{argmax}}\frac{1}{N}\sum_{k=1}^{N} H(\boldsymbol{B}_k)\ln f(\boldsymbol{B}_k;\boldsymbol{q}), \tag{9}$$

where $\boldsymbol{B}_1,\ldots,\boldsymbol{B}_N$ are random samples generated from the nominal pdf $f(\cdot;\boldsymbol{p})$.

In the case of the distributions from the exponential families (for example Bernoulli or exponential) the stochastic program (9) has a simple solution. The elements of the vector $\hat{\mathbf{q}} = (\hat{q}_1,\ldots,\hat{q}_n)$ can be computed as

$$\hat{q}_i = \frac{\sum_{k=1}^{N} H(\boldsymbol{B}_k)B_{ki}}{\sum_{k=1}^{N} H(\boldsymbol{B}_k)},\quad i = 1,\ldots,n, \tag{10}$$

where B_{ki} means the ith coordinate of the vector $\mathbf{B}_k$ (Kroese, Taimre, & Botev 2013).

There is also a possibility to use the CE method as an iterative method and gradually refine the vector of parameters of the IS pdf. The iterative version for the distributions from the exponential families is given by the following Algorithm 1.

Algorithm 1 Cross-entropy iterative method

```
Inputs: sample size $N$, vector $\boldsymbol{p}$ of parameters of the nominal pdf, functions $f$ and $H$, stopping criterion (e.g. a predetermined number of steps).
1. Choose the initial vector $\boldsymbol{q}^{(0)}$, for example $\boldsymbol{q}^{(0)} = \boldsymbol{p}$. Set $j = 1$. Set $I = \{1,\ldots,n\}$.
2. Generate the samples $\boldsymbol{B}_1,\ldots,\boldsymbol{B}_N$ from the pdf $f\left(\cdot;\boldsymbol{q}^{(j-1)}\right)$.
3. For all $i \in I$ compute the elements of the vector $\boldsymbol{q}^{(j)} = \left(q_1^{(j)},\ldots,q_n^{(j)}\right)$ as
   $q_i^{(j)} = \frac{\sum_{k=1}^{N} H(\boldsymbol{B}_k)W\left(\boldsymbol{B}_k;\boldsymbol{p},\boldsymbol{q}^{(j-1)}\right)B_{ki}}{\sum_{k=1}^{N} H(\boldsymbol{B}_k)W\left(\boldsymbol{B}_k;\boldsymbol{p},\boldsymbol{q}^{(j-1)}\right)}$,
   where $W(\cdot;\boldsymbol{p},\boldsymbol{q}) = f(\cdot;\boldsymbol{p})/f(\cdot;\boldsymbol{q})$.
4. If the stopping criterion is not fulfilled, continue with step 2.
Output: vector $\boldsymbol{q} = \boldsymbol{q}^{(j)}$ of parameters of the IS pdf.
```

At first sight it seems that the CE algorithm provides a straightforward way to find an appropriate vector $\boldsymbol{q}$ of parameters of the IS pdf, however the algorithm should be used with caution.

In the step 2. the samples $\boldsymbol{B}_1,\ldots,\boldsymbol{B}_N$ are generated from the pdf $f\left(\cdot;\boldsymbol{q}^{(j-1)}\right)$. If the system is operational for all of these samples, the new vector $\boldsymbol{q}^{(j)}$ cannot be determined. This situation is often caused by low values of vector $\boldsymbol{q}^{(j-1)}$ elements. It is possible to add the step

- If $H(\boldsymbol{B}_1) = \cdots = H(\boldsymbol{B}_N)$, repeat step 2.

between steps 2. and 3., however this would prolong the computation time and may lead to the likelihood ratio degeneracy, for more information about the degeneracy of the likelihood ratio see (Rubinstein & Kroese 2011). To avoid this situation, it is therefore necessary to pay sufficient attention to the choice of the initial vector $\boldsymbol{q}^{(0)}$. It is evident, that when the elements of the vector $\boldsymbol{p}$ of parameters of the nominal distribution are rare event probabilities, the choice $\boldsymbol{q}^{(0)} = \boldsymbol{p}$ is not convenient. A proper way of choosing the vector $\boldsymbol{q}^{(0)}$ will be proposed in section 6.

Consider also the case, when for some component index i the values B_{ki} equal 0 for all $k = 1,\ldots,N$. In this case the ith element of vector $\boldsymbol{q}^{(i)}$ would be set to zero and the IS pdf would not satisfy the condition (7). The screening algorithm, which will be described below, is primarily intended to prevent the likelihood ratio degeneracy, but it also solves this problem.

2.4 *Screening algorithm*

The screening algorithm is often used to identify the non-bottleneck (i.e. unimportant) components of the vector $\boldsymbol{q}^{(j)}$. If the relative difference between the nominal and the new parameter

$$\frac{q_j^{(i)} - p_j}{p_j} \tag{11}$$

is smaller than some threshold ε, the ith component is marked as non-bottleneck and the value of $q_j^{(i)}$ is set to p_j. This means that the ith component does not influence the likelihood ratio in the IS estimator (6). When applying this algorithm to the highly reliable systems simulation, it is sufficient to set ε to zero. Consequently the value of q_i cannot be smaller than the nominal value p_i. Since p_i denotes the (usually rare event) probability of the ith component failure, it would be unreasonable to decrease this probability in the IS pdf.

The use of this type of the screening algorithm means to insert the following step after the step 4. of the Algorithm 1,

- For all $i \in I$ check, if $q_i^{(j)} < p_i$. If this condition is fulfilled, set $q_i^{(j)} = p_i$ and remove i from I.

3 SYSTEM REPRESENTATION AND CORRESPONDING SYSTEM FUNCTION

Every system is a set of components, however each type of system requires different kind of representation and different approach to the simulation. For easier work with systems, many types of system representation have already been developed. An example is explained in (Briš 2010), where the system is represented as a directed acyclic graph. Other ways to construct the system function for special types of systems are discussed in (Ross 2007).

3.1 *System representation using adjacency matrix*

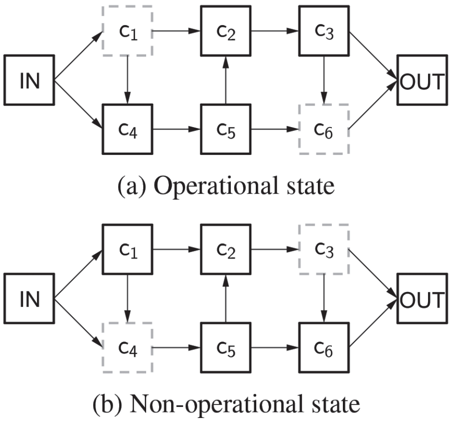
In this case a representation using an adjacency matrix is suggested. This representation is intended for systems, that can be interpreted as a collection of n components and two special elements called IN and OUT, where some of these $n+2$ elements are connected to (or "in relation with") other elements. Each of the n components is either operational or failed. We say that a system of this kind is operational, if there exists a path from IN to OUT leading only through the operational components. Such a system can be depicted as a directed graph with $n+2$ nodes.

Example 2. The schemes in Figure 1 show an example of a system with independent components, that can be represented using an adjacency matrix. The corresponding adjacency matrix for this system is

$$\mathbf{M} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

In the language of relations the system and corresponding system function can be represented as follows. A system is determined by a set $\Gamma = \{\mathrm{IN}, c_1, \ldots, c_n, \mathrm{OUT}\}$ and relation R on set Γ, the set $C = \{c_1, \ldots, c_n\}$ is a set of all system components. For each two elements in Q it can be decided, whether they are in relation or not. If $\gamma_i \in \Gamma$ is in relation with $\gamma_j \in \Gamma$, we write $(\gamma_i, \gamma_j) \in R$. Generally this relation is not symmetric, if $(\gamma_i, \gamma_j) \in R$ holds, then $(\gamma_j, \gamma_i) \in R$ does not necessarily hold. The computer representation of relation R can be easily realised using a $(n+2) \times (n+2)$ matrix of logical values 0 and 1, which we call the matrix of the relation or simply the adjacency matrix. First row and column belongs to the IN elements, last row and column to the OUT element and the remaining belong to the components. The value 1 at position (i, j) indicates, that $(\gamma_i, \gamma_j) \in R$.

The system function H is specified by the following condition. The system is operational if there



(a) Operational state

(b) Non-operational state

Figure 1. System example.

exists a sequence $k_1, \ldots, k_d$ ($d \in \mathbb{N}$) of indexes of **operational** components that

$$\begin{aligned}&\left(\mathrm{IN}, c_{k_1}\right) \in R, \left(c_{k_d}, \mathrm{OUT}\right) \in R,\\&\forall i \in \{1,\ldots,d-1\} : \left(c_{k_i}, c_{k_{i+1}}\right) \in R\end{aligned} \qquad (12)$$

holds and failed otherwise.

This representation is suitable for example for reliability networks studied in (Kroese, Taimre, & Botev 2013), which are usually represented as undirected graphs with components represented by edges. This reference also describes a simulation method based on the Conditional Monte Carlo designed for reliability networks. This method will be modified for the more general system representation and compared with the simulation method based on the importance sampling.

3.2 *System function evaluation algorithm*

For the system function evaluation we created the Algorithm 2 based on our previous results in (Briš & Domesová 2014). Even though the former algorithm was originally intended for systems with symmetric adjacency matrix, it can be modified to reflect the more general case, i.e. the asymmetric adjacency matrix.

Algorithm 2 Function $H(\boldsymbol{b})$ evaluation

```
Inputs: vector $\boldsymbol{b}$, relation $R$ (determined by $\boldsymbol{M}$).
  1. Initialization:
     (a) create an empty set $M_1$
     (b) create a set $M_2$ containing operational components $c_i$, for which $(\mathrm{IN}, c_i) \in R$
     (c) create a set $M_3$ containing failed components
  2. If $M_2$ is empty, the system is failed, $H(\boldsymbol{b}) = 1$. Terminate this algorithm.
  3. If $M_2$ contains a component $c_i$, for which $(c_i, \mathrm{OUT}) \in R$, the system is operational, $H(\boldsymbol{b}) = 0$. Terminate this algorithm.
  4. Insert components $c_i$, that are not in $M_3$ and fulfil the condition $\exists \gamma \in M_2 : (\gamma, c_i) \in R$, into $M_1$.
  5. Move all elements from $M_2$ into $M_3$. Move all elements from $M_1$ into $M_2$.
  6. Continue with step 2.
Output: system state (operational or failed).
```

4 IMPLEMENTATION AND PARALLELIZATION

The IS-based simulation method presented in section 2 in combination with the function H evaluation forms a useful tool for the simulation of systems specified by the adjacency matrix. It's principle is based on the generation of independent samples, therefore we can easily reduce the simulation time using parallel computing.

There are many ways to implement the method. For comfortable work with the simulation results in the form of graphs it is convenient to use the Matlab environment. However, to reduce the computing time it is better to focus on lower-level programming languages. There is a possibility to combine the advantages of both approaches, to use Matlab for the user-friendly work with results and implement the most important algorithms in other languages. The MEX interface of Matlab allows to call functions written in C or C++ from Matlab as easily as if they were usual Matlab functions.

After consideration of possible solutions the two following alternatives of Matlab implementation acceleration were chosen:

1. parallel computing on CPU using the OpenMP library (via the MEX interface),
2. parallel computing on GPU using the CUDA technology (via the Parallel Computing Toolbox).

In the first alternative the source codes of the accelerated functions are written in the C++ language and for random numbers generation the Boost library, version 1.56, is used. The second alternative uses source codes written in the CUDA C extension of the C language, random numbers are generated via the cuRAND library [NVIDIA 2014].

4.1 *System function implementation*

The process of the system function evaluation consumes most of the simulation time, therefore the efficiency of the implementation of this function determines the computation time of the simulation to some extent.

For function H evaluation the Algorithm 2 is used. To reduce the memory requirements and the consumption of computation time of the implementation, bitwise operations are used. Matrix $\boldsymbol{M}$ and each of the sets M_1, M_2 and M_3 are implemented as arrays of `unsigned int` data type, each variable has the length of 32 bits. In the case of the sets M_1, M_2 and M_3 the individual bits determine the presence of certain component in the set and in the case of the matrix $\boldsymbol{M}$ the bits symbolize relations between the components and the elements IN and OUT. For example the representation of the matrix $\boldsymbol{M}$ from the example 2 as an unsigned int array is

$$[18 \quad 20 \quad 8 \quad 192 \quad 32 \quad 68 \quad 128 \quad 0]. \tag{13}$$

If the simulated system has 32 components or less, the implementation works with only one `unsigned int` variable for each of the sets M_1, M_2 and M_3.

4.2 *Simulation method implementation*

The simulation is divided into two basic steps, first one is the cross-entropy algorithm for the determination of distribution parameters and second one is the importance sampling method itself. The basic scheme of the CE algorithm is written in the Matlab language and it runs in m iterations.

The most important part of the implementation is formed by the function `simulation_run`, which includes the function H algorithm. The function `simulation_run` is accelerated using the two ways mentioned above and it is executed m + 1 times in total, once for every CE method iteration and once for the IS method. Its input arguments are the matrix $\boldsymbol{M}$ converted into an array of `unsigned int` variables, number of components n, number of samples N to be performed during one CE method iteration/IS method execution, vector $\boldsymbol{p}$ of parameters of the nominal pdf, vector $\boldsymbol{q}$ of parameters of the IS pdf and a vector of bottleneck components indexes. The function outputs the value $\sum_{k=1}^{N} H(\boldsymbol{B}_k) W(\boldsymbol{B}_k; \boldsymbol{p}, \boldsymbol{q})$ and in the case of the CE method iterations the second output is a vector of values $\sum_{k=1}^{N} H(\boldsymbol{B}_k) W(\boldsymbol{B}_k; \boldsymbol{p}, \boldsymbol{q}) B_{ki}$ for every $i \in \{1,\ldots,n\}$.

4.3 *Accuracy of likelihood ratio calculation*

The CE algorithm and the IS method contain calculating the likelihood ratio as

$$W(\boldsymbol{b}; \boldsymbol{p}, \boldsymbol{q}) = \frac{f(\boldsymbol{b}; \boldsymbol{p})}{f(\boldsymbol{b}; \boldsymbol{q})} = \frac{\prod_{i=1}^{n} p_i^{b_i} (1-p_i)^{1-b_i}}{\prod_{i=1}^{n} q_i^{b_i} (1-q_i)^{1-b_i}}. \tag{14}$$

The problem with the accuracy arises while computing $1 - p_i$ for small values of p_i. If $p_i < 0.1^8$, the expression $1 - p_i$ returns 1 in the single precision. To prevent this, we define five sets of component indexes. The set S_1 contains indexes of components, for which $b_i = 1$. The set P_ε contains indexes of components, for which $b_i = 0 \wedge p_i < \varepsilon$ and $P = \{1,\ldots,n\} \setminus (S_1 \cup P_\varepsilon)$. Similarly Q_ε contains indexes of components, for which $b_i = 0 \wedge q_i < \varepsilon$ and $Q = \{1,\ldots,n\} \setminus (S_1 \cup Q_\varepsilon)$. It is convenient to choose for example $\varepsilon = 0.1^5$ as a threshold. With this notation the likelihood ratio can be written as

$$W(\boldsymbol{b}; \boldsymbol{p}, \boldsymbol{q}) = W_1 \cdot W_2, \tag{15}$$

where

$$W_1 = \prod_{i \in S_1} \frac{p_i}{q_i} \frac{\prod_{i \in P} 1 - p_i}{\prod_{i \in Q} 1 - q_i} \tag{16}$$

and

$$W_2 = \frac{\prod_{i \in P_\varepsilon} 1 - p_i}{\prod_{i \in Q_\varepsilon} 1 - q_i}. \tag{17}$$

For the second factor we can write

$$\log W_2 = \sum_{i \in P_\varepsilon} \log(1 - p_i) - \sum_{i \in Q_\varepsilon} \log(1 - q_i) \tag{18}$$

and using the Taylor series we can approximate

$$\log W_2 \doteq \sum_{i \in Q_\varepsilon} \left(q_i + \frac{q_i^2}{2} \right) - \sum_{i \in P_\varepsilon} \left(p_i + \frac{p_i^2}{2} \right). \tag{19}$$

4.4 *Computation time*

In this section we compare the computation time of the two ways of the function `simulation_run` implementation. Computation time of one function run is examined, therefore there are no iterations of the CE method and the vector $\boldsymbol{q}$ is pre-chosen.

Two testing systems shown in Figure 2 and Figure 3 were chosen for the experiments. System A is a reliability network with a regular structure, it has 60 components represented as edges of this network. System B is a reliability network with 18 components taken from (Kroese, Taimre, & Botev 2013), where it was used as a testing problem for the demonstration of the Permutation Monte Carlo method efficiency. This method will be discussed in section 5 as an alternative to the IS approach. For simplicity we consider $\boldsymbol{p} = 0.01$ and $\boldsymbol{q} = 0.1$ for both systems.

The graphs at Figure 4 show the dependence of the computation time on number of threads for the first way of approximation. This implementation uses OpenMP and it was tested on a double Intel Sandy Bridge E5-2470 processor (16 cores). Number of samples in each simulation was $7.488 \cdot 10^7$. The values of computation time for few different thread counts are written in Table 1. Each of the simulations was executed three times, the values of computation time are averages of results obtained from this three simulations.

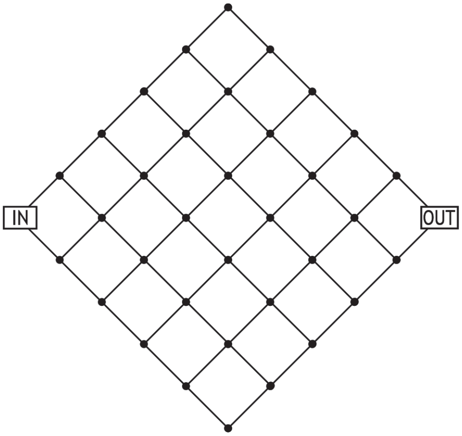


Figure 2. System A.

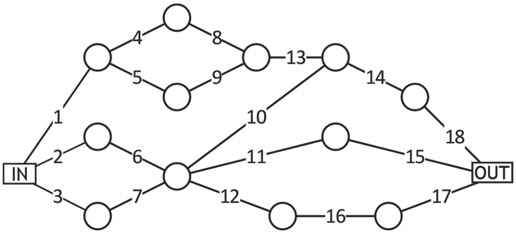


Figure 3. System B.

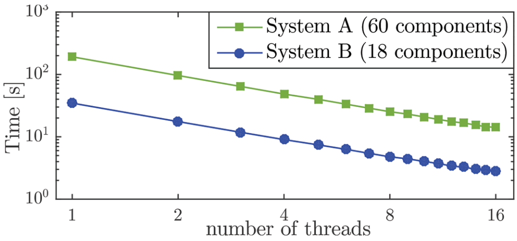


Figure 4. OpenMP implementation scalability graph.

Table 1. IS method implementation acceleration.

		Time [s]	
	Number of threads	System A	System B
OpenMP	1	191.69	34.65
	2	96.05	17.53
	4	48.48	9.02
	8	25.27	4.77
	16	14.20	2.83
CUDA		2.42	1.06

In the case of the CUDA implementation 192 threads per block, 390 blocks per grid and batches of 1000 samples per each thread are used, this gives us the same total of $7.488 \cdot 10^7$ samples as in the previous case. As a testing device a NVIDIA Kepler K20 accelerator was used. Values of the computation time for both testing systems are written in Table 1.

5 COMPARISON WITH THE PMC METHOD

The abbreviation PMC stands for the Permutation Monte Carlo method, which is used for network reliability calculations in (Kroese, Taimre, & Botev 2013). However the principle of the method can be used for a wider group of systems. With a modification based on the Algorithm 2 we applied this method to general systems determined by an adjacency matrix and we accelerated it using the CUDA technology.

The PMC method is based on a variance reduction technique called Conditional Monte Carlo, which also provides an unbiased estimator of the value $\ell = \mathbb{E}(Y)$. The Conditional Monte Carlo technique is based on the fact that $\mathbb{E}(\mathbb{E}(Y \mid \boldsymbol{Z})) = \mathbb{E}(Y)$ for any random variable Y and random vector $\mathbf{Z}$. It is assumed that there exists such a random vector $\boldsymbol{Z}$ that $\mathbb{E}(H(\boldsymbol{B}) \mid \boldsymbol{Z} = z)$ can be computed analytically for any value of $\mathbf{z}$. The value of ℓ is estimated as

$$\hat{\ell}_C = \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}(Y \mid \boldsymbol{Z}_k), \tag{20}$$

where $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ are samples of the random vector $\boldsymbol{Z}$.

The PMC method uses a different formulation of the problem. The static system described above is interpreted as an alternative evolution model captured at a specific point of time. The alternative system is described using the adjacency matrix and random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$. In this case the component states change in time. At time $t = 0$ all of the components are failed. The random variable X_i has the exponential distribution with parameter λ_i and it describes the time to repair of the ith component c_i. It is defined by the relation $\mathbb{P}(X_i < 1) = p_i$, which is fulfilled for $\lambda_i = -\ln(1 - p_i)$. Using this formulation the unavailability ℓ of the system can be expressed as

$$\ell = \mathbb{E}\left(I_{S(\mathbf{X}) \geq 1}\right) = \mathbb{P}\left(S(\mathbf{X}) \geq 1\right), \tag{21}$$

where $S(\mathbf{X})$ is a function that returns the time when the system starts operating for a sample of $\boldsymbol{X}$.

Let us define $\Pi = (\Pi_1, \ldots, \Pi_n)$ as a permutation of values $1, \ldots, n$ obtained by sorting the values $X_1, \ldots, X_n$ in ascending order, i.e. $X_{\Pi_1} < \cdots < X_{\Pi_n}$. The random variables $X_1, \ldots, X_n$ denote the times to repair of the components, therefore the random vector Π denotes the order in which the components were put into operation. The function $crit(\Pi)$ returns the smallest number of components that must be put into operation until the whole system is operational. Therefore $S(\boldsymbol{X}) = X_{crit(\Pi)}$.

At this point all the random variables needed to use the Conditional Monte Carlo method are defined. In this case $Y = I_{S(\mathbf{X}) \geq 1}$ and $\boldsymbol{Z} = \Pi$. It is necessary to compute the value

$$G(\pi) = \mathbb{E}\left(I_{S(\mathbf{X}) \geq 1} \mid \Pi = \pi\right) \tag{22}$$

analytically. In (Kroese, Taimre, & Botev 2013) the value of $G(\pi)$ is computed as

$$G(\pi) = \sum_{j=1}^{c} \omega_{c,j} \exp\left(-\nu_{c-j+1}\right), \tag{23}$$

where $c = crit(\pi)$,

$$\nu_i = \sum_{j=1}^{i} \lambda_{\Pi_j} \text{ for } i \in \{1, \ldots, c\} \tag{24}$$

and the values $\omega_{c,j}$ are given by a recursive formula

$$\omega_{1,1} = 1,\ \omega_{k+1,j} = \omega_{k,j} \frac{\nu_{c-k}}{\nu_{c-k} - \nu_{c-j+1}}, \tag{25}$$

$$\omega_{k+1,k+1} = 1 - \sum_{j=1}^{k} \omega_{k+1,j} \tag{26}$$

for $k \in \{1, \ldots, c-1\}$ and $j \in \{1, \ldots, k\}$.

5.1 *CUDA accelerated implementation*

The Matlab implementation presented in (Kroese, Taimre, & Botev 2013) uses pre-counted values ν_k ($k \in 1, \ldots, c$) for the calculation of values $\omega_{k,j}$, that are saved at a form of a matrix. This approach is not suitable for the CUDA implementation because the threads can only use a limited amount of the local memory. However for the calculation of the values $\omega_{c,j}$, $j \in \{1, \ldots, b\}$, an explicit formula

$$\omega_{c,j} = \prod_{i=1, i \neq j-1}^{c} \frac{\nu_i}{\nu_i - \nu_{c-j+1}} \tag{27}$$

can be derived, the use of this formula leads to the reduction of memory requirements.

The calculation of $S(\boldsymbol{x})$ presented in (Kroese, Taimre, & Botev 2013) is based on the sequential construction of the incidence matrix and in each step it is decided whether the system is operational or not. For the CUDA implementation it is not convenient to use this method of calculation, because the process of the construction of the incidence matrix differs for the specific vectors $\boldsymbol{x}$ and every thread would need to record its own adjacency matrix. For this reason a new implementation based on the Algorithm 2 was chosen. This implementation (see Algorithm 3) uses the matrix M which is common for all the threads and therefore can be stored in the global or constant memory.

5.2 *Computation time*

The CUDA accelerated implementation was compared to the original Matlab implementation using the testing reliability network with 18 components, see Figure 3. The same hardware as in section 4.4 was used for testing. Due to the higher computation time we chose N = 7.488 · 10^6 and we also considered $\boldsymbol{p} = 0.01$. The original Matlab implementation uses a loop over the samples, therefore it can be easily parallelized using the `parfor loop` . Results of this three versions of implementation are shown in Table 2.

Algorithm 3 Function $S(\boldsymbol{x})$ evaluation

```
Inputs: vector $\boldsymbol{x}$ as a sample of $\boldsymbol{X}$, matrix $M$.
1. Sort $\boldsymbol{x}$ and determine the permutation $\boldsymbol{\pi}$ (a sample of $\boldsymbol{\Pi}$).
2. Initialize the zero state vector $\boldsymbol{b}$ of the length $n$, initialize $k = 1$.
3. Write 1 at the $(\pi_k)^{\text{th}}$ position of vector $\boldsymbol{b}$.
4. Evaluate $H(\boldsymbol{b})$ using the Algorithm 2.
5. If $H(\boldsymbol{b}) = 0$, set $crit(\boldsymbol{\pi}) = k$ and terminate the algorithm.
6. Increase $k$ by 1 and continue with step 3.
Output: $x_{crit(\boldsymbol{\pi})}$.
```

Table 2. PMC method implementation acceleration.

	Time [s]
Matlab	2332.18
Matlab + parfor (16 threads)	183.58
Matlab + CUDA	1.91

6 APPLICATIONS

The proposed approach based on the IS method was applied to the testing systems presented in section 4.4. For the experiments the CUDA accelerated version of the implementation was used. For comparison the same problems were also solved using the PMC method, also accelerated using CUDA.

The unreliability of all components is identical and equals p. Both problems are solved for ten different values of p, specifically p $\left\{0.1, 0.1^2, \ldots, 0.1^{10}\right\}$. The following inputs of the CE method are chosen: sample size $N_{CE} = 7.488 \cdot 10^5$, as the stopping criterion 10 iterations are predetermined and the initial vector $\boldsymbol{q}^{(0)}$ differs depending on the vector $\boldsymbol{p}$. Sample size is $N = 7.488 \cdot 10^6$ for both simulation methods.

6.1 *Reliability network of 60 components*

Network graph of this system is shown in Figure 2.

For $\boldsymbol{p} = \overline{0.1}$ the initial vector $\boldsymbol{q}^{(0)} = \boldsymbol{p}$ was chosen and as a result of the CE method we obtained an optimal vector $\boldsymbol{q}$ of parameters of the IS pdf, denote $\boldsymbol{q} = \boldsymbol{q}_{0.1}$. This vector was used as an input of the IS method, the estimation of the unavailability of the whole system is $\hat{\ell}_{IS} = 2.44 \cdot 10^{-2}$. For $\boldsymbol{p} = 0.01$ we have many possibilities of choosing the initial vector, however the vector $\boldsymbol{q}_{0.1}$ obtained for the previous value of p appeared to be an appropriate choice. For the remaining values $p \in \{0.1^3, \ldots, 0.1^{10}\}$, the procedure is analogous, as the initial vector $\boldsymbol{q}^{(0)}$ we always use the optimal vector $\boldsymbol{q}$, that was obtained for the higher value of unreliability p.

The results of the IS method are written in Table 4, for comparison the results of the PMC method are listed to. The approximations of the system unavailability are almost equal, both methods work properly. Differences are in the accuracy of this results, in the Table 4 the accuracy is represented by the Relative Standard Deviation (RDS) estimated as

$$\frac{s}{\hat{\ell} \cdot \sqrt{N}} \cdot 100, \tag{28}$$

where s is the standard deviation and $\hat{\ell}$ is the approximation of the system unavailability, smaller value of RSD is better. The graph at Figure 5 shows the ratio of the variance achieved by both methods, for $p < 0.1^3$ the variance achieved by the IS estimator is more than 100 times lower than the variance of the PMC estimator. For suitability of the IS approach it is especially convenient that the ratio of the variances grows with lower unavailability, i.e. it is particularly suitable for highly reliable systems.

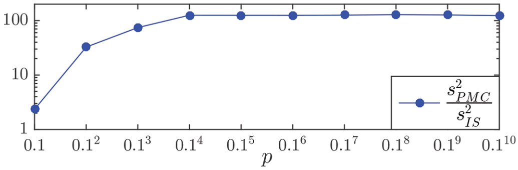


Figure 5. Ratio of the variances, higher value is better.

Table 3. Achieved variance reduction.

	Variance s^2			Variance reduction	
p	MC	IS + CE	PMC	IS + CE	PMC
0.1	2.38e-02	5.40e-03	1.28e-02	4.4 ×	1.9 ×
0.1^2	2.02e-04	5.49e-07	1.79e-05	368 ×	11 ×
0.1^3	2.00e-06	4.34e-11	3.23e-09	46156 ×	621 ×

The Table 3 compares both methods with the simple MC simulation in terms of the achieved variance reduction. Results for $p < 0.1^3$ are not listed because the sample size N was not sufficient to capture the rare event.

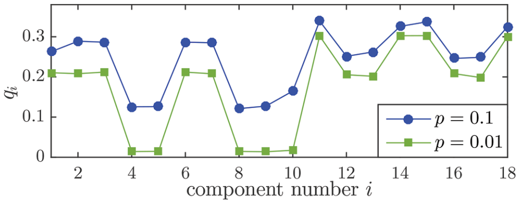


Figure 6. Vector $\boldsymbol{q}$ for the non-directed system B.

6.2 *Reliability network of 18 components*

This system is given by the network graph in Figure 3.

We applied the same procedure to the series of problems depending on $p \in \left\{0.1, 0.1^2, \ldots, 0.1^{10}\right\}$. For $p = 0.1$ we chose $\boldsymbol{q}^{(0)} = \boldsymbol{p}$ and the CE method returned the optimal vector $\boldsymbol{q}_{0.1}$ of IS pdf parameters, that is demonstrated by the upper graph of the Figure 6. For $p = 0.01$ we chose $\boldsymbol{q}^{(0)} = \boldsymbol{q}_{0.1}$ and obtained $\boldsymbol{q}_{0.01}$ as a result of the CE method, see the lower graph of the Figure 6. Graphs for $p = 0.001$ and lower are not plotted, they would coincide with the lower graph. We can notice, that values of this vectors correspond to the "importance" of individual components, i.e. high value of

Table 4. Comparisons of the proposed IS approach and the PMC method.

	System A (60 components)				System B (18 components)				System B—directed			
	IS		PMC		IS		PMC		IS		PMC	
p	$\hat{\ell}_{IS}$	RSD	$\hat{\ell}_{PMC}$	RSD	$\hat{\ell}_{IS}$	RSD	$\hat{\ell}_{PMC}$	RSD	$\hat{\ell}_{IS}$	RSD	$\hat{\ell}_{PMC}$	RSD
0.1	2.44e-02	0.110	2.44e-02	0.169	1.90e-02	0.129	1.91e-02	0.090	2.34e-02	0.112	2.35e-02	0.083
0.1^2	2.03e-04	0.133	2.05e-04	0.753	2.00e-05	0.294	1.99e-05	0.206	2.75e-05	0.213	2.77e-05	0.175
0.1^3	2.00e-06	0.120	1.98e-06	1.051	2.01e-08	0.279	1.99e-08	0.228	2.80e-08	0.218	2.79e-08	0.192
0.1^4	2.00e-08	0.097	2.00e-08	1.083	2.00e-11	0.288	2.00e-11	0.230	2.79e-11	0.216	2.80e-11	0.194
0.1^5	2.00e-10	0.098	1.98e-10	1.092	2.00e-14	0.282	2.00e-14	0.230	2.82e-14	0.216	2.81e-14	0.193
0.1^6	2.00e-12	0.097	1.98e-12	1.093	1.99e-17	0.267	1.99e-17	0.231	2.81e-17	0.215	2.81e-17	0.193
0.1^7	2.00e-14	0.097	1.99e-14	1.089	1.99e-20	0.293	1.99e-20	0.231	2.81e-20	0.218	2.80e-20	0.194
0.1^8	2.00e-16	0.097	2.03e-16	1.077	2.00e-23	0.284	2.00e-23	0.231	2.78e-23	0.216	2.79e-23	0.194
0.1^9	2.00e-18	0.097	2.01e-18	1.084	2.00e-26	0.274	2.00e-26	0.230	2.79e-26	0.222	2.80e-26	0.194
0.1^{10}	2.00e-20	0.097	1.96e-20	1.096	2.01e-29	0.271	2.00e-29	0.231	2.80e-29	0.220	2.79e-29	0.194

q_i means that if the component i is failed, the whole system is failed with high probability.

Computed values of the system unavailability and RSD for both simulation methods are written in Table 4. The values of the unavailability agree with results reported in (Kroese, Taimre, & Botev 2013). The values of RSD are now slightly lower for the PMC method, however we can see that the IS approach is suitable for this reliability network, that serves as a testing problem for the PMC method.

6.3 *System with oriented edges*

Consider again the system given by the network graph in Figure 3. We are interested in a system with similar structure, but the edges are now treated as oriented. Obviously only the edge number 10 is working in both directions. Let's say this edge will communicate only in the direction "from left down to up right", e.g. the path going from IN to OUT through edges 2, 6, 10, 14, 18 is valid but the path going from IN to OUT through edges 1, 4, 8, 13, 10, 11, 15 is not valid. It is expected that after this restriction the system unavailability will grow.

The IS approach was applied to this system using the same procedure as in the previous cases, see Figure 7. We can observe the effect of this structural change on the "importance" of the individual components. After the modifications in implementation, the PMC method can also be used for directed systems. For both methods results see Table 4. As expected, the values of unavailability are higher than in the previous case.

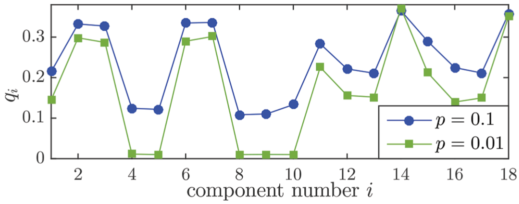


Figure 7. Vector q for the directed system B.

7 CONCLUSIONS

The simulation method based on the importance sampling technique with IS pdf parameters estimation using the cross-entropy method was successfully applied to the highly reliable systems with independent components. The proposed procedure of choosing the initial vector of the CE method has proven to be beneficial. In section 6 it was used for solving a series of problems with decreasing unreliability, however this process can be used generally when estimating very low values of the unavailability ℓ. The procedure is summarized in our general Algorithm 4 for rare event probability quantification.

Notice, that in section 6 we worked with a sequence $\alpha_k = \left(10^9, 10^8, \ldots, 10^0\right)$.

The results show that this IS-based approach is well suited for rare events quantification in the field of highly reliable systems simulation due to its massive variance reduction. For example in the case of the testing system with unavailability $\ell = 2 \cdot 10^{-6}$ the variance was reduced more than $4 \cdot 10^4$ times in comparison to the MC method. It was

Algorithm 4 CE method for rare event probability quantification using failure probability decreasing

```
Input: vector $\boldsymbol{p}_{nom}$ of parameters of the nominal pdf.
  1. Choose an appropriate decreasing sequence $\{\alpha_k\}_{k=1}^{m}$ of length $m$, where $a_m = 1$. Set $k = 1$ and $\boldsymbol{q}^{(0)} = \alpha_1 \cdot \boldsymbol{p}_{nom}$.
  2. Determine $\boldsymbol{q}^{(k)}$ as an output of the Algorithm 1 with inputs $\boldsymbol{p} = \alpha_k \cdot \boldsymbol{p}_{nom}$ and initial vector $\boldsymbol{q}^{(k-1)}$.
  3. If $k < m$, increase $k$ by 1 and go to step 2.
Output: vector $\boldsymbol{q} = \boldsymbol{q}^{(m)}$ of parameters of the IS pdf.
```

not possible to apply the simple MC method to more reliable systems, however it was shown that the variance reduction increases with increasing system reliability.

The approach was verified by applying the PMC method to the same series of testing problems. The results obtained by both methods were comparable, the IS-based method was successful especially in the case of the system of 60 components with different impact to the system reliability, where it achieved about 100 times lower variance.

Significant acceleration of the simulations was achieved using CPU and GPU parallel computing. Especially CUDA has proven to be a powerful technology for simulation based algorithms. For example the computation time of the CUDA accelerated PMC method was approximately 1200 times shorter than the computation time of the non-accelerated Matlab implementation. The modified implementation of this successful simulation method also brought a generalization for directed systems given by an adjacency matrix.

ACKNOWLEDGEMENT

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science—LQ1602".

REFERENCES

Briš, R. (2007). *Inovation methods for reliability quantification of systems and elements*. Ostrava, Czech Republic Vysoká škola báňská—Technická univerzita Ostrava.

Briš, R. (2010). Exact reliability quantification of highly reliable systems with maintenance. *Reliability Engineering & System Safety* 95(12), 1286–1292.

Briš, R. & S. Domesová (2014). New computing technology in reliability engineering. *Mathematical Problems in Engineering*.

Dubi, A. (2000). *Monte Carlo applications in systems engineering*. Wiley.

Kleijnen, J.P.C., A.A.N. Ridder, & R.Y. Rubinstein (2010). Variance reduction techniques in monte carlo methods.

Kroese, D.P., T. Taimre, & Z.I. Botev (2013). *Handbook of Monte Carlo Methods*. John Wiley & Sons.

NVIDIA (2014). Cuda c best practices guide. URL: http://docs.nvidia.com/cuda/cuda-c-best-practicesguide/.

Ross, S.M. (2007). *Introduction to probability models*. Elsevier Inc.

Rubinstein, R.Y. & D.P. Kroese (2011). *Simulation and the Monte Carlo method*. John Wiley & Sons.

Rubinstein, R.Y. & D.P. Kroese (2013). *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media

This page intentionally left blank

Network and wireless network reliability

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Advanced protocol for wireless information and power transfer in full duplex DF relaying networks

Xuan-Xinh Nguyen
Wireless Communications Research Group, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Duy-Thanh Pham & Thu-Quyen Nguyen
Ton Duc Thang University, Ho Chi Minh City, Vietnam

Dinh-Thuan Do
Ho Chi Minh City of Technology and Education, Ho Chi Minh City, Vietnam

ABSTRACT: An issue of energy consumption in the wireless network communication systems has attracted much attention from the researchers in recent years. The problem of effective energy consumption for cellular networks becomes an important key in the system design process. This paper proposes a new protocol for power tranfer, named Time Switching Aware Channel protocol (TSAC), in which the system can be aware channel gain to adjust the proper time for the power transfer. This paper also investigates the throughput optimization problem of energy consumption in Decode and Forward (DF) based cooperative network approach in which an allocated relay power transmission is proposed. By assuming that the signal at a relay node is decoded correctly when there is no outage, the optimal throughput efficiency of the system is analytically evaluated.

1 INTRODUCTION

The trend of the researchers towards energy consumption in wireless communication systems has experienced a drastic change over recent years. There have been the increasing energy-aware radio access solutions on energy consumption where a prudent use of energy is one of the decisive design elements. Besides, applications involving wireless sensor networks are becoming increasingly popular in today's demanding life.

The sensor networks and cellular networks, wireless devices are equipped by replaceable or rechargeable batteries in conventional wireless networks. However, the lifetime of these battery powered devices are usually limited. Due to lots of inconvenient circumstances such as a sensor network with thousands of distributed sensor nodes, devices located in toxic environments, medical sensors implanted inside human bodies, replacing or recharging the available batteries periodically may not be the reasonable option.

For those reasons, obtaining permanent power supply over the energy harvesting (Ozel et al., 2011, Chin Keong and Rui, 2012) has become an attractive methodology to prolong these wireless network lifetime. Nowadays, solar and wind are utilized as typical energy resources in our life. In addition, by taking advantage of an idea that the wireless signals can carry both energy as well as information (Varshney, 2008), the surrounding radio signal considered as a novel viable source is achieving more and more research attention in the field of wireless communication systems.

Basing on the first announced approach in (Varshney, 2008), more practical receiver architectures have been developed by supposing that the receiver has two circuits to separately perform energy harvesting and information decoding (Zhou et al., 2013, Nasir et al., 2013). Especially, with a strategy called time switching, the receiver can either switch on the two circuits at separate times. For a power splitting strategy, it is possible to divide its observations into two streams which are directed to the two circuits at the same time. The work (Zhou et al., 2013) has been taken into account a simple single-input single-output scenario, and the upgrading to multi-input multi-output broadcasting scenarios has been considered in (Zhang and Ho, 2013).

In the paper (Xiaoming et al., 2014), a time allocation policy is carried out for two transmitters, the efficiency of the energy transfer is maximized by means of an energy beamformer that exploits a quantized version of the CSI which is received in the uplink by the energy transmitter. Moreover, G. Yang investigates the optimal time and power allocations strategies (Gang et al., 2014) so that the total amount of harvested energy is maximized and takes into account the effect of the CSI accuracy on the latter quantity.

A relay assisted system with energy transfer has focused on two main directions: a) Simultaneous Wireless Information and Power Transfer (SWIPT) scenarios where the employed relay, (Ng and Schober, 2013, Ng et al., 2013a) (or the source terminal (Ng et al., 2013b)) salvages energy from the radiated signal incident from the source terminal (or the employed relay). b) Multi-hop energy transfer scenarios in which the energy is transferred to remote terminals (Xiaoming et al., 2014, Gang et al., 2014). The results in (Xiaoming et al., 2014) show that multi-hop energy transfer can decrease the high path-loss of the energy-bearing signal, while (Gang et al., 2014) deals with the case where a multi-antenna relay feeds two separate terminals with information and power, respectively, and studies the transmission rate and outage probability that is sacrificed in the remote energy transfer.

The principle of Full Duplex (FD) technique, which allows the communication node to transmit and receive simultaneously signals over the same frequency, has been announced and discussed (Choi et al., 2010, Duarte et al., 2012, Yingbo et al., 2012, Rui et al., 2010) and (Krikidis et al., 2012). In comparison with the Half Duplex (HD) mode, the FD mode has the ability to double the spectral efficiency due to its efficient exploitation in the limited resources. The self-interference of FD mode, however, leaking from node's transmission to its own reception, reduces the performance of FD communication.

An objective of the energy harvesting communications, called throughput optimization, has been broadly studied in volumes of literatures. In (Tutuncuoglu and Yener, 2012) and (Jing and Ulukus, 2012), the throughput optimization for transmitter with a deadline constraint was investigated over a static channel condition. In addition, the problems of throughput optimization were extended and applied to fading and multiple access channels (Chin Keong and Rui, 2010, Ozel et al., 2011, Jing and Ulukus, 2011). Besides, the cooperation between nodes is also introduced and considered to throughput optimization in the energy harvesting communications. In (Chuan et al., 2013), the problem of throughput maximization was investigated for the orthogonal in Gaussian relay channel with the energy harvesting constraints.

In particular, apart from the aforementioned literature, this paper considers a FD DF relaying networks underlay wireless energy transfer. We use and improve the Time Switching (TS) receiver mechanism so that relays harvest the energy from the source RF radiation, in which the system can be aware channel gain to adjust the proper time for the power transfer and information communication. The main contributions can be described as follows

1. This paper proposed a new protocol for energy harvesting at energy constrained relay that can be aware Channel State Information (CSI) to allocating the time for a fix pre-defined power.
2. A close-form for analytical expression in term of system's throughput and numerical result for optimal relay transmission power allocation is also derived.

The rest of this paper, section 2 introduces the system model and presents the energy harvesting protocol. Section 3 derives the outage probability and throughput analysis and power allocation policy for optimization throughput of the system. The numerical result is presented in section 4. Finally, Section 5 concludes the proposed protocol of this paper.

2 SYSTEM MODEL AND PROTOCOL DESCRIPTION

As shown in Fig. 1, we consider a system model which includes a source denoted by S, a destination denoted by *D* and an intermediate assistance relay denoted by *R*. Each node, i.e. *S* and *D* are installed an antenna, therefore, it works at half-duplex mode, *R* is equipped two antennas, and operate at full-duplex mode.

Channel assumptions: *h*, *f*, *g* are independent and identically distributed (i.i.d.) exponential random variables with mean λ_h, λ_f and λ_g, respectively. The channel gain can be get by using trainning sequences.

2.1 *Signal model*

In the Wireless Information Transfer phase (WIT), the received signal at *R*, $y_{R,i}$ and *D*, $y_{D,i}$ in the time slot *i*th, respectively, are given by

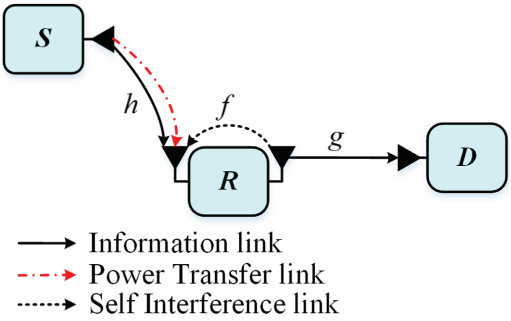


Figure 1. System model of full duplex relay communication.

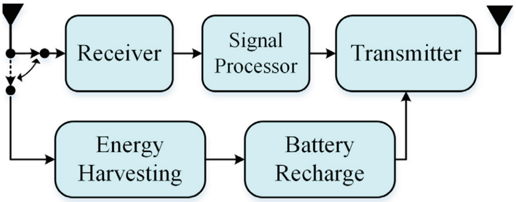


Figure 2. Energy harvesting system model.

$$y_{R,i} = \frac{h_i}{\sqrt{d_1^m}} \sqrt{P_S} x_i + f_i \sqrt{P_R} x_i' + n_{R,i} \tag{1}$$

$$y_{D,i} = \frac{g_i}{\sqrt{d_2^m}} \sqrt{P_R} x_i' + n_{D,i} \tag{2}$$

where $(.)_i$ is the block time index *i*th of (.); x_i and x'_i is message symbol at *S* and decoded symbol at *R* respectively, that with unit power and zero average. Supposing that the relay node always exactly decodes when $S \rightarrow R$ link have no outage. Besides, h_i, g_i is the source to relay channel gain and the relay to destination channel gain at slot time *i*th, respectively; P_S, P_R is transmitted power from the source and the relay, respectively; *m* is the path loss exponent; d_1 is the source to relay distance, d_2 is the relay to destination distance; $n_{R,i}$ and $n_{D,i}$ are respective AWGNs at *R* và *D* in block time *i*th.

From (1) and (2) the SINR at R and in the time slot *i*th, respectively, are determined by

$$\gamma_{R,i} = \frac{P_S |h_i|^2}{P_R d_1^m |f_i|^2 + d_1^m \sigma_{R,i}^2} \tag{3}$$

$$\gamma_{D,i} = \frac{P_R |g_i|^2}{d_2^m \sigma_{D,i}^2} \tag{4}$$

where: $\sigma_{R,i}^2$, $\sigma_{D,i}^2$ are the variances of AWGNs $n_{R,i}$ and $n_{D,i}$, respectively.

Based on the DF relaying scheme, the end-to-end SINR at block time *i*th, is proven as bellow

$$\gamma_{e2e,i} = \min\left(\gamma_{R,i}, \gamma_{D,i}\right) \tag{5}$$

where $\gamma_{R,i}$, $\gamma_{D,i}$ are the performances of the first and second hop, given by (3) and (4) respectively.

2.2 *Energy harvesting protocol*

In the Wireless Power Transfer (WPT) time slot, we focus the performance of scheme that uses the new energy harvesting protocol which is named Time Switching Aware Channel (TSAC). The protocol can be described as below (see Figure 2).

In this protocol, the antenna which is responsible for received signal, absorbs the RF signal to convert to DC signal. Therefore, the received signal at the energy constrained relay is given by

$$y_{R,i} = \frac{h_i}{\sqrt{d_1^m}} \sqrt{P_S} x_i + n_{R,i} \tag{6}$$

As in (Zhou et al., 2013, Nasir et al., 2013), the absorbable energy can be saved in an extreme-capacitor and then entirely is used for transmission stage. Hence, the transmission energy at *R* can be described as

$$E_i = \eta \alpha_i T \frac{P_S |h_i|^2}{d_1^m} \tag{7}$$

where $0 < \eta < 1$ is the energy conversion efficiency that depends on the rectification process and the energy harvesting circuitry.

In this protocol, the transmitted power in the relay node is predefined with a fixed value, denoded by P_R. Hence, the utilizable energy to work is given by

$$E_i = P_R \left(1 - \alpha_i\right) T \tag{8}$$

Setting (7) equal (8), the allocated fraction of time for power transfer in any block time is derived as follow

$$\alpha_i = \frac{P_R d_1^m}{\eta P_S |h_i|^2 + P_R d_1^m} \tag{9}$$

In equation (9), the duration of time allocated in energy harvesting phase is a function of some parameters including channel gains h_i (can be estimated as (Love et al., 2008))[1], preset power at relay node P_R, distance between *S* and *R* d_1, absorbable coefficient and power transmission at source P_S. Specially, this fraction time is always less than one, i.e. $\alpha_i < 1$, that implies the allocated time can respond for wireless information and power transfer.

[1] For the channel gain, source and relay node has to obtain them CSI. At the beginning of each block (e.g. block time *i*th), the CSI acquisition is achieved in two steps. In the first step, the source transmits its pilot signal to the relay, and the relay estimates h_i. In the second step, the relay *R* feeds back h_i, to the source node, *S*. In order to reduce the feedback overhead, the relay can feed back their quantized version to the source.

3 OUTAGE PROBABILITY AND THROUGHPUT ANALYSIS

In this section, the throughput and outage probability of the proposed protocol are analyzed. In this mode, the outage probability occurs when the system performance, i.e. $\gamma_{e2e,i}$, drops below the threshold value γ_0, it is defined as $\gamma_0 = 2^{Rc} - 1$ with R_c is transmission rate of system. So that the expression of outage probability can be obtained by

$$\begin{aligned} OP_i &= \Pr\left\{\gamma_{e2e,i} < \gamma_0\right\} \\ &= \Pr\left\{\min\left(\gamma_{R,i}, \gamma_{D,i}\right) < \gamma_0\right\} \end{aligned} \tag{10}$$

Because $\gamma_{R,i}$ independent with $\gamma_{D,i}$, so that the outage can be rewritten as

$$OP_i = 1 - \Pr\left\{\gamma_{R,i} > \gamma_0\right\}\Pr\left\{\gamma_{D,i} > \gamma_0\right\} \tag{11}$$

In this transmission mode, i.e. delay-constrained mode, the throughput efficiency of system at the time slot *i*th, is the function of outage probability and the EH duration, which is formulated by

$$t_i = \left(1 - OP_i\right)\left(1 - \alpha_i\right) \tag{12}$$

And the average throughput efficiency of system is

$$t = E\left\{t_i\right\} \tag{13}$$

where $E\{x\}$ is the expectation function of variable x.

The optimal value of transmitted power for maximization throughput, which can be modified with various system parameters, i.e. P_R, P_S, η, γ_0 … is determined via resolve of function below

$$\begin{aligned} &P_R^{opt} = \arg\max_{P_R}\left\{t\left(P_R\right)\right\} \\ &\quad \text{Subject to: } P_R \geq 0 \end{aligned} \tag{14}$$

Throughput of system in this case can be determined as (13) and after some algebraic manipulations, we have below proposition as

Proposition 1: the average throughput of full-duplex relaying energy harvesting network with TSAC protocol can be expressed as

$$\begin{aligned} t = {} & \exp\left(-\frac{\gamma_0 d_2^m \sigma_D^2}{\lambda_g P_R}\right) \times \left[\exp\left(-\frac{\gamma_0 d_1^m \sigma_R^2}{\lambda_h P_S}\right)\right. \\ & \times \frac{\lambda_h P_S}{\lambda_h P_S + \gamma_0 P_R d_1^m \lambda_f} - \frac{X}{\lambda_h}\exp\left(\frac{X}{\lambda_h}\right) \times \left\{\exp\left(\frac{X+\theta}{\bar{\omega}\lambda_f}\right)\right. \\ & \left.\left. \times Ei\left(-\frac{\left(X+\theta\right)}{\bar{\omega}\lambda_f} - \frac{X+\theta}{\lambda_h}\right) - Ei\left(-\frac{X+\theta}{\lambda_h}\right)\right\}\right] \end{aligned} \tag{15}$$

where $X = \frac{P_R d_1^m}{\eta P_S}$, $\bar{\omega} = \frac{\gamma_0 P_R d_1^m}{P_S}$, $\theta = \frac{\gamma_0 d_1^m \sigma_R^2}{P_S}$ and Ei is the exponential integral function as eq. 8.211 in (Jeffrey and Zwillinger, 2007).

Proof: the proposition 1 can be derived as below

Substituting (9), (11) into (12) the average throughput of system can be expressed as

$$\begin{aligned} t = {} & E_{|g_i|^2}\left\{\Pr\left\{\gamma_{D,i} > \gamma_0\right\}\right\} \\ & \times\left(E_{|h_i|^2|f_i|^2}\left\{\Pr\left\{\gamma_{R,i} > \gamma_0\right\}\right\}\right. \\ & \left. - E_{|h_i|^2|f_i|^2}\left\{\Pr\left\{\gamma_{R,i} > \gamma_0\right\}\alpha_i\right\}\right) \\ = {} & t_1 \times \left(t_2 - t_3\right) \end{aligned} \tag{16}$$

The first item could be given by

$$\begin{aligned} t_1 &= E_{|g_i|^2}\left\{\Pr\left\{\gamma_{D,i} > \gamma_0\right\}\right\} \\ &= E_{|g_i|^2}\left\{\Pr\left\{P_R\left|g_i\right|^2 > \gamma_0 d_2^m \sigma_D^2\right\}\right\} \\ &= \exp\left(-\frac{\gamma_0 d_2^m \sigma_D^2}{\lambda_g P_R}\right) \end{aligned} \tag{17}$$

And the second item is

$$\begin{aligned} t_2 &= E_{|h|^2,|f|^2}\left\{\Pr\left\{\gamma_{R,i} > \gamma_0\right\}\right\} \\ &= E_{|h|^2,|f|^2}\left\{\Pr\left\{P_S\left|h_i\right|^2 > \gamma_0\left(P_R d_1^m\left|f_i\right|^2 + d_1^m \sigma_R^2\right)\right\}\right\} \\ &\overset{(a)}{=} \exp\left(-\frac{\gamma_0 d_1^m \sigma_R^2}{\lambda_h P_S}\right)\frac{1}{\lambda_f}\int_0^\infty \exp\left(-\frac{\gamma_0 P_R d_1^m x}{\lambda_h P_S} - \frac{x}{\lambda_f}\right)dx \\ &= \exp\left(-\frac{\gamma_0 d_1^m \sigma_R^2}{\lambda_h P_S}\right)\frac{\lambda_h P_S}{\lambda_h P_S + \gamma_0 P_R d_1^m \lambda_f} \end{aligned} \tag{18}$$

where step (a) can be derived by the channels, i.e. h, f, are follow the (i.i.d) exponential random variables with p.d.f $f_X(x) = \lambda_X^{-1}\exp(-x/\lambda_X)$ and c.d.f $F_X(x) = 1 - \exp(-x/\lambda_X)$.

Finally, the third element is determined as

$$\begin{aligned} t_3 &= E_{|h_i|^2,|f_i|^2}\left\{\Pr\left\{\gamma_{R,i} > \gamma_0\right\}\alpha_i\right\} \\ &= E_{|h_i|^2,|f_i|^2}\left\{\Pr\left\{\varpi\left|f_i\right|^2 < \left|h_i\right|^2 - \theta\right\}\frac{X}{\left|h_i\right|^2 + X}\right\} \\ &= \frac{X}{\lambda_h}\int_\theta^\infty \exp\left(-\frac{x}{\lambda_h}\right)\left(1 - \exp\left(-\frac{x-\theta}{\varpi\lambda_f}\right)\right)\frac{1}{x+X}dx \end{aligned}$$

$$= \frac{X}{\lambda_h}\exp\left(\frac{X}{\lambda_h}\right)\times\left\{\exp\left(\frac{X+\theta}{\varpi\lambda_f}\right)\right.$$

$$\left.\times Ei\left(-\frac{X+\theta}{\varpi\lambda_f}-\frac{X+\theta}{\lambda_h}\right)-Ei\left(-\frac{X+\theta}{\lambda_h}\right)\right\} \quad (19)$$

where $X = P_R d_1^m / \eta P_S$, $\varpi = \gamma_0 P_R d_1^m / P_S$, $\theta = \gamma_0 d_1^m \sigma_R^2 / P_S$, Ei is the exponential integral function as eq. 8.211 in (Jeffrey and Zwillinger, 2007). The last integral can be direved by appling eq. 3.352.2 given in (Jeffrey and Zwillinger, 2007).

Substituting (17), (18) and (19) into (16). The proposition 1 is easily derived. This is complete the proof.

Sloving (14) by using (15), the optimal transfer power at relay with aim maximization throughput can be obtained. Because of complexity in expression, so a close-form of optimal expression cannot be derived. However, an optimal value can be proven by using numerical simulation that is gathered in the next section.

4 NUMERICAL RESULT

This section uses the derived analytical results to provide the comparison with the Monte Carlo simulation. Interestingly, there is strict agreement among both cases. As a result, the validity of analytical results is verified.

We set the SINR threshold, $\gamma_0 = 10$ dB; the average, $\lambda_h = \lambda_g = 0$ dB, $\lambda_f = -20$ dB; the distance of first hop $S \rightarrow R$, $d_1 = 3$ m; distance between R node and D node, $d_2 = 1$ m; path loss, $m = 3$; the energy harvesting efficiency $\eta = 0.8$; the transmission power at source, $P_S = 26$ dB; variance of noise at destination and relay node, $\sigma_D^2 = -5$ dB and $\sigma_R^2 = -10$ dB, respectively.

Fig. 3 plots the throughput efficiency versus the different transmitted power values with the flexible changes of $\eta = 0.6, 0.8, 1$, respectively.

The throughput effeciency is maximum at approximate $P_R = 9$ dB of transmitted power, with t ≈ 0.35, 0.32, 0.27 at $\eta = 1, 0.8, 0.6$, respectively, it can be called the optimal power allocation. Besides, the throughput efficiency can be enhanced by increasing the energy conversion efficiency.

As your observation, Fig. 4 examines the impact of threshold value, γ_0 (or tranmission rate) on the throughput of systems. It can be observed from Fig. 4 that the greater throughput can be obtained with the small threshold γ_0, e.g. at $\eta = 1$, t ≈ 0.65 at $\gamma_0 = 0$ dB and t ≈ 0.35 at $\gamma_0 = 10$ dB. For the value of threshold γ_0 which is less than 16 dB, the higher the energy conversion efficiency is, the higher optimal throughput becomes. On the other hand, for the higher value of η which is greater than 16 dB, the different values of energy conversion efficiency probably don't affect the throughput of the systems.

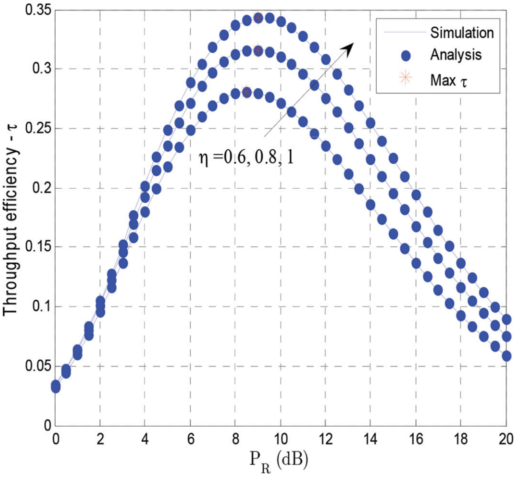


Figure 3. Throughput efficiency versus P_R with ($\eta = 0.6, 0.8, 1$).

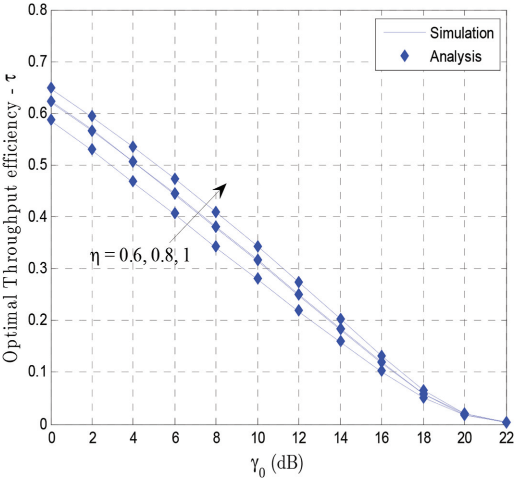


Figure 4. Throughput effeciency vesus γ_0. Other parameters $P_R = 10$ dB.

5 CONCLUSION

In this paper, a decode-and-forward wireless cooperative or sensor network with new power transfer protocol, i.e. TSAC protocol, has been considered where the harvested energy of relay node from the source can be used effectively to forward the source signal to the destination node. In order to determine the optimal throughput at the destination, analytical expression for the outage probability is derived. The simulated results in this paper have given the practical insights into the impact of various system parameters, i.e. P_R, η, γ_0, on the performance of wireless energy harvesting and information processing using DF relay nodes.

REFERENCES

Chin Keong, H. & Rui, Z. 2012. Optimal energy allocation for wireless communications with energy harvesting constraints. *Signal Processing, IEEE Transactions on,* 60, 4808–4818.

Chin Keong, H. & Rui, Z. Optimal energy allocation for wireless communications powered by energy harvesters. Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on, 13–18 June 2010. 2368–2372.

Choi, J.I., Jain, M., Srinivasan, K., Levis, P. & Katti, S. Achieving single channel, full duplex wireless communication. Proceedings of the sixteenth annual international conference on Mobile computing and networking, 2010. ACM, 1–12.

Chuan, H., Rui, Z. & Shuguang, C. 2013. Throughput Maximization for the Gaussian Relay Channel with Energy Harvesting Constraints. *Selected Areas in Communications, IEEE Journal on,* 31, 1469–1479.

Duarte, M., Dick, C. & Sabharwal, A. 2012. Experiment-Driven Characterization of Full-Duplex Wireless Systems. *Wireless Communications, IEEE Transactions on,* 11, 4296–4307.

Gang, Y., Chin Keong, H. & Yong Liang, G. 2014. Dynamic Resource Allocation for Multiple-Antenna Wireless Power Transfer. *Signal Processing, IEEE Transactions on,* 62, 3565–3577.

Jeffrey, A. & Zwillinger, D. 2007. Preface to the Seventh Edition A2 - Ryzhik, Alan JeffreyDaniel Zwillinger, I.S. Gradshteyn, I.M. *Table of Integrals, Series, and Products (Seventh Edition).* Boston: Academic Press.

Jing, Y. & Ulukus, S. 2012. Optimal Packet Scheduling in an Energy Harvesting Communication System. *Communications, IEEE Transactions on,* 60, 220–230.

Jing, Y. & Ulukus, S. Optimal Packet Scheduling in a Multiple Access Channel with Rechargeable Nodes. Communications (ICC), 2011 IEEE International Conference on, 5–9 June 2011. 1–5.

Krikidis, I., Suraweera, H.A., Smith, P.J. & Chau, Y. 2012. Full-Duplex Relay Selection for Amplify-and-Forward Cooperative Networks. *Wireless Communications, IEEE Transactions on,* 11, 4381–4393.

Love, D.J., Heath, R.W., Lau, V.K.N., Gesbert, D., Rao, B.D. & Andrews, M. 2008. An overview of limited feedback in wireless communication systems. *Selected Areas in Communications, IEEE Journal on,* 26, 1341–1365.

Nasir, A.A., Xiangyun, Z., Durrani, S. & Kennedy, R.A. 2013. Relaying Protocols for Wireless Energy Harvesting and Information Processing. *Wireless Communications, IEEE Transactions on,* 12, 3622–3636.

Ng, D.W.K. & Schober, R. Spectral efficient optimization in OFDM systems with wireless information and power transfer. Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European, 9–13 Sept. 2013. 1–5.

Ng, D.W.K., Lo, E.S. & Schober, R. 2013b. Wireless Information and Power Transfer: Energy Efficiency Optimization in OFDMA Systems. *Wireless Communications, IEEE Transactions on,* 12, 6352–6370.

Ng, D.W.K., Lo, E.S. & Schober, R. Energy-efficient resource allocation in multiuser OFDM systems with wireless information and power transfer. Wireless Communications and Networking Conference (WCNC), 2013 IEEE, 7–10 April 2013 2013a. 3823–3828.

Ozel, O., Tutuncuoglu, K., Jing, Y., Ulukus, S. & Yener, A. 2011. Transmission with Energy Harvesting Nodes in Fading Wireless Channels: Optimal Policies. *Selected Areas in Communications, IEEE Journal on,* 29, 1732–1743.

Rui, X., Hou, J. & Zhou, L. 2010. On the performance of full-duplex relaying with relay selection. *Electronics letters,* 46, 1674–1676.

Tutuncuoglu, K. & Yener, A. 2012. Optimum Transmission Policies for Battery Limited Energy Harvesting Nodes. *Wireless Communications, IEEE Transactions on,* 11, 1180–1189.

Varshney, L.R. Transporting information and energy simultaneously. Information Theory, 2008. ISIT 2008. IEEE International Symposium on, 6–11 July 2008. 1612–1616.

Xiaoming, C., Chau, Y. & Zhaoyang, Z. 2014. Wireless Energy and Information Transfer Tradeoff for Limited-Feedback Multiantenna Systems With Energy Beamforming. *Vehicular Technology, IEEE Transactions on,* 63, 407–412.

Yingbo, H., Ping, L., Yiming, M., Cirik, A.C. & Qian, G. 2012. A Method for Broadband Full-Duplex MIMO Radio. *Signal Processing Letters, IEEE,* 19, 793–796.

Zhang, R. & Ho, C.K. 2013. MIMO Broadcasting for Simultaneous Wireless Information and Power Transfer. *Wireless Communications, IEEE Transactions on,* 12, 1989–2001.

Zhou, X., Zhang, R. & Ho, C.K. 2013. Wireless Information and Power Transfer: Architecture Design and Rate-Energy Tradeoff. *Communications, IEEE Transactions on,* 61, 4754–4767.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

A performance analysis in energy harvesting full-duplex relay

Tam Nguyen Kieu, Tuan Nguyen Hoang, Thu-Quyen T. Nguyen & Hung Ha Duy
Ton Duc Thang University, Ho Chi Minh City, Vietnam

D.-T. Do
Ho Chi Minh City of Technology and Education, Ho Chi Minh City, Vietnam

M. Vozňák
VSB-Technical University of Ostrava, Ostrava, Czech Republic

ABSTRACT: In this paper, we compare the impact of some relay parameters on two relaying schemes: Amplify-and-Forward (AF) and Decode-and-Forward (DF) in full-duplex cooperative networks. Especially, closed-form expressions for the outage probability and throughput of the system is derived. Furthermore, we evaluate the dependence of system performance, in term of the outage probability and throughput, on the noise at nodes, transmission distance and relay transmission power.

1 INTRODUCTION

In recent years, a number of radio system applications which require long lifetime are facing with an obstructive challenge on energy consumption. In order to determine the throughput, a delay-limited transmittance mode is usually used to derive analytical expressions for outage probability and throughput. The Simultaneous Wireless Power and Information Transfer (SWPIT) for duplex relaying has been introduced, in which two resources communicate together over a power collecting relay. By examining the Time Switching Relay (TSR) receiving structure, the TS-based Two-Way Relaying (TS-TWR) protocol is intro-duced (Ke, Pingyi & Ben Letaief 2015a, b, Krikidis et al. 2014, Nasir 2013).

There has been research which deeply studied about the interference from Secondary Uses (SUs) to primary receivers (PTs) and from PTs to SUs in cognitive radio networks. The secondary users are able to not only transmit a packet on a licensed channel to a primary user when the selected channel is idle or occupied by the primary user but also harvest RF (radio frequency) energy from the primary users' transmissions when the channel is busy (Mousavifar et al. 2014, Sixing et al. 2014, Sixing et al. 2015, Tong et al. 2014).

In other situation, we investigate a Decode-and-Forward (DF) and Amplified-and Forward (AF) relaying system relied on radio power collection. The power constrained relay node early collects power over Radio-Frequency (RF) signals from the source node (Nasir 2014, Yiyang et al. 2013, Yuanwei et al. 2014).

The remainder of this paper is arranged as follows. Section 2 shows the system model of the EH enabled FD relaying network using delayed-limit mode in TSR. In Section 3, the outage probability and throughput analysis. Simulation results are introduced in section 4. Finally, conclusion is given in Section 5 of this paper.

2 SYSTEM MODEL

In this section, we describe the Time Switching–Based Relaying (TSR) protocol and derive expressions for the outage probability and throughput, which are considered in delay-limited transmission mode.

As in Figure 1, the suggested model is comprising of three nodes. The source node is denoted by *S*, the destination node is denoted by *D* and the relay node is denoted by *R*. Each node is equipped with two antennas, one of them is responsible for signal transmission and the other is for signal reception. The cooperative relay is assumed to be an energy constrained device so that it must harvest energy from the source, and use that energy to transfer the source information to the destination node (D). Terms g_1 and g_2 respectively represent the quasi-static block-fading channel from the source to the relay and from the relay to the destination node. In addition, terms l_1 and l_2 denote the distance from the source to the relay and from the relay to the destination, respectively.

The TSR protocol for the proposed system is illustrated in Figure 2.

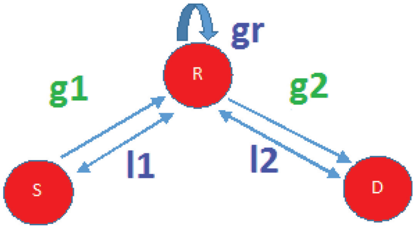


Figure 1. System model of one way full duplex relaying.

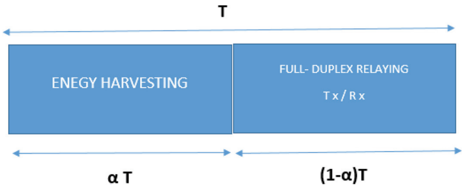


Figure 2. The parameters of TSR protocol.

The information process is separated into two stages. At first, the energy is transferred from the source to the relay within a duration of $\alpha T, (0<\alpha<1)$. The remaining time, $(1-\alpha)T$, is employed to convey information, where α is time switching coefficient and T is the duration of one signal block.

During the energy harvesting phase, the received signal at the relay node can be expressed as

$$y_R = \sqrt{\frac{P_S}{l_1^m}} g_1 x_S + n_R \tag{1}$$

where P_S is the source transmission power and m is the path loss exponent.

In this work, we assume a normalized path loss model in order to show the path loss degradation effects on the system performance. For simplicity, n_R and n_d are the zero mean Additive White Gaussian Noise (AWGN) with variance 1.

Regarding wireless received power, the harvested energy at the relay is given by

$$E_h = \eta \alpha T \frac{P_s |g_1|^2}{d_1^m} \tag{2}$$

where η is the energy conversion efficiency.

For the information transfer phase, assume that the source node transmits the signal x_S to R and R forwards signal x_R to the destination node. Both signals have unit energy and zero–mean, i.e, $E\left[|x_i|^2\right]=1$ and $E[x_i]=0$, for $i \in \{S, R\}$. Therefore, the signal received signal received at the relay under self-interference source is rewritten as

$$y_R = \sqrt{\frac{P_S}{l_1^m}} x_S g_1 + \overset{\prec}{g}_R x_r + n_R \tag{3}$$

where $\overset{\prec}{g}_R$ is the residual self-interference factor at R.

In this paper, we investigate both AF and DF schemes in duplex relaying system. For AF, the relay amplifies the signal with an amplification factor

$$K^{-1} = \sqrt{\frac{P_S}{l_1^m}|g_1|^2 + P_R \left|\overset{\prec}{g}_r\right|^2 + I} \tag{4}$$

With DF, the relay decodes signal before retransmitting it. So the transmitted signal from the relay can be expressed as follows.

$$x_R(i) = \begin{cases} K\, y_R[i-\tau] \text{ with AF} \\ \sqrt{\frac{P_R}{P_S}} x_S[i-\tau] \text{ with DF} \end{cases} \tag{5}$$

where τ accounts for the time delay bred by relay processing.

It is costly seeing that harvested power then assist operation for the next stage transmission, P_R is advanced by

$$P_R = \frac{E_h}{(1-\alpha)T} = \rho P_S \frac{|g_1|^2}{l_1^m} \tag{6}$$

where ρ is defined as $\rho = \alpha\eta/(1-\alpha)$.

Therefore, the received signal at the destination is given by

$$y_d(\mathrm{k}) = \frac{g_2}{\sqrt{l_2^m}} x_r[\mathrm{k}] + n_d[\mathrm{k}] \tag{7}$$

With AF, we have

$$y_D(k) = \underbrace{\frac{g_2}{\sqrt{l_2^m}} K \sqrt{P_R} \frac{g_1\sqrt{P_S}}{\sqrt{l_1^m}} x_S}_{signal} + \underbrace{\frac{g_2}{\sqrt{l_2^m}} K \sqrt{P_R}\, \overset{\prec}{g}_r x_R}_{RSI} + \underbrace{\frac{g_2}{\sqrt{l_2^m}} K \sqrt{P_R} n_R + n_D}_{noise} \tag{8}$$

With DF we obtain

$$y_D(t) = \sqrt{P_R} g_2 x_S(t-\tau) + n_d(t) \tag{9}$$

In the above results, the instantaneous received SINR at D through R is determined as

$$\gamma = \frac{E\left\{|signal|^2\right\}}{E\left\{|noise|^2\right\} + E\left\{|RSI|^2\right\}} \tag{10}$$

We have

$$\gamma = \frac{\dfrac{P_S|g_1|^2 P_R|g_2|^2}{l_1^m l_2^m P_R\left|\overset{\prec}{g}_r\right|^2}}{\dfrac{IP_S|g_1|^2}{P_R\left|\overset{\prec}{g}_r\right|^2 l_1^m} + \dfrac{P_R|g_2|^2}{l_2^m} + I} \tag{11}$$

We assume that the channel gains $|g_1|^2, |g_2|^2$ are independent and identically distributed (i.i.d.) exponential random variables.

3 OUTAGE PROBABILITY AND THROUGHPUT ANALYSIS

In this section, we compare the outage probability and throughput of full-duplex one-way relaying with energy harvesting and information transfer in two relaying modes: AF and DF. Based on these analytical expressions, we can see clearly some of factors imfluencing factors on system performance and learn how to deploy it in different situations.

3.1 *Outage probability analysis*

The outage probability of FD relaying network in delay-limitted model is calculated as

$$P_{out} = \Pr(\gamma \le H) \tag{12}$$

where R is target rate and $H = 2^R - 1$.

Proposition 1: the outage probability of the energy-harvesting-enabled two-way full-duplex relay with AF protocol is derived as

$$P_{out}^{AF} = \Pr\left\{\frac{\dfrac{P_S|g_1|^2 P_R|g_2|^2}{l_1^m l_2^m P_R\left|\overset{\prec}{g}_r\right|^2}}{\dfrac{IP_S|g_1|^2}{P_R\left|\overset{\prec}{g}_r\right|^2 l_1^m} + \dfrac{P_R|g_2|^2}{l_2^m} + I} < H\right\}$$

$$= 1 - \int_0^{1/\rho H} 2\sqrt{\frac{l_1^m l_2^m H\left(\dfrac{I}{\rho} + Iy\right)}{\lambda_s\lambda_d\left(P_S - \rho P_S H y\right)}} \times K_1\left(2\sqrt{\frac{l_1^m l_2^m Z\left(\dfrac{I}{\rho} + Iy\right)}{\lambda_s\lambda_d\left(P_S - \rho P_S H y\right)}}\right)\frac{1}{\lambda_r} e^{-\frac{y}{\lambda_r}} dy \tag{13}$$

where $\lambda_s, \lambda_d, \lambda_r$ are the mean value of the exponential random variables $g_1, g_{2,}, g_r$, respectively, and $K_1(x)$ is Bessel function defined as (8.423.1) in (David H. A. 1970).

Proof: We denote $x = |g_1|^2|g_2|^2$ and $y = \left|\overset{\prec}{g}_r\right|^2$. If x and y are dependent then we have

$$P_{out}^{AF} = \begin{cases} \Pr\left\{x < \dfrac{l_1^m l_2^m H\left(\dfrac{I}{\rho} + Iy\right)}{P_S - \rho P_S H y}\right\}, & y < \dfrac{1}{\rho H} \\ 1, & y > \dfrac{1}{\rho H} \end{cases} \tag{14}$$

Interestingly, the cumulative distribution function of x is calculated by

$$F_X(a) = \Pr(X < a) = 1 - 2\sqrt{a / \lambda_s\lambda_d}\, K_1\left(2\sqrt{a / \lambda_s\lambda_d}\right) \tag{15}$$

and Y can be modeled with probability distribution function $f_Y(b) = (1/\lambda_r)e^{(b/\lambda_r)}$. Then the **Proposition 1** is achieved after some simple manipulations.

Proposition 2: the outage probability of the energy-harvesting-enabled two-way full-duplex relaying with DF protocol is derived as

$$P_{out}^{DF} = 1 - \left(1 - e^{\frac{-1}{\rho\lambda_r H}}\right)\left(2\sqrt{\frac{l_1^m l_2^m Z\left(\dfrac{I}{\rho} + Iy\right)}{\lambda_s\lambda_d\left(P_S - \rho P_S Z y\right)}}\right) \times K_1\left(2\sqrt{\frac{l_1^m l_2^m H\left(\dfrac{I}{\rho} + Iy\right)}{\lambda_s\lambda_d\left(P_S - \rho P_S H y\right)}}\right) \tag{16}$$

Proof: Base on

$$\gamma_{DF} = \min\left\{\frac{1}{\rho y}, \frac{\rho P_S x}{l_1^m l_2^m I}\right\} \tag{17}$$

and the above effect, we obtain the desired result after doing some algebras.

3.2 *Optimal throughput analysis*

In Propositions 1 and 2, the outage probability of the considered model, when the relay harvests energy from the source signal and employs that power to forward source signal to the destination, is a function of distance l_1 and noise factor, I, and increases when l_1 increases from 0 to 2 and I increases from 0.1 to 1. In the delay-limited transmission mode, the transmitter is communicating at a fix transmission rate R (in bits/sec/Hz) and $(1-\alpha)T$ is the efficient information interval. Hence, the throughput of system can be written as

$$\tau = (1 - P_{out}) R \frac{(1-\alpha)T}{T} \tag{18}$$

Unfortunately, it is as high complexity to get the optimal throughput mathematically. However we can get the optimal value by numerical method as given in the next part.

4 SIMULATION RESULTS

In this section, we employ the results of derived analysis to offer perception into the variety of design options. The energy harvesting efficiency is set to be $\eta = 1$, the path loss exponent is set to be $m = 3$. For simplicity, we set $\lambda_s = \lambda_d = 1; \lambda_r = 0.1$ and $l_1 = l_2 = 1$ (except Figure 3, Figure 4) as well as $I = 1$ (except Figure 5 and Figure 6).

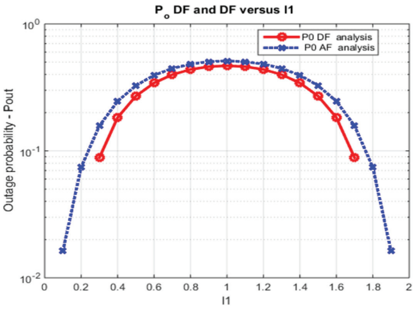


Figure 3. Outage probability of AF and DF model versus distant.

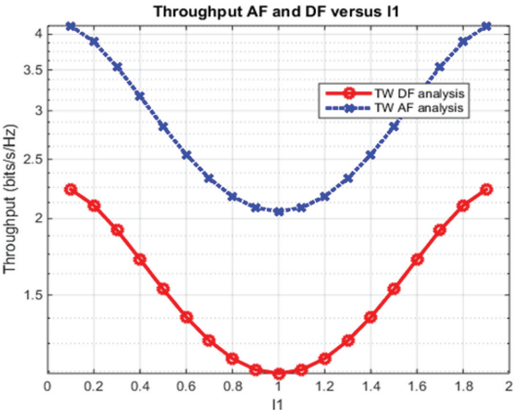


Figure 4. Throughput of AF and DF model versus distant.

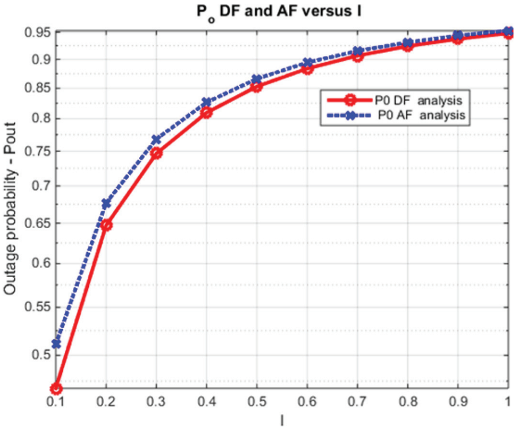


Figure 5. Outage probability of AF and DF model versus *I*.

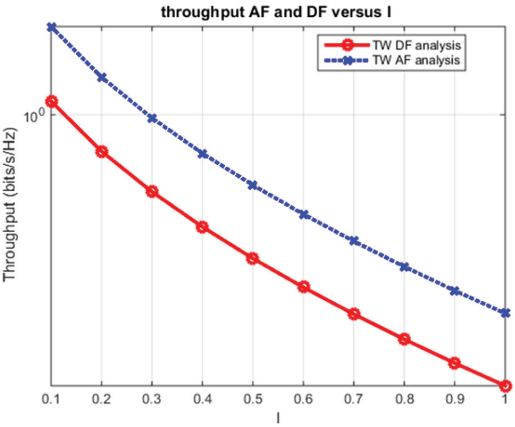


Figure 6. Throughput of AF and DF model versus I.

It can be seen from Figure 3 and Figure 4 that the outage probability of DF model is better than AF model while its throughput is worse. As close and intermediate distance, the outage probability is gradually increasing but throughput of them is contrariwise. The outage is maximum at some specific distance from S node.

The same thing happens in Figure 5 and Figure 6, the outage probability of DF model is still better than AF model but its throughput is worse than AF. This is due to noise at relay node which has impact on system performance.

5 CONCLUSION

In this paper, the mathematical and numerical analysis have shown practical insight of full-duplex relaying system in term of the effect of different system parameters on the performance of wireless energy collecting and information processing system, which employs AF and DF relay modes. The throughput results in this paper accounts for the upper bound on the realistically attainable throughput. Moreover, we also find that AF model outperforms DF model in delay—limited scheme of full-duplex relaying network.

REFERENCES

David H.A. 1970. *Order Statistics*. New York, NY, USA: Wiley.

Ke, X., Pingyi F. & Ben Letaief K. 2015a. Time-switching based SWPIT for network-coded two-way relay transmission with data rate fairness. *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 5535–5539.

Ke, X., Pingyi F. & Ben Letaief K. 2015b. Wireless Information and Energy Transfer for Two-Hop Non-Regenerative MIMO-OFDM Relay Networks. *Selected Areas in Communications, IEEE Journal on,* vol. 33, pp. 1595–1611.

Krikidis I., Timotheou S., Nikolaou S., Gan Z., Ng D.W.K. & Schober R. 2014. Simultaneous wireless information and power transfer in modern communication systems. *Communications Magazine, IEEE,* vol. 52, pp. 104–110.

Mousavifar S.A., Yuanwei L., Leung C., Elkashlan M., and Duong T.Q. 2014. Wireless Energy Harvesting and Spectrum Sharing in Cognitive Radio. *Vehicular Technology Conference (VTC Fall), 2014 IEEE 80th*, pp. 1–5.

Nasir A.A., Xiangyun Z., Durrani S. & Kennedy R.A. 2013. Relaying Protocols for Wireless Energy Harvesting and Information Processing. *Wireless Communications, IEEE Transactions on*, vol. 12, pp. 3622–3636.

Nasir A.A., Xiangyun Z., Durrani S. & Kennedy R.A. 2014. Throughput and ergodic capacity of wireless energy harvesting based DF relaying network. *Communications (ICC), 2014 IEEE International Conference on,* pp. 4066–4071.

Sixing Y., Erqing Z., Zhaowei Q., Liang Y., & Shufang L. 2014. Optimal Cooperation Strategy in Cognitive Radio Systems with Energy Harvesting. *Wireless Communications, IEEE Transactions on,* vol. 13, pp. 4693–4707.

Sixing Y., Zhaowei Q. & Shufang L. 2015. Achievable Throughput Optimization in Energy Harvesting Cognitive Radio Systems. *Selected Areas in Communications, IEEE Journal on,* vol. 33, pp. 407–422.

Tong C., Zhiguo D. & Guiyun T. 2014. Wireless information and power transfer using energy harvesting relay with outdated CSI. *High Mobility Wireless Communications (HMWC), 2014 International Workshop on,* pp. 1–6.

Yiyang N., Shi J., Ran T., Kai-Kit W., Hongbo Z. & Shixiang S. 2013. Outage analysis for device-to-device communication assisted by two-way decode-and-forward relaying. *Wireless Communications & Signal Processing (WCSP), 2013 International Conference on,* pp. 1–6.

Yuanwei L., Lifeng W., Elkashlan M., Duong T.Q. & Nallanathan A. 2014. Two-way relaying networks with wireless power transfer: Policies design and throughput analysis. *Global Communications Conference (GLOBECOM), 2014 IEEE*, pp. 4030–4035.

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

A stochastic model for performance analysis of powered wireless networks

Nhi Dang Ut & Dinh-Thuan Do
Department of Computer and Communications Engineering, HCMC University of Technology and Education, Ho Chi Minh City, Vietnam

ABSTRACT: A wireless network using a relay node to harvest energy and process information simultaneously is considered in this paper. The relay node uses the harvested energy from the source signal then it amplifies and forwards that signal to destination node. Based on two receiver architectures, namely time switching and power switching, this paper introduces stochastic model for analysis of the Time Switching based Relaying protocol (TSR) and the Time Power Switching based Receiver (TPSR), respectively. To determine the throughput at destination node, the analytical expression for the outage probability is derived for the delay-limited transmission mode. The numerical results confirm the effect of some system parameters to the optimal throughput at destination node for the network, such as the time fraction for energy harvesting, the power splitting ratio, the source transmission rate, the noise power, and the energy harvesting efficiency. More particularly, we compare the throughput at destination node between TSR protocol and ideal receiver, TSR protocol and TPSR receiver for the delay-limited transmission mode.

1 INTRODUCTION

In recent years, energy harvesting through Radio Frequency (RF) solution has been received significant research as a solution to keep the lifetime of a wireless network longer. In contrast with traditional energy supplies such as batteries or internal charging sources, energy harvesting would enable the wireless networks to operate by using energy harvested from external source such as RF signals (Varshney 2008) due to the fact that RF signals can carry energy and information at the same time.

The concept of energy harvesting and process information for an ideal receiver was first introduced by Varshney, where the author studied about the fundamental performance tradeoff for simultaneous information and power transfer. But this approach has been proved that is not available in practice since practical receivers still have their limitation to decode the carried information directly (Zhou, Zhang and Ho 2012). Other works have been done by using realizable receivers with separate receivers for energy harvesting and information processing. The work by Varshney has been extended to a frequency-selective channels with additive white Gaussian noise (Grover & Sahai 2010). Then a hybrid network is studied using stochastic-geometry model where base stations (BSs) and PBs form independent homogeneous Poisson Point Processes (PPPs) and mobiles are uniformly distributed in corresponding Voronoi cells with respect to BSs is introduced in Kaibin Huang & Vincent Lau (2014). A MIMO wireless broadcast system consisting of three nodes, where one receiver harvests energy and another receiver decodes information separately from signals is considered by Rui Zhang and Chin Keong Ho (2011). Then this work is extended by considering an imperfect channel state information at the transmitter (Xiang & Tao 2012).

In this paper, we introduce an wireless cooperative network where a relay node harvests energy from the source signal, then it Amplify-and-Forward (AF) that signal to destination node. Based on the time switching, power switching architectures (Zhou, Zhang & Ho 2012), and AF relaying protocol (Laneman, Tse & Wornell 2004), we introduce the Time-Switching based Relaying (TSR) protocol for energy harvesting and information processing at the relay node in delay-limited transmission mode. We also compare the optimal throughput observed at destination node between TSR and the time power switching relaying receiver (Dinh-Thuan Do 2015) to have a deep look inside the behavior of the system.

2 SYSTEM MODEL

Figure 1 shows the wireless system under study, where the information from the source (denoted by S), is transmitted to destination node (denoted

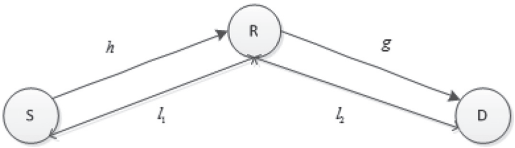


Figure 1. System model for energy constrained relay wireless network.

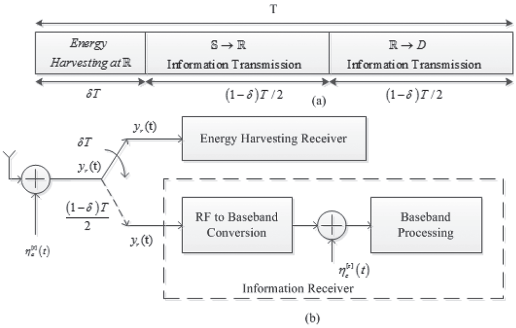


Figure 2. TSR protocol (a) TSR model and (b) Block diagram.

by *D*), through a relay node (denoted by *R*). The distance from source to relay node and from relay node to destination node are denoted by l_1 and l_2, respectively. The channel gain source to relay node and from relay node to destination node are denoted by *h* and *g*, respectively.

Based on the system model and the time switching receiver architecture, this paper introduces the time switching-based relaying and the time power switching based relaying receiver for energy harvesting and information processing from source to destination at the relay node with delay-limited transmission mode. The delay-limited transmission mode means that the destination node has to decode the received information block-by-block and as the result, the code length can not be larger than the transmission block time (Liu, Zhang & Chua 2012). The figure of merit for the system under study is the throughput at the destination node, which is defined as the number of bits are successfully decoded per unit time per unit.

3 TIME SWITCHING-BASED RELAYING (TSR) PROTOCOL

The time switching-based protocol (TSR) for energy harvesting and information processing at relay node can be seen in Figure 2.

In Figure 2, *T* is the block time to transmit information from *S* to *D*, δ is the fraction of *T* in which the relay node harvests energy from the source signal, where $0 \le \delta \le 1$. The rest of time $(1-\delta)T$, is used to transmit information from *S* to *D*. The first half remaining $(1-\delta)T/2$ time is used to transmit information from *S* to *D*, and the rest remaining half $(1-\delta)T$ is used to transmit information from *R* to *D*. The value of δ that we choose will affect to the throughput at destination node, which will be illustrated in the following sections.

3.1 *Energy harvesting and information processing at the relay node*

Figure 2b shows the block diagram for TSR protocol. The received signal at relay node, $y_r(t)$, is corrupted by a noise signal $\eta_a^{[r]}(t)$ generated by the antenna, is first sent to energy harvesting receiver. Then, in the remaining time $(1-\delta)T/2$, it is sent to information receiver. The energy harvesting receiver rectifies the received signal to get the direct current and uses that harvested energy for information processing.

The received signal, $y_r(t)$, at the relay node is given by:

$$y_r(t) = \frac{1}{\sqrt{l_1^\mu}}\sqrt{P_s}\,h\,s(t) + \eta_a^{[r]}(t) \tag{1}$$

where *h* is the channel gain from the source to relay node, l_1 is the distance from the source node to relay node, P_s is the power transmitted from the source, μ is the pathloss exponent, and $s(t)$ is the normalized information signal from the source node.

The energy harvested at relay node, denoted by E_r, is defied by:

$$E_r = \frac{\theta P_s |h|^2}{l_1^\mu}\delta T \tag{2}$$

where θ is the energy conversion efficiency and $0 \le \theta \le 1$.

After harvesting energy from $y_r(t)$, then information receiver converts $y_r(t)$ to baseband signal and processes it, this introduces an additive noise due to conversion from RF signal to baseband signal, denoted as $\eta_c^{[r]}(\mathrm{t})$. The sampled baseband signal at relay node after converted is given by:

$$y_r(k) = \frac{1}{\sqrt{l_1^\mu}}\sqrt{P_s}\,h\,s(k) + \eta_{a,r}(k) + \eta_{c,r}(k) \tag{3}$$

where *k* is the index value, $s(k)$ is the sampled-normalized signal from the source, $\eta_{a,r}(\mathrm{t})$ is the AWGN noise introduced by receiving antenna at relay node, and $\eta_{c,r}$(t) is the AWGN noise introduced by the conversion process from RF to baseband signal. The relay node then amplifies this sampled signal and transmits it. The transmitted signal from relay node, denoted as $x_r(k)$, can be expressed as:

$$x_r(k) = \frac{\sqrt{P_r}\, y_r(k)}{\sqrt{\dfrac{P_s\,|h|^2}{l^{1}_{\mu}} + \phi^2_{\eta_{a,r}} + \phi^2_{\eta_{c,r}}}} \tag{4}$$

where $\dfrac{P_s\,|h|^2}{l^{1}_{\mu}} + \phi^2_{\eta_{a,r}} + \phi^2_{\eta_{c,r}}$ is the power constraint factor, $\phi^2_{\eta_{a,r}}$ and $\phi^2_{\eta_{c,r}}$ are the variances of $\eta_{a,r}(k)$ and $\eta_{c,r}(k)$, respectively, P_r is the power transmitted from the relay node.

The signal received at destination node after sampled, denoted as $y_d(k)$, is given by:

$$y_d(k) = \frac{1}{\sqrt{l_2^{\mu}}}\, g\, x_r(k) + \eta_{a,d}(k) + \eta_{c,d}(k) \tag{5}$$

where $\eta_{a,d}(k)$ and $\eta_{c,d}(k)$ are the AWGN noises introduced by the antenna and conversion at destination node, respectively, and g is the channel gain from R to D.

Substituting $x_r(k)$ from (4) into (5), we have:

$$y_d(k) = \frac{g\sqrt{P_r l_1^{\mu}}\, y_r(k)}{\sqrt{l_2^{\mu}}\sqrt{P_s\,|h|^2 + l_1^{\mu}(\phi^2_{\eta_{a,r}} + \phi^2_{\eta_{c,r}})}} + \eta_{a,d}(k) + \eta_{c,d}(k) \tag{6}$$

And by substituting $y_r(k)$ in (3) into (6), we have:

$$y_d(k) = \frac{\sqrt{P_r\, P_s}\, h\, g\, s(k)}{\sqrt{l_2^{\mu}}\sqrt{P_s\,|h|^2 + l_1^{\mu}\phi^2_{\eta_r}}} + \frac{\sqrt{P_r\, l_1^{\mu}}\, g\, \eta_r(k)}{\sqrt{l_2^{\mu}}\sqrt{P_s\,|h|^2 + \phi^2_{\eta_r}}} + \eta_d(k) \tag{7}$$

where $\eta_r(k)$ is defined as $\eta_r(k) = \eta_{a,r}(k) + \eta_{c,r}(k)$, and $\eta_d(k)$ is defined as $\eta_d(k) = \eta_{a,\mathrm{d}}(k) + \eta_{c,\mathrm{d}}(k)$, are the overall AWGN noises at the relay and destination node, respectively, $\phi^2_{\eta_r} = \phi^2_{\eta_{a,r}} + \phi^2_{\eta_{c,r}}$ is the overall variance at relay node.

From (2), we can calculate the power transmitted from relay node as:

$$P_r = \frac{E_r}{(1-\delta)T/2} = \frac{2\theta P_s\,|h|^2\,\delta}{l_1^{\mu}(1-\delta)} \tag{8}$$

Finally, substitute P_r from (8) into (7):

$$y_d(k) = \underbrace{\frac{\sqrt{2\theta\,|h|^2\,\delta P_s}\, h\, g\, s(k)}{\sqrt{(1-\delta)l_1^{\mu}l_2^{\mu}}\sqrt{P_s\,|h|^2 + l_1^{\mu}\phi^2_{\eta_r}}}}_{signal\ part} + \underbrace{\frac{\sqrt{2\theta P_s\,|h|^2\,\delta}\, g\, \eta_r(k)}{\sqrt{(1-\delta)l_2^{\mu}}\sqrt{P_s\,|h|^2 + l_1^{\mu}\phi^2_{\eta_r}}} + \eta_d(k)}_{overal\ \mathrm{n}oise} \tag{9}$$

The received signal at destination node $y_d(k)$, is expressed by (9) in terms of P_s, θ, δ, l_1, l_2, h and g.

3.2 *Throughput analysis*

The SNR at destination node, denoted by ψ_D, can be calculated using (9) by $\psi_D = \dfrac{E\left\{|signal\ \mathrm{part}\ in(9)|^2\right\}}{E\left\{|overall\ noise\ in(9)|^2\right\}}$, is expressed by:

$$\psi_D = \frac{\dfrac{2\theta\,|h|^4\,P^2_s\,|g|^2\,\delta}{(1-\delta)l_1^{\mu}l_2^{\mu}\left(P_s\,|h|^2 + l_1^{\mu}\phi^2_{\eta_r}\right)}}{\dfrac{2\theta P_s\,|h|^2\,|g|^2\,\phi^2_{\eta_r}\,\delta}{(1-\delta)l_2^{\mu}\left(P_s\,|h|^2 + l_1^{\mu}\phi^2_{\eta_r}\right)} + \phi^2_{\eta_d}}$$
$$= \frac{2\theta P^2_s\,|h|^4\,|g|^2\,\delta}{2\theta P_s\,|h|^2|g|^2\,\phi^2_{\eta_r}\,\delta + P_s\,|h|^2\,l_1^{\mu}l_2^{\mu}\phi^2_{\eta_d}\,(1-\delta) + l_1^{2\mu}l_2^{\mu}\phi^2_{\eta_r}\,\phi^2_{\eta_d}\,(1-\delta)} \tag{10}$$

where $\phi^2_{\eta_d} = \phi^2_{\eta_{d,a}} + \phi^2_{\eta_{d,c}}$.

The throughput at destination node, denoted by ω, is determined by evaluating the outage probability, denoted as ∂_{out}, given a constant transmission rate from source node *R bits/sec/Hz*, $R = \log_2(1+\lambda_0)$, λ_0 is the threshold value of SNR for data detection at destination node. The outage probability at destination node for TSR protocol is given by:

$$\partial_{out} = p\left(\lambda_D < \lambda_0\right) \tag{11}$$

where $\lambda_0 = 2^R - 1$.

The outage probability at destination node, can be expressed analytically by:

$$\partial_{out} = 1 - \frac{1}{\gamma_h}\int_{z=d/c}^{\infty} e^{-\left(\frac{z}{\gamma_h} + \frac{a\,z+b}{\left(c\,z^2 - d\,z\,\gamma_g\right)}\right)} dz \tag{12a}$$

$$\approx 1 - e^{-\frac{d}{c\gamma_h}}\, u\, K_1(u) \tag{12b}$$

where:

$$a = P_s l_1^{\mu} l_2^{\mu} \phi_{\eta_d}^2 \lambda_0 \left(1-\delta\right) \quad (13a)$$

$$b = l_1^{2\mu} l_2^{\mu} \phi_{\eta_r}^2 \phi_{\eta_d}^2 \lambda_0 \left(1-\delta\right) \quad (13b)$$

$$c = 2\theta P_s \delta \quad (13c)$$

$$d = 2\theta P_s l_1^{\mu} \phi_{\eta_r}^2 \lambda_0 \delta \quad (13d)$$

$$u = \sqrt{\frac{4a}{c\gamma_h \gamma_g}} \quad (13e)$$

γ_h is the mean value of $|h|^2$, γ_g is the mean value of $|g|^2$ and $K_1(\bullet)$ is the first order modified Bassel function (Gradshteyn & Ryzhik 1980).

Finally, the throughput at destination node is given by:

$$\omega = \left(1-\partial_{out}\right) R \frac{\left(1-\delta\right)T/2}{T} = \frac{\left(1-\partial_{out}\right)R\left(1-\delta\right)}{2} \quad (14)$$

This is based on the fact that the transmission rate from the source is *R bits/sec/Hz* and $\left(1-\delta\right)T/2$ is the effective time to transmit information from the source node to the destination node. The throughput ω is depended on P_s, θ, δ, l_1, l_2, R, $\phi_{\eta_r}^2$ and $\phi_{\eta_d}^2$.

Following is the demonstration (Ali Nasir, Xiangyun Zhou, Salman Durrani & Rodney Kennedy 2013) for equation (12) and (13).

Substituting the value of SNR in (10) into (11), we got:

$$\begin{aligned}
\psi_D &= p\left(\frac{2\theta P_s^2 |h|^4 |g|^2 \delta}{2\theta P_s |h|^2 |g|^2 l_1^{\mu} \phi_{\eta_r}^2 \delta + P_s |h|^2 l_1^{\mu} l_2^{\mu} \phi_{\eta_d}^2 \left(1-\delta\right) + l_1^{2\mu} l_2^{\mu} \phi_{\eta_r}^2 \phi_{\eta_d}^2 \left(1-\delta\right)} < \lambda_0 \right) \\
&= p\left(|g|^2 < \frac{P_s l_1^{\mu} l_2^{\mu} \phi_{\eta_d}^2 \lambda_0 \left(1-\delta\right) |h|^2 + l_1^{2\mu} l_2^{\mu} \phi_{\eta_r}^2 \phi_{\eta_d}^2 \lambda_0 \left(1-\delta\right)}{2\theta P_s^2 \delta |h|^4 - 2\theta P_s l_1^{\mu} \phi_{\eta_r}^2 \lambda_0 \delta |h|^2} \right) \\
&= p\left(|g|^2 < \frac{a|h|^2 + b}{c|h|^4 - d|h|^2} \right) \quad (A1)
\end{aligned}$$

where:

$$a = P_s l_1^{\mu} l_2^{\mu} \phi_{\eta_d}^2 \lambda_0 \left(1-\delta\right)$$

$$b = l_1^{2\mu} l_2^{\mu} \phi_{\eta_r}^2 \phi_{\eta_d}^2 \lambda_0 \left(1-\delta\right)$$

$$c = 2\theta P_s \delta$$

$$d = 2\theta P_s l_1^{\mu} \phi_{\eta_r}^2 \lambda_0 \delta$$

The denominator in (P1), $c|h|^4 - d|h|^2$, can be even positive or negative, thus ∂_{out} is given by:

$$\begin{aligned}
\partial_{out} &= p\left(\left(c|h|^4 - d|h|^2\right)|g|^2 < \left(a|h|^2 + b\right)\right) \\
&= \begin{cases} p\left(|g|^2 < \dfrac{a|h|^2 + b}{c|h|^4 - d|h|^2}\right), & |h|^2 < d/c \\ p\left(|g|^2 > \dfrac{a|h|^2 + b}{c|h|^4 - d|h|^2}\right) = 1, & |h|^2 > d/c \end{cases}
\end{aligned} \quad (A2)$$

The second leg in (A2) is due to the fact that if $|h|^2 > d/c$, then $c|h|^4 - d|h|^2$ is a negative number and the probability of $|g|^2$ greater than negative numbers is always 1. Because of (A2), ∂_{out} is given:

$$\begin{aligned}
\partial_{out} &= \int_{z=0}^{d/c} f_{|h|^2}(z) p\left(|g|^2 > \frac{az+b}{cz^2 - dz}\right) dz \\
&\quad + \int_{z=d/c}^{\infty} f_{|h|^2}(z) p\left(|g|^2 < \frac{az+b}{cz^2 - dz}\right) dz \\
&= \int_{z=0}^{d/c} f_{|h|^2}(z) dz + \int_{z=d/c}^{\infty} f_{|h|^2}(z) \left(1 - e^{-\frac{az+b}{\left(cz^2 - dz\right)\gamma_g}}\right) dz
\end{aligned} \quad (A3)$$

where z is the integration variable, $f_{|h|^2}(z) = \frac{1}{\lambda_h} e^{-z/\gamma_h}$ is the Probability Density Function (PDF) of exponential random variable $|h|^2$, $F_{|g|^2}(z) = p(|g|^2 < z) = 1 - e^{-z/\gamma_g}$ is the Cumulative Distribution Function (CDF) of the exponential random variable $|g|^2$ and γ_g is the mean of the exponential random variable $|g|^2$. By substituting $f_{|h|^2}(z) = \frac{1}{\gamma_h} e^{-z/\gamma_h}$ in (A3), ∂_{out} is given by:

$$\partial_{out} = 1 - \frac{1}{\gamma_h} \int_{z=d/c}^{\infty} e^{-\left(\frac{z}{\gamma_h} + \frac{a z + b}{\left(c z^2 - d z\right)\gamma_g}\right)} dz \tag{A4}$$

The analytical expression of ∂_{out} for the TSR protocol presented in (12) is presented by (A4).

The integration in (A4) can not be written shorter any more. However, we can apply a high SNR approximation to get further simplified for ∂_{out} because at high SNR, the third factor in the denominator (10), $l_1^{2\mu} l_2^{\mu} \phi_{\eta_r}^2 \phi_{\eta_d}^2 (1-\delta)$, is very small when compared to the other factors, $2\theta P_s |h|^2 |g|^2 \phi_{\eta_r}^2 \delta$, and $P_s |h|^2 l_1^{\mu} l_2^{\mu} \phi_{\eta_d}^2 (1-\delta)$. So, we can re-write:

$$\psi_D \approx \frac{2\theta |h|^4 P_s^2 |g|^2 \delta}{2\theta P_s |h|^2 |g|^2 \phi_{\eta_r}^2 \delta + P_s |h|^2 l_1^{\mu} l_2^{\mu} \phi_{\eta_d}^2 (1-\delta)} \tag{A5}$$

Or we can say that at high SNR, b can be replaced by 0. Due to this, ∂_{out} in (A4) can be re-written as:

$$\partial_{out} = 1 - \frac{1}{\gamma_h} \int_{z=d/c}^{\infty} e^{-\left(\frac{z}{\gamma_h} + \frac{a z}{(c z - d)\gamma_g}\right)} dz \tag{A6}$$

Let's define $x = cz - d$. The approximated outage probability at high SNR is:

$$\begin{aligned} \partial_{out} &\approx 1 - \frac{e^{-\frac{d}{c\gamma_h}}}{c\gamma_h} \int_{x=0}^{\infty} e^{-\left(\frac{x}{\gamma_h c} + \frac{a}{x\gamma_g}\right)} dx \\ &= 1 - e^{-\frac{d}{c\gamma_h}} u K_1(\mathrm{u}) \end{aligned} \tag{A7}$$

where $u = \sqrt{\frac{4a}{c\gamma_h\gamma_g}}$, $K_1(\bullet)$ is the first-order modified Bessel function of the second kind and the last equality is obtained by using the formula,

$$\int_0^{\infty} e^{-\frac{\beta}{4x} - \lambda x} dx = \sqrt{\frac{\beta}{\lambda}} K_1\left(\sqrt{\beta\lambda}\right) \qquad [17]$$

3.3 *Time power-switching-based relaying receiver (TPSR)*

Figure 3 illustrates the block diagram for the Time Power Switching based (TPSR) receiver in which T is the block time for information to transmit from source node to destination node, δ is the fraction of the block time and $0 \le \delta \le 1$, in which δT is used for energy harvesting and the remaining time is used to transmit the signal to destination node. In term of power splitting, P is the power of the transmitted signal and $0 \le \varepsilon \le 1$, denotes the fraction of power splitting ratio for the harvested energy. The received power is divided into two parts, εP and $(1-\varepsilon)P$, which is used for energy harvesting and for signal transmission from source to destination node, respectively.

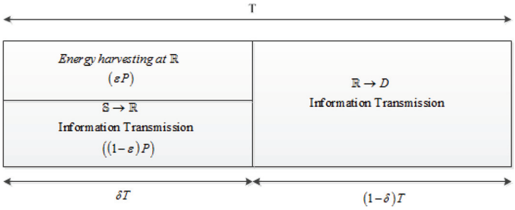


Figure 3. Block diagram for TPSR receiver.

Based on the previous analysis for TSR protocol, the harvested energy at TPSR receiver is given by:

$$E_r = \frac{\theta\delta\varepsilon P_s |h|^2}{l_1^{\mu}} \tag{15}$$

The transmitted power from relay node is:

$$P_r = \frac{E_r}{(1-\delta)T} = \frac{\theta\varepsilon P_s |h|^2 \delta}{l_1^{\mu}(1-\delta)} \tag{16}$$

The received signal at destination node after sampled is:

$$\begin{aligned} y_d(k) = &\underbrace{\frac{\sqrt{\theta\delta\varepsilon|h|^2} P_s h g s(k)}{\sqrt{(1-\delta) l_1^{\mu} l_2^{\mu}} \sqrt{P_s |h|^2 + l_1^{\mu} \phi_{\eta_r}^2}}}_{signal\ part} \\ &+ \underbrace{\frac{\sqrt{\theta\delta\varepsilon P_s |h|^2}\, g \eta_r(k)}{\sqrt{(1-\delta) l_2^{\mu}} \sqrt{P_s |h|^2 + l_1^{\mu} \phi_{\eta_r}^2}} + \eta_d(k)}_{Overall\ noise} \end{aligned} \tag{17}$$

Next, we can calculate the SNR at destination node, $\psi_D = \frac{E\left\{|\text{signal part in}\,(17)|^2\right\}}{E\left\{|\text{overall noise in}\,(17)|^2\right\}}$:

$$\psi_D = \frac{\theta\delta\varepsilon P_s^2 |h|^4 |g|^2}{\theta\delta\varepsilon P_s |h|^2 |g|^2 l_1^{\mu} \phi_{\eta_r}^2 + P_s |h|^2 l_1^{\mu} l_2^{\mu} \phi_{\eta_d}^2 (1-\delta) + l_1^{2\mu} l_2^{\mu} \phi_{\eta_d}^2 \phi_{\eta_r}^2 (1-\delta)} \tag{18}$$

The throughput at destination node for TPSR receiver with delay-limited transmission mode can be calculated based on (12) as:

$$\omega = \left(1 - \partial_{out}\right) R \frac{(1-\delta)T}{T} = \left(1 - \partial_{out}\right) R (1-\delta) \qquad (19a)$$

where:

$$a = P_s \, l_1^{\mu} \, l_2^{\mu} \, \phi_{\eta_d}^2 \, \lambda_0 (1-\delta) \qquad (19b)$$

$$b = l_1^{2\mu} \, l_2^{\mu} \, \phi_{\eta_r}^2 \, \phi_{\eta_d}^2 \, \lambda_0 (1-\delta) \qquad (19c)$$

$$c = \theta \varepsilon P_s^2 \, \delta \qquad (19d)$$

$$d = \theta \varepsilon P_s \, l_1^{\mu} \, \phi_{\eta_r}^2 \, \lambda_0 \, \delta \qquad (19e)$$

$$u = \sqrt{\frac{4a}{c \, \gamma_h \, \gamma_g}} \qquad (19f)$$

3.4 *Ideal receiver*

In this section, we introduce a ideal relay receiver, which harvests energy and processes information from the same received signal (Varshney 2008). In the first half of the block time *T*/2, it harvests energy and processes information the received signal from source node and in the remaining half, it transmits the source signal to the destination node.

The energy harvested in *T*/2 is:

$$E_r^i = \frac{\theta P_s^i |h|^2}{l_1^{\mu}} \qquad (20)$$

The power transmitted from the relay node, using harvested energy E_r^i is:

$$P_r^{\,i} = \frac{E_r^i}{T/2} = \frac{\theta P_s^i |h|^2}{l_1^{\mu}} \qquad (21)$$

At destination node, the received signal, $y_d(k)$, is expressed by:

$$y_d^i(k) = \underbrace{\frac{\sqrt{\theta |h|^2} \, P_s^i \, h \, g \, s(k)}{\sqrt{l_1^{\mu} l_2^{\mu}} \sqrt{P_s^i |h|^2 + l_1^{\mu} \phi_{\eta_r}^2}}}_{\textit{signal part}} + \underbrace{\frac{\sqrt{\theta P_s^i |h|^2} \, g \, \eta_r(k)}{\sqrt{l_2^{\mu}} \sqrt{P_s^i |h|^2 + l_1^{\mu} \, \phi_{\eta_r}^2}} + \eta_d(k)}_{\textit{Overall noise}} \qquad (22)$$

The SNR at destination node, $\psi^{i}{}_{D}$, can be calculated as $\psi^{i}{}_{D} = \frac{E\left\{|\text{signal part in (22)}|^2\right\}}{E\left\{|\text{overall noise in (22)}|^2\right\}}$, is given by:

$$\psi^{i}{}_{D} = \frac{\theta \left(P_s^i\right)^2 |h|^4 |g|^2}{\theta P^{i}{}_{s} |h|^2 |g|^2 \, l_1^{\mu} \, \phi_{\eta_r}^2 + l_1^{\mu} \, l_2^{\mu} \, \phi_{\eta_d}^2 (P_s^i |h|^2 + l_1^{\mu} \, \phi_{\eta_r}^2)} \qquad (23)$$

The throughput at the destination node for the ideal receiver with SNR given in (23) for delay-limited transmission mode is calculated by:

$$\omega^{i} = \frac{\left(1 - \partial^{i}{}_{out}\right) R}{2} \qquad (24)$$

This is due to the fact that the effective time for communication between source and destination node is *T*/2.

The outage probability, $\partial^{i}{}_{out}$, is calculated based on (12), where:

$$a = P_s^i \, l_1^{\mu} \, l_2^{\mu} \, \phi_{\eta_d}^2 \, \lambda_0 \qquad (25)$$

$$b = l_1^{2\mu} \, l_2^{\mu} \, \phi_{\eta_r}^2 \, \phi_{\eta_d}^2 \, \lambda_0 \qquad (26)$$

$$c = \theta \left(P_s^i\right)^2 \qquad (27)$$

$$d = \theta P_s^i \, l_1^{\mu} \, \phi_{\eta_r}^2 \, \lambda_0 \qquad (28)$$

4 NUMERICAL RESULS

In this section, numerical results are provided to illustrate the TSR protocol and TPSR receiver with delay-limited transmission mode. The distance from source node to relay node and distance from relay node to destination node are denoted as l_1 and l_2, respectively, $0 \le \delta \le 1$ is the fraction of block time *T*, θ is the energy harvesting efficiency.

For simplicity, we choose the source tranmission rate is $R = 3$ for default, the energy harvesting efficiency $\theta = 1$, the power transmitted from source $P_s = 1$, and the pathloss exponent $\mu = 2.7$ (corresponding to an urban cellunar network). We also assume that $\phi_{\eta_a}^2 = \phi_{\eta_{a,r}}^2 = \phi_{\eta_{a,d}}^2$, $\sigma_{n_c}^2 = \sigma_{n_{c,r}}^2 = \sigma_{n_{c,d}}^2$, and the mean value γ_h and γ_g of random variables $|h|^2$ and $|g|^2$, are set equal to 1.

Figure 4 shows the optimal throughput at destination node for TSR protocol with delay-limited transmission mode for different values of δ. As we can see, the throughput increases when δ increases,

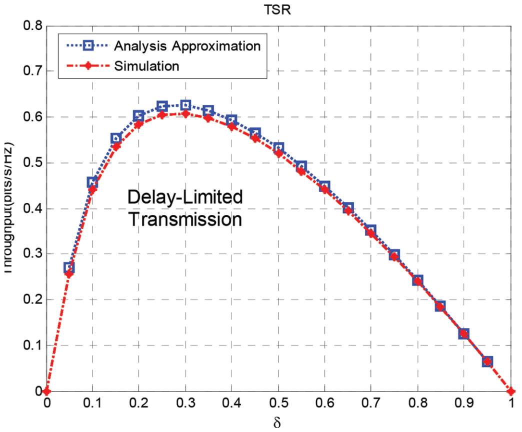


Figure 4. Throughput at destination node for TSR protocol, with $\phi^2_{\eta_a} = \phi^2_{\eta_c} = 0.01$, $P_s = 1$, $\theta = 1$, and $l_1 = l_2 = 1$.

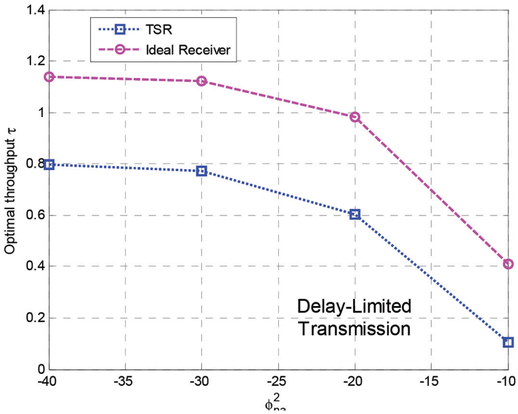


Figure 5. Optimal throughput value for TSR protocol and the ideal receiver for different values of atenna noise variance, $\phi^2_{\eta_c} = 0.01$, $P_s = 1$, $\theta = 1$, and $l_1 = l_2 = 1$.

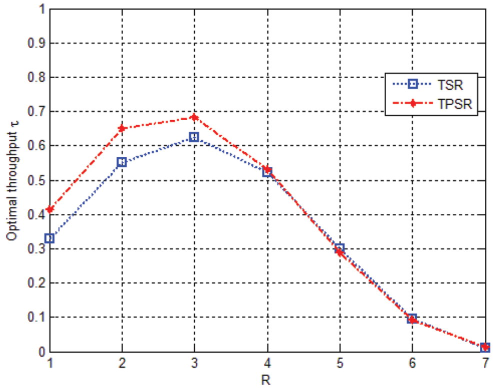


Figure 6. Optimal throughput value for TSR protocol and TPSR receiver with delay-limited transmission mode for different values of R.

but when it reaches its maximum value, approximately at $\delta \approx 0.28$, it starts to decrease. This is because when δ exceeds its optimal value, there is not much time for harvesting energy from the source signal. This leads to smaller throughput observed at destination node. For conclusion, the greater the value of δ than its optimal value, the more time is used for energy harvesting and less time to transmit signal to destination node. And the result is smaller throughput observed at destination node.

Figure 5 shows the optimal throughput for TSR protocol and the ideal receiver in comparison with the delay-limited transmission mode for different values of antenna noise variance $\phi^2_{\eta_a}$, the conversion noise variance is set to $\phi^2_{\eta_c} = 0.01$. The ideal receiver harvests and processes information from the same signal, so its optimal throughput is much better than TSR protocol. As observation, the optimal throughput for both TSR protocol and the ideal receiver increase when the antenna noise variance decreases. This is simply understood because the less noise in the signal, the better quality of signal received for information processing.

Figure 6 shows the optimal throughput value for TSR protocol and the TPSR receiver with delay-limited transmission mode for different values of the source transmission rate, R *bits/sec/Hz*. It can be seen that the throughput increases as R increases, but it will start to decrease as R is getting greater than 3. This is because the throughput in (14) depends on R, when R becomes larger, the receiver at destination node failed to decode a large mount of data incoming in the block time T. This causes the increase of the outage probability but decrease of the throughput at destination node. And, we can observe that when R is low, the optimal throughput of TPSR receiver is better than TSR protocol, but there is not much different between them when R is large.

5 CONCLUSIONS

In this paper, we consider a wireless network where a relay node harvests energy from the source signal and uses that harvested energy to forward the source signal to destination node. The throughput for delay-limited transmission mode for TSR protocol and TPSR receiver has been discussed. The throughput at destination node is determined by analytical expressions for the outage probability for the delay-limited transmission mode. The analytical results for the optimal throughput at destination node for TSR protocol and TPSR receiver is provided to let us have a deep look into the system, and to understand the effect of some system parameters to the optimal throughput value at destination node.

REFERENCES

Chalise, B.K., Zhang, Y.D., and Amin, M.G., Energy harvesting in an OSTBC based amplify-and-forward MIMO relay system, inProc. 2012 IEEE ICASSP.

Do, Dinh-Thuan., Time Power Switching based Relaying Protocol in Energy Harvesting Mobile Node: Optimal Throughput Analysis, *Mobile Information Systems Journal*, Article ID 769286, 2015.

Fouladgar, A.M., and Simeone, O., On the transfer of information and energy in multi-user systems, accepted for publication in IEEE Commun. Lett., 2012.

Laneman, J.N., Tse, D.N.C., and Wornell, G.W., Cooperative diversity in wireless networks: efficient protocols and outage behavior, IEEE Trans. Inf. Theory, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

Grover, P. and Sahai, A., Shannon meets Tesla: Wireless information and power transfer, inProc. 2010 IEEE ISIT.

Ho, C.K. and Zhang, R., Optimal energy allocation for wireless communications with energy harvesting constraints, IEEE Trans. Signal Process., vol. 60, no. 9, pp. 4808–4818, Sept. 2012.

Kaibin Huang and Vincent Lau, K.N., Enabling Wireless Power Transfer in Cellular Networks: Architecture, Modeling and Deployment, Wireless Communications, IEEE Transactions on, vol.13, 2014.

Laneman, J.N., Tse, D.N.C., and Wornell, G.W., Cooperative diversity in wireless networks: efficient protocols and outage behavior, IEEE Trans. Inf. Theory, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

Liu, L., Zhang, R., and Chua, K.C., Wireless information transfer with opportunistic energy harvesting, accepted for publication inIEEE Trans. Wireless Commun., 2012.

Luo, S., Zhang R., and Lim, T.J., Optimal save-then-transmit protocol for energy harvesting wireless transmitters, accpeted in IEEE Trans. Wireless Commun., 2013.

Nasir, Ali A., Xiangyun Zhou., Salman Durrani and Rodney Kennedy, A., Relaying Protocols for Wireless Energy Harvesting and Information Processing, IEEE transaction on wireless communications, vol. 12, no. 7, July 2013.

Rui Zhang and Chin Keong Ho, MIMO Broadcasting for Simultaneous Wireless Information and Power Transfer, IEEE Global Communications Conference (Globecom), December 5–9, 2011, Houston, USA.

Varshney, L.R., Transporting information and energy simultaneously, in Proc. 2008 IEEE ISIT.

Xiang, Z. and Tao, M., Robust beamforming for wireless information and power transmission, IEEE Wireless Commun. Lett., vol.1, no.4, pp. 372–375, 2012.

Xu, J. and Zhang, R., Throughput optimal policies for energy harvesting wireless transmitters with non-ideal circuit power, accpepted in IEEE J. Sel. Area. Commun., 2013.

Xun Zhou, Rui Zhang, and Chin Keong Ho, Wireless Information and Power Transfer: Architecture Design and Rate-Energy Tradeoff, IEEE Global Communications Conference (Globecom), December 3–7, 2012, California, USA.

Zhou, X., Zhang, R., and Ho, C.K., Wireless information and power transfer: architecture design and rate-energy tradeoff, 2012.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Energy harvesting in amplify-and-forward relaying systems with interference at the relay

Thanh-Luan Nguyen & Dinh-Thuan Do
HCMC University of Technology and Education, Ho Chi Minh, Vietnam

ABSTRACT: Harvesting energy from Radio-Frequency (RF) signals is an arising solution for prolonging the lifetime of wireless networks where relay node is energy-constrained. In this paper, an interference aided energy harvesting scheme is proposed for cooperative relaying systems, where relay harvests energy from signals transmitted from source and co-channel interferences, and then consumes that energy for forwarding the information signal to the destination. A Time Switching-based Relaying Protocol (TSR) is proposed to allow energy harvesting and information processing at the relay. Applying the proposed approach to an amplify-and-forward relaying system with the three-terminal model—the source, the relay and the destination, the approximated analytical results expressed in closed-form of the outage probability is derived to analyze the performance of the system. Furthermore the ergodic capacity, which expressed in integral-form, is derived in order to determine the achievable throughputs. In addition, the achievable throughput of the system is investigated.

1 INTRODUCTION

Nowadays wireless communication devices are developing with an incredible speed, and existing all over the world. Both size and amount of such devices are increasing every year. However, a major, and maybe a leading problem is that they consume tremendous amount of energy (Hasan 2011). Some inefficient solutions are using traditional recharging and wiring method because of numerous small devices. Another considerable answer to that problem is applying far-field microwave power transfer technique for long-distance transmission or settling and deploying additional power beacons. But the fact remains that such technology is not ready work for today communication systems and not feasible for releasing throughout the world.

A compelling solution interested by many researchers is to harvest energy from Radio Frequency (RF) radiation captured by the receive antennas to support energy-constrained communication devices. Ambient RF signal from communication devices is widely available in urban areas and can present through the night, from indoors to outdoors. These characteristics make energy harvesting a highly promising technology. In this technique, the receive antennas convert ambient RF radiation into Direct Current (DC) voltage and supply appropriate circuits (Paing 2008 & Rajesh 2011).

Since the limitation of the circuitry, one is unable to process information decoding and energy harvesting simultaneously. Inversely, the source signal carries both information and energy at the same time. Furthermore, the receiver is assumed to decode the information and harvest energy from the same signal (Varshney 2011 & Grover 2010). There are two protocols for harvesting energy and decoding information separately (Zhang, Nasir & Liu 2013, Medepally 2010), one is the Time Switching-based Relaying protocol (TSR), where the relay switch over time between decoding and harvesting; the other is the Power Splitting-based Relaying protocol (PSR), where a portion of the received power is used for energy harvesting and remaining power is used for information processing.

In cooperative networks, by setting up an intermediate relay between the source and the destination, the coverage area and capacity of communication system can effectively be enhances (Laneman 2004). However, since the relay is energy-constrained, it is difficult to prolong the lifetime of relaying systems. Optimistically, one can apply energy harvesting approach in order to achieve the desirable performance, since such information decoding and energy harvesting has advantages in wireless networks when the nodes cooperative together in transmitting the sources signal to destination. For both protocols, the Co-Channel Interference (CCI) signals supply energy in the energy harvesting phase and act as noise in the information decoding phase.

In this paper, an Amplify-and-Forward (AF) wireless cooperative network is investigated, where the relay harvests energy from the received RF signals broadcasted by a source. Impacts of Co-Channel Interferences (CCI) signals are considered. Specifically, the relays harvest energy from the information signal and the CCI signals and utilize that harvested energy to forward the source signal to its destination. The TSR receiver architecture is adopted, and the corresponding protocol is then proposed.

A three-terminal model of AF relaying is proposed, where the source node communicates with destination node through an intermediate relay node. Due to the effect of CCI signals, the outage probability is derived approximately in order to determine the outage capacity, defined as the maximum constant rate that can be maintained over fading blocks with a given outage probability. First the ergodic capacity is illustrated numerically. The corresponding achievable throughputs of the proposed energy harvesting system are also studied.

2 SYSTEM MODEL

Consider a cooperative AF relaying system, where the source, S communicates with the destination, D through an intermediate relay, R. The relay is assumed to be energy-constrained as illustrated in. Figure 1. A single antenna operated in the half-duplex mode is equipped in each node.

Figure 1 shows the system model for cooperative relaying, where the relay harvests energy from the signal transmitted from the source S and interferers and then use that energy to charge its battery.

Both, the source-to-relay and relay-to-destination transmission experience independent Rayleigh fading, with the channel gain h_S and h_D with $\mathbb{E}\{|h_S|^2\} = \Omega_s$ and $\mathbb{E}\{|h_D|^2\} = \Omega_D$ respectively, in which $\mathbb{E}\{\cdot\}$ denotes expectation operator and $|\cdot|$ is the absolute value operator.

We assume that there are M CCI signals affecting the relay. The complex channel fading gain between the *ith* interferer and the relay is denoted as l_i, with $\mathbb{E}\{|l_i|^2\} = \Omega_i$. In this paper, all channels follow a Rayleigh distribution.

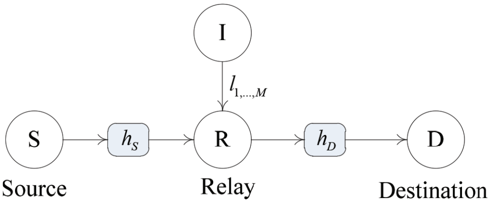


Figure 1. System model for two hop channel state information amplify-and-forward system, with co-channel interference at the relay.

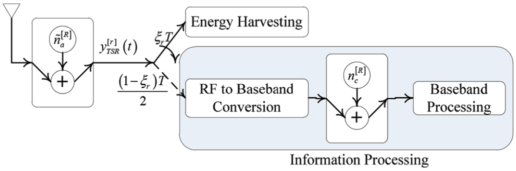


Figure 2. Block diagram of TSR protocol for energy harvesting and information processing at the relay.

Figure 2 illustrates the key parameters in the TSR protocol. Where T is the block time in which a certain block of information is transmitted from the source node to the destination node. The relay spends $\xi_r T$ block time for energy harvesting, where $0 \le \xi_r \le 1$ denotes the energy harvesting ratio and the remaining block time is divided into two equal parts, that is $(1-\xi_r)T/2$, for source-to-relay and relay-to-destination transmission, respectively.

3 TIME SWITCHING-BASED RELAYING (TSR) PROTOCOL

3.1 *Energy harvesting and information processing at the relay node*

Firstly, the source transmits its signal $s(t)$ to the relay. Accordingly, in the presence of the co-channel interference, the received signal $y_R^{TSR}(t)$ [1] at the relay node can be expressed as:

$$y_R^{TSR}(t) = h_S s(t) + \sum_{i=1}^{M} l_i s_i(t) + \tilde{n}_a^{[R]}(t) \tag{1}$$

where $\tilde{n}_a^{[R]}(t)$ is the narrowband Gaussian noise due to the receiving antenna[2], $s_i(t)$ is the signal transmitted from the *ith* interferer.

After down conversion, the sampled baseband signal at the relay node, $y_R^{TSR}(k)$ is given by

$$\begin{aligned} y_R^{TSR}(k) = h_S s(k) + \sum_{i=1}^{M} l_i s_i(k) \\ + \underbrace{n_a^{[R]}(k) + n_c^{[R]}(k)}_{\triangleq n_R^{TSR}(k)} \end{aligned} \tag{2}$$

where $s(k)$ and $s_i(k)$ is the sampled information signal from the source and the *ith* interferer, respectively; $n_a^{[R]}(k)$ denotes the baseband additive white Gaussian noise (AWGN) introduced by the receiving antenna at the relay, and $n_c^{[R]}(k)$ is the sampled AWGN due to the conversion from RF band to baseband signal, both with zero mean and variances of $N_a^{[R]}$ and $N_c^{[R]}$, respectively; and $n_R^{TSR}(k)$ is the overall AWGNs at the relay.

The harvested energy during the harvesting time, $\xi_r T$ is given by

$$E_h = \xi_e \left(\mathcal{P}_S |h_S|^2 + \sum_{i=1}^{M} \mathcal{P}_i |l_i|^2 \right) \xi_r T \tag{3}$$

where ξ_e, with $0 \le \xi_e \le 1$ is the energy conversion efficiency, its value depends upon the harvesting circuitry, $\mathcal{P}_S = \mathbb{E}\{|s(t)|^2\}$ and $\mathcal{P}_i = \mathbb{E}\{|s_i(t)|^2\}$, is the transmit power of the source and the interference sources, respectively.

The transmit power of the relay

$$\mathcal{P}_R = \frac{E_h}{(1-\xi_r)T/2} = \frac{2\xi_r \xi_e}{1-\xi_r} \left(\mathcal{P}_S |h_S|^2 + \sum_{i=1}^{M} \mathcal{P}_i |l_i|^2 \right) \tag{4}$$

Before forwarding $y_R^{TSR}(k)$ to D, the relay amplifies the received signal by multiplying it with the gain, G, which can be expressed as:

$$G = \frac{\sqrt{\mathcal{P}_R}}{\sqrt{\mathcal{P}_S |h_S|^2 + \sum_{i=1}^{M} \mathcal{P}_i |l_i|^2 + N_R}} \tag{5}$$

where $N_R = N_a^{[R]} + N_c^{[R]}$ are the variances of the overall AWGNs at the relay.

Hence, the received signal at the destination node after the sampling process, $y_D^{TSR}(k)$ is given by

$$y_D^{TSR}(k) = y_R^{TSR}(k) h_D G + \underbrace{n_a^{[D]}(k) + n_c^{[D]}(k)}_{\triangleq n_D^{TSR}(k)} \tag{6}$$

where $n_a^{[D]}(k)$ and $n_c^{[D]}(k)$ are the AWGNs at the destination node due to the antenna and conversion, both with zero mean and variances of $N_a^{[D]}$ and $N_c^{[D]}(k)$, respectively, and $n_D^{TSR}(k)$ is the overall AWGNs at the destination. By substituting $y_R^{TSR}(k)$ from (2) into (6), $y_D^{TSR}(k)$ is given by

$$\begin{aligned} y_D^{TSR}(k) &= h_S s(k) h_D G \\ &\quad + \left(\sum_{i=1}^{M} l_i s_i(k) + n_R^{TSR}(k) \right) h_D G + n_D^{TSR}(k) \end{aligned} \tag{7}$$

As a result, the SINR of the decision is given by (8) (see next page).

3.2 *Outage probability*

In this paper, the outage probability is defined as the probability that Ψ_{S2D}^{TSR} drops below an acceptable SINR threshold, γ_{th}. This concept is mathematically illustrated by

$$P_{outage} = \Pr(\Psi_{S2D}^{TSR} < \gamma_{th}) = F_{\Psi_{S2D}^{TSR}}(\gamma_{th}) \tag{8}$$

At high SNR, the outage probability at the destination node is approximately given by

$$\begin{aligned} P_{outage} &= 1 - \frac{2}{\overline{\gamma}_g} \left(\frac{\gamma_{th} \overline{\gamma}_g}{\overline{\gamma}_1} \right)^{1/2} \exp\left\{ -\frac{N_{R/D} \gamma_{th}}{\overline{\gamma}_1} \right\} K_1 \left(2\sqrt{\frac{\gamma_{th}}{\overline{\gamma}_1 \overline{\gamma}_g}} \right) \\ &\quad \times \sum_{i=1}^{\upsilon(A)} \sum_{j=1}^{\tau_i(A)} \chi_{i,j}(\mathbf{A}) \frac{\Gamma(j) \mu_{\langle i \rangle}^{-j}}{(j-1)!} \left(\frac{\gamma_{th}}{\overline{\gamma}_1} + \frac{1}{\mu_{\langle i \rangle}} \right)^{-j} \end{aligned} \tag{9}$$

where $\mathbf{A} = \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_M)$, $\mu_i = \frac{\mathcal{P}_i}{N_D} \mathbf{\Omega}_i$, $\upsilon(\mathbf{A})$ is the number of distinct diagonal elements of $\mathbf{A}$, $\mu_{\langle 1 \rangle} > \mu_{\langle 2 \rangle} > \ldots > \mu_{\langle \upsilon(\mathbf{A}) \rangle}$ are the distinct diagonal elements in decreasing order, $\tau_i(\mathbf{A})$ is the multiplicity of $\mu_{\langle i \rangle}$, and $\chi_{i,j}(\mathbf{A})$ is the (*i*, *j*)th characteristic coefficient of A (Gu & Aissa 2014).

When the interfering signals are statistically independent and identically distributed (i.i.d.), i.e., $\mu_i = \mu$, $i = 1, 2, \ldots, M$, then $\upsilon(A) = 1$ and $\tau_i(A) = M$, the outage probability, P_{outage}, is then reduced to

$$\begin{aligned} P_{outage} &= 1 - \frac{2}{\overline{\gamma}_g} \left(\frac{\gamma_{th} \overline{\gamma}_g}{\overline{\gamma}_1} \right)^{1/2} K_1 \left(2\sqrt{\frac{\gamma_{th}}{\overline{\gamma}_1 \overline{\gamma}_g}} \right) \exp\left\{ -\frac{N_{R/D} \gamma_{th}}{\overline{\gamma}_1} \right\} \\ &\quad \times \frac{\Gamma(M) \mu^{-M}}{(M-1)!} \left(\frac{\gamma_{th}}{\overline{\gamma}_1} + \frac{1}{\mu} \right)^{-M} \end{aligned} \tag{10}$$

where

$$\overline{\gamma}_1 = \frac{\mathcal{P}_S}{N_D} \Omega_S \tag{10a}$$

$$\overline{\gamma}_g = \frac{2\xi_r \xi_e}{1-\xi_r} \Omega_D \tag{10b}$$

$$N_{R/D} = \frac{N_R}{N_D} \tag{10c}$$

where $\overline{\gamma}_1$ is defined as the average signal-to-noise ratio (SNR).

Proof: See Appendix A

3.3 *Ergodic capacity and the achievable throughput*

The second parameter used to evaluate the performance of the cooperative network is the throughput which is determined by evaluating the ergodic capacity, C_E in the unit of bit/s/Hz, at the destination. In the AF-cooperative communication, using the received SINR at the destination, Ψ_{S2D}^{TSR} in (8), C_E is given by

$$C_E = \mathbb{E}\left[\log_2\left(1+\Psi_{S2D}^{TSR}\right)\right] \tag{11}$$

$$= \int_0^\infty \log_2(1+\varpi) f_{\Psi_{S2D}^{TSR}}(\varpi)\, d\varpi \tag{12}$$

$$\Psi_{S2D}^{TSR} = \frac{2\xi_r\xi_e\mathcal{P}_S|h_D|^2|h_S|^2\left(\mathcal{P}_S|h_S|^2+\sum_{i=1}^{M}\mathcal{P}_i|l_i|^2\right)}{2\xi_r\xi_e|h_D|^2\left(\mathcal{P}_S|h_S|^2+\sum_{i=1}^{M}\mathcal{P}_i|l_i|^2\right)\left(\sum_{i=1}^{M}\mathcal{P}_i|l_i|^2+N_R\right)+N_D(1-\xi_r)\left(\mathcal{P}_S|h_S|^2+\sum_{i=1}^{M}\mathcal{P}_i|l_i|^2\right)+(1-\xi_r)N_DN_R} \tag{13}$$

where $f_{\Psi_{S2D}^{TSR}}(\varpi)$ stands for the PDF of the random variable Ψ_{S2D}^{TSR} Using the integration-by-parts method, the expression in (13) can be rewritten as

$$C_E = \left\{\log_2(1+\varpi)\left[F_{\Psi_{S2D}^{TSR}}(\varpi)-1\right]\right\}_0^\infty - \frac{1}{\ln 2}\int_0^\infty \frac{1}{1+\varpi}\left[F_{\Psi_{S2D}^{TSR}}(\varpi)-1\right]d \tag{14}$$

$$= \frac{1}{\ln 2}\int_0^\infty \frac{1}{1+\varpi}\left[1-F_{\Psi_{S2D}^{TSR}}(\varpi)\right]d\varpi \tag{15}$$

where $\{f(x)\}_a^b \triangleq f(b)-f(a)$.

The throughput at the destination depends only on the effective information time, $(1-\xi_r)T/2$ and is given by

$$\tau_E = \frac{(1-\xi_r)T/2}{T}C_E = \frac{1-\xi_r}{2}C_E \tag{16}$$

3.4 *Outage Capacity and the achievable throughput*

Outage capacity, in the unit of bit/s/Hz, is defined as the maximum constant rate that can be maintained over fading blocks with a specified outage probability (Gu & Aissa 2015). In the AF cooperative communication system under study, the outage capacity is expressed as

$$C_O = \left[1-P_{outage}(\gamma_{th})\right]\log_2(1+\gamma_{th}) \tag{17}$$

The achievable throughput at the destination relates only to the transmission time is given by

$$\tau_O = \frac{1-\xi_r}{2}C_O \tag{18}$$

4 NUMERICAL RESULTS

In this section, the approximated analytical results are derived. Monte Carlo simulation results illustrated to corroborate the proposed analysis. To evaluate the effects of the interferences on the system's performance, we define the average signal-to-interference ratio as $SIR \triangleq \frac{\bar{\gamma}_1}{\bar{\gamma}_{INF}}$. Here after, and unless stated, the variances are assume to be identical, that is, $N_a^{[R]} = N_c^{[R]} = N_a^{[D]} = N_c^{[D]}$, number of interferers is set to 1 ($M = 1$) and the values of the energy conversion efficiency is set to 1 ($\xi_e = 1$).

Figure 3 shows the throughput τ_E and τ_O versus the energy harvesting ratio ξ_r for different values of average SIR received at the relay, where the average SNR is fixed at 20 dB and $\gamma_{th} = 5$ dB. The analytical and simulation results of ergodic capacity are from (15) and (12), respectively. It is observed that the analytical results match well the simulation results. In general, the throughput increases as ξ_r increases to some optimal value. But later,

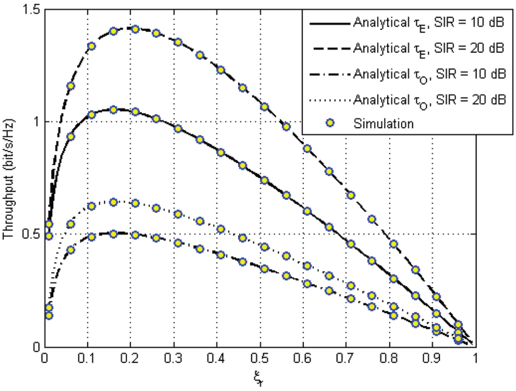


Figure 3. Throughput τ_E and τ_O versus the energy harvesting ratio, ξ_r for different values of SIR, where SNR is fixed at 20 dB, with $N_R = N_D = 1$ and $\gamma_{th} = 5$ dB.

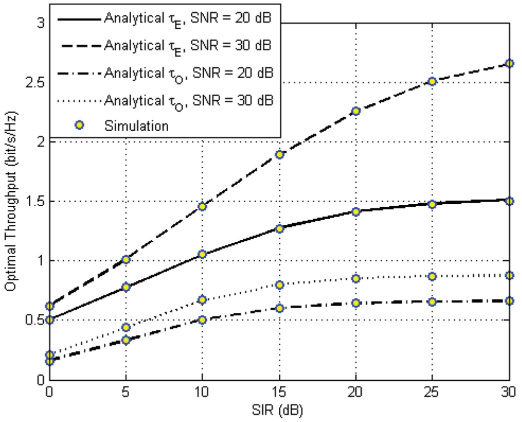


Figure 4. Optimal throughput τ_E and τ_O versus the average SIR for different values of average SNR, with $N_R = N_D = 1$ and $\gamma_{th} = 5$ dB.

as ξ_r increases from its optimal value, more data is wasted on energy harvesting resulting that the throughput of the system gradually drops down from its maximum value. Furthermore, when the average SIR increases the optimal throughput is also increase. This is implies that an increase in power of the CCI signals can deteriorate the system performance, but reduces the ξ_r required to achieve the same value of throughput.

Figure 4 shows the Optimal throughput τ_E and τ_O versus the average SIR for different values of average SNR, where $\gamma_{th} = 5$ dB. It also shows that for a given average SNR, the optimal throughput increases as the average SIR increases. This implies that an increasing in the average SIR can effectively enhance the system throughput. This implies that in order to enhance the system's throughput, we can either increase the signal power or by decreasing the noise variances.

5 CONCLUSION

In this paper, an interference aided energy harvesting amplify-and-forward relaying system was proposed, where the energy-constrained relay harvests energy from the received information signal and Co-Channel Interference (CCI) signals, then uses that harvested energy to forward the signal to destination after multiply it the gain. The time switching-based relaying protocol was adopted here for circuit simplicity.

The achievable throughput of the system is numerically derived from the ergodic capacity and analytically derived from the outage capacity. The outage probability is calculated approximately at high SNR for simplicity.

It is shown that when the SNR is fixed, an increase in the power of the CCI signals can reduce the system performance, but required less time for energy harvesting at the relay. In order to enhance the system throughput, one can either increase the power of the information signal or decrease the noise variances.

APPENDIX A

At high SNR, the third factor in the denominator of (8), $(1-\xi_r)N_D N_R$ can be ignored since its value is too small compared to the other two factors in the denominator, $2\xi_r\xi_e\left|h_D{}^2\right|\left(\mathcal{P}_S\left|h_S\right|^2+\sum_{i=1}^{M}\mathcal{P}_i\left|l_i\right|^2\right)\left(\sum_{i=1}^{M}\mathcal{P}_i\left|l_i\right|^2+N_R\right)$ and $N_D(1-\xi_r)(\mathcal{P}_S\,|\,h_S\,|^2+\sum_{i=1}^{M}\mathcal{P}_i\,|\,l_i\,|^2$. As a result, the approximated SINR at the destination applying high SNR approximation is given by

$$\Psi_{S2D}^{TSR} \approx \frac{\gamma_1}{\gamma_{INF}+\dfrac{1}{\gamma_g}+N_{R/D}} \tag{A.1}$$

where,

$$\gamma_1 = \frac{\mathcal{P}_S}{N_D}\left|h_S\right|^2 \tag{A.1a}$$

$$\gamma_{INF} = \sum_{i=1}^{M}\frac{\mathcal{P}_i}{N_D}\left|l_i\right|^2 \tag{A.1b}$$

$$\gamma_g = \frac{2\xi_r\xi_e}{1-\xi_r}\left|h_D\right|^2 \tag{A.1c}$$

$$N_{R/D} = \frac{N_R}{N_D} \tag{A.1d}$$

In order to find P_{outage}, the cumulative density function, $F_{\Psi_{S2D}^{TSR}}(\gamma_{th})$ is approximately given by

$$F_{\Psi_{S2D}^{TSR}}(\gamma_{th}) = \int_0^\infty\int_0^\infty \Pr\left(\gamma_1 < \gamma_{th}\left(y+\frac{1}{z}+N_{R/D}\right)\right) \times f_{\gamma_g}(z)f_{\gamma_{INF}}(y)dydz \tag{A.2}$$

where $f_{\gamma_g}(z)$ and $f_{\gamma_{INF}}(y)$ denotes the probability density function (PDF) of γ_g and γ_{INF}, respectively. The PDF of γ_{INF} is given by (for details on this analysis, see Bletsas, H. Shin, and M. Z. Win (2007))

$$f_{\gamma_{INF}}(y) = \sum_{i=1}^{v(A)}\sum_{j=1}^{\tau_i(A)}\chi_{i,j}(A)\frac{\mu_{\langle i\rangle}^{-j}}{(j-1)!}y^{j-1}\exp\left\{-\frac{y}{\mu_{\langle i\rangle}}\right\} \tag{A.3}$$

If the interfering signals are i.i.d., the CDF of γ_{INF} reduces to

$$f_{\gamma_{INF}}(y) = \frac{\mu^{-M}}{(M-1)!} y^{M-1} \exp\left\{-\frac{y}{\mu}\right\} \tag{A.4}$$

In addition, to evaluate $F_{\Psi_{S2D}^{TSR}}(\gamma_{th})$, we also need to determine the CDF and PDF of RVs, γ_1 and γ_g, respectively. Note that, the CDF of γ_1 and PDF of γ_g can be expressed as $F_{\gamma_1}(\gamma) = 1 - \exp\{-\frac{\gamma}{\overline{\gamma}_1}\}$ and $f_{\gamma_g}(z) = \frac{1}{\overline{\gamma}_g}\exp\{-\frac{z}{\overline{\gamma}_g}\}$, respectively. Substituting (A.3) into (A.2) and with the given CDF of γ_1 and PDF of γ_g we have

$$\begin{aligned} F_{\Psi_{S2D}^{TSR}}(\gamma_{th}) = 1 - \frac{1}{\overline{\gamma}_g}\exp\left\{-\frac{N_{R/D}\gamma_{th}}{\overline{\gamma}_1}\right\} \sum_{i=1}^{v(A)} \sum_{j=1}^{\tau_i(A)} \chi_{i,j}(A) \frac{\mu_{\langle i \rangle}^{-j}}{(j-1)!} \\ \times \int_0^\infty y^{j-1} \exp\left\{-\left(\frac{\gamma_{th}}{\overline{\gamma}_1} + \frac{1}{\mu_{\langle i \rangle}}\right) y\right\} \\ \times dy \int_0^\infty \exp\left\{-\frac{\gamma_{th}}{\overline{\gamma}_1 z} - \frac{z}{\overline{\gamma}_g}\right\} dz \end{aligned} \tag{A.5}$$

The two integrals can be evaluated as follow (Prudnikov, Brychkov & Marichev 1986. *Integrals and Series,* eq. (2.3.3.1) and eq. (2.3.16.1))

$$I_1 = \int_0^\infty y^{j-1} \exp\left\{-\left(\frac{\gamma_{th}}{\overline{\gamma}_1} + \frac{1}{\mu_{\langle i \rangle}}\right) y\right\} dy = \Gamma(j)\left(\frac{\gamma_{th}}{\overline{\gamma}_1} + \frac{1}{\mu_{\langle i \rangle}}\right)^{-j} \tag{A.6}$$

$$I_2 = \int_0^\infty \exp\left\{-\frac{\gamma_{th}}{\overline{\gamma}_1 z} - \frac{z}{\overline{\gamma}_g}\right\} dz = 2\left(\frac{\gamma_{th}\overline{\gamma}_g}{\overline{\gamma}_1}\right)^{1/2} K_1\left(2\sqrt{\frac{\gamma_{th}}{\overline{\gamma}_1 \overline{\gamma}_g}}\right) \tag{A.7}$$

where $K_1(\cdot)$ stands for the first-order modified Bessel function of the second kind, $\Gamma(j)$ denotes the Gamma function.

The approximated outage probability, P_{outage} in (10) is achieved by substituting (A.6) and (A.7) into (A.5).

REFERENCES

Bletsas, H. Shin, & Win, M.Z. 2007. Cooperative communications with outage-optimal opportunistic relaying. *IEEE Trans. Wireless Commun.,* 6: 3450–3460.

Grover, P. & Sahai, A. 2010. Shannon meets Tesla: Wireless information and power transfer. *Proc. 2010 IEEE Int. Symp. Inf. Theory*: 2363–2367.

Gu, Y. & Aissa, S. 2015. RF-Based energy harvesting in Decode-and-Forward relaying systems: ergodic and outage capacities. *Proc. IEEE Int. Conf. Commun.*: 6425–6434.

Gu, Y. & Aissa, S. 2014. Interference aided energy harvesting in decode-and forward relaying systems. *Proc. IEEE Int. Conf. Commun.*: 5378–5382.

Hasan, Z.H., Boostanimehr, & Bhargava, V. K. 2011. Green cellular networks: A survey, some research issues and challenges. *IEEE Commun.Surveys Tuts.,* 13: 524–540.

Laneman, J.N., Tse, D.N.C. & Wornell, G.W. 2004. Cooperative diversity in wireless networks: Efficient protocols and outage behaviour. *IEEE Trans. Inf. Theory,* 50: 3062–3080.

Liu, L., R. Zhang, & Chua, K.C. 2013. Wireless information transfer with opportunistic energy harvesting. *IEEE Trans. Wireless Commun.,* 12: 288–300.

Medepally, B. & Mehta, N.B. 2010. Voluntary energy harvesting relays and selection in cooperative wireless networks. *IEEE Trans. Wireless Commun.,* 9: 3543–3553.

Nasir, A.A., Zhou, X., Durrani, S. & Kennedy, R.A. 2013. Relaying protocols for wireless energy harvesting and information processing. *IEEE Trans. Wireless Commun.,* 7: 3622–3636.

Paing, T., Shin, J., Zane, R. & Popovic, Z. 2008. Resistor emulation approach to low-power RF energy harvesting. *IEEE Trans. Power Elec.,* 23: 1494–1501.

Prudnikov, A.P., Brychkov, Y.A. & Marichev, O.I. (1 & 2) 1986. *Integrals and Series*. New York: Gordon and Breach Science Publishers.

Rajesh, R., Sharma, V. & Viswanath, P. 2011. Information capacity of energy harvesting sensor nodes. *Proc. 2011 IEEE Int. Symp. Inf. Theory*: 2363–2367.

Varshney, L.R. 2008. Transporting information and energy simultaneously. *Proc. 2008 IEEE Int. Symp. Inf. Theory*: 1612–1616.

Zhang, R. & Ho, C.K. 2013. MIMO broadcasting for simultaneous wireless information and power transfer. *IEEE Trans. Wireless Commun.,* 12: 1989–2001.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

On a tandem queueing network with breakdowns

A. Aissani
Department of Computer Science, University of Science and Technology Houari Boumediene (USTHB), El Alia, Bab-Ez-Zouar, Algiers, Algeria

ABSTRACT: The purpose of this paper is to provide a method for finding the probability distribution of the virtual waiting time of a customer in a closed queueing network with two stations in tandem and unreliable servers. We obtain the joint probability distribution of the server state (busy or out of order) and the residual work in each station. Then we derive the probability distribution of the virtual waiting time of a customer in the network in terms of its Laplace-Stieltjes transform. These results are interesting to provide some performance metrics such as the utilization or the load of a central node (or base station) in physical networks such as mobile or wireless networks (WSN), data bases and other telecommunication or computer systems.

1 INTRODUCTION

Queueing network models are interesting tools when we want to take into account the effect of packet traffic or routing protocols on the performance of a real physical network (Boucherie & Dijk 2010, Medvediev 1978, Demirkol, Ersoy, Alagz, & Deli 2009, Qiu, Feng, Xia, Wu, & Zhou 2011, Senouci, Mellouk, & Aissani 2012). We consider a Closed Queueing Network of two single server nodes (or stations) S_1 and S_2 in tandem in which circulates a constant number N of requests (customers, packets, etc). Such a model (see Figure 1) is used in many systems with multiple access in which we consider a Central Node (or Base Station) against the rest of the network considered as a bloc of data transmission. For example, the work (Osman & Knottenbelt 2012) presents a categorization of queueing network performance models of database systems. They considers amongst others the transaction processing model in which the central node or server represents the hardware components (for example the CPU) of a centralized database or a site in a distributed database.

A more elaborated example concerns modeling issues in Wireless Sensor Networks (WSNs) for performance evaluation purpose. WSNs are widely used to implement low cost non-attended monitoring of different environments. WSNs operate in a complex real-time and real world noisy environment, which gives rise to several modeling challenges regarding to the quantitative protocol evaluation for QoS (Quality of Service) goals. Recent applied and theoretical research focus amongst others on: (i) deployment of the sensors (location and coverage issues); (Senouci, Mellouk, & Aissani 2014, Senouci, Mellouk, Oukhellou, & Aissani 2015) (ii) energy efficiency due to low-power and low-cost devices with limited capabilities (sensing, data processing, transmission range, memory, communication); (iii) saturation throughput analysis; (iv) end-to-end delay reflecting the time needed by a message to traverse one hop of its multi-hop path to the sink node (Qiu, Xia, Feng, Wu, & Jin 2011). Such an aspect copes with retrial behavior of the message-sending node and the active/sleep periods of the potential next-hope nodes (Phung-Duc 2012); (v) routing protocols (Qiu, Xia, Feng, Wu, & Jin 2011); (vi) reliability and maintainability issues, which can be understood in different ways according to the routing protocol used for the physical purpose (Senouci, Mellouk, & Aissani 2012) etc.

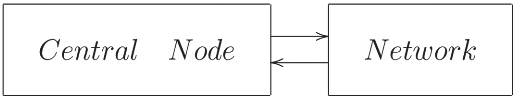


Figure 1. Model of a network with central node.

In this context of WSN's we assume that the deployment is achieved yet. Moreover, since the second node represents the rest of the network, so the exponential and Poisson assumptions about arrivals and breakdowns are reasonable by virtue of limit theorems of probability. The network lifetime strongly depends on the routing protocol used and can be defined in several ways. For example, the lifetime can be defined as the time elapsed until the first node (or last) depletes its energy (dies and cannot provides service). In some scenarios, such as intrusion or fire detection, it is necessary that all nodes stay alive as long as possible, since network quality decreases as soon one node dies. In these scenarios, it is important to know when the first node dies, the FND (resp.LND) metric (First node dies) gives an estimated value for this event. The HNA metric (Half of the nodes Alive) gives an estimated value for the case when the loss of a single or a few nodes does not automatically reduce the QoS. Now, since the sensor nodes are placed at different distances from the base station, the network lifetime distribution can be studied using these three metrics from the spatio-temporal point of view (Senouci, Mellouk, & Aissani 2012).

In this paper, we introduce the probability distributions of the lifetime of the node which fits to any of the above metrics according to the modeling level. When a node "die", a random interruption period (for corrective maintenance) begins in order to renewal the node to the state "as-good-as-new".

We assume that the lifetime of the central node S_1 is arbitrary distributed with Probability Distribution Function (PDF) $D_1(t)$ while the lifetime of the node S_2 is exponentially distributed with parameter θ_2 .

After breakdown, the renewal of the node begins immediately and it's duration is a random variable arbitrary distributed with PDF $R_1(t)$ in S_1 and exponentially distributed with parameter ν_2 in S_2 . The service time of a request is also exponentially distributed with parameter μ_i in the station $S_i, i=1,2$. We assume that the three sequences of lifetimes, renewal times, service times are mutually independent and identically distributed sequences of random variables.

After service in the node S_i , the served request joins immediately the node $S_j, i, j \in 1,2, j \neq i$. If a new request finds in S_j the server available i.e. free of requests and in functioning state, then its service begins immediately. Otherwise, he joins a First in-First out queue and waits the beginning of service without any constraint on the duration of the waiting time.

Concerning the evolution of the request whose service was interrupted, we consider the following two schemes.

i. After the renewal of the node, the service of the interrupted request continues from the point at which it was interrupted (Scheme 1).
ii. After renewal, a new service begins (Scheme 2).

The purpose of this paper is to provide a method for finding the probability distribution of the virtual waiting time of a request in such a Queueing Network. In the following section, we provide a technical remark which help us to simplify the considered problem. In section 3, we describe the basic stochastic process describing the evolution of our network of queue. In section 4 we derive a partial differential system of equations for the probability distribution of the system state of the basic stochastic process describing the evolution of our queueing network. Section 5 is devoted to the resolution of this system of equations in stationary regime. In section 6, we show how this solution can be used to derive the probability distribution of the virtual waiting time of a request in the central node of the network. Finally, in section 7 we show some applications and numerical examples.

2 SIMPLIFICATION OF THE PROBLEM

Let $G_1(t)$ be the distribution function of the generalized service time (Gaver 1962, Taleb. & Aissani 2010) (also called completion time (Aissani & Artalejo 1998)) of an arbitrary request in the base station S_1 i.e the time from a first access of the request at the server until he leaves the system with completed service.

Denote also by $H_i(t)$ the distribution functions of the service time in station S_i , $D_i(t)$ the distribution of the lifetime of the server in station S_i and $R_i(t)$ the distribution of the renewal time in S_i , $i=1$ or 2.

The corresponding Laplace-Stieltjes transforms are denoted by $g_i(s)$, $d_i(s)$, $r_i(s)$, $h_i(s)$, $Re(s)>0$, $i=1$ or $i=2$.

It can be shown that the Laplace-Stieltjes transform of the distribution $G_1(t)$ is given by

- Scheme 1.

$$g_1(s)=\int_0^\infty e^{-st}P\{r_1(s),t\}dH_1(t),$$

$$\widetilde{P}(x,s)=\int_0^\infty e^{-st}P(x,t)dt=\frac{1}{s}\times\frac{1-d_1(s)}{1-xd_1(s)},$$

Here $p_k(t)$ is the probability that k breakdowns occurs during the service of the marked customer and

$$P(x,t)=\sum_0^\infty x^k P_k(t).$$

is the generating function (or z -transform) of this probability distribution relatively to the discrete variable k . So, $\widetilde{P}(x,s)$ represents the Laplace transform relatively to the continuous real variable t .

- Scheme 2.

$$g_1(s)=\frac{\alpha_{12}(s)}{1-\alpha_{11}(s)r_1(s)},$$

where

$$\alpha_{11}(s) = \int_0^\infty \int_0^t [1 - H_1(x)]dD_1(x),$$

$$\alpha_{12}(s) = \int_0^\infty \int_0^t [1 - D_1(x)]dH_1(x).$$

Recall that in section 1, we have assumed that the functions $D_2(t)$, $R_2(t)$, $H_2(t)$ corresponds to exponential distribution functions. So, by elementary algebra, we show that

$$d_2(s) = \int_0^\infty e^{-st} dD_2(t) = \frac{\theta_2}{s + \theta_2},$$

$$r_2(s) = \int_0^\infty e^{-st} dR_2(t) = \frac{\nu_2}{s + \nu_2},$$

$$d_2(s) = \int_0^\infty e^{-st} dD_2(t) = \frac{\mu_2}{s + \mu_2}.$$

Now, according to the remark of several authors (Gaver 1962) (for FIFO Queues) the waiting time of an arbitrary (but marked) request can be decomposed into the sum of the proper service time, and all the renewal times of the breakdowns occurring during this service.

From a request point of view, it is not important what time has been devoted to the service itself, or to the reparation of the breakdown. The only important thing is at what time the server will be available to accept a new request in service. So, we will assume that the server in the station S_1 is absolutely reliable and the service time in this station follows a probability distribution $G_1(t)$.

Similar arguments can be provided from the point of view of the second node S_2.

3 THE BASIC STOCHASTIC PROCESS

We introduce the following notations. Let $e(t) = 0$, if the server in S_1 is free of request and available; $e(t) = 1$, if the server is busy or out of order.

$\beta(t) = 0$, if the server in S_2 is free of requests; $\beta(t) = 1$, if the server is busy.

$\alpha(t) = 0$, if the server in S_2 is available; $\alpha(t) = 1$, if it is out of order.

We introduce also $\gamma(t)$ a continuous random variable which is equal to

- the period from t until the moment of breakdown of the server S_1, if $e(t) = 0$ and after t there is no arrival in S_1,
- period from t until the beginning of a request which will be arrived at time t if $e(t) = 1, \beta(t) \neq 0$,
- period from t until the server S_1 achieve the service of all customers if $e(t) = 1, \beta(t) = 0$.

Consider the following stochastic process $\zeta(t) = \{e(t), \alpha(t), \beta(t); \gamma(t)\}$ defined on the state space $E = \{0,1\} \otimes \{0,1\} \otimes \{0,1\} \otimes \mathrm{R}^+$

It is not difficult to show that the process $\{\zeta(t), t \geq 0\}$ is a linear Markov process with spontaneous variations of state in the sense of (Gnedenko & Kovalenko 1989). This process is visibly ergodic, since the embedded Markov chain is finite, irreducible and aperiodic.

4 SYSTEM STATE OF EQUATIONS

In this section we derive the system state equations.

The functions

$$F_{ij}^{(k)}(x,t) = P\{e(t) = k, \alpha(t) = i, \beta(t) = j, \gamma(t) < x\}, \quad i, j, k \in \{0,1\}; x \geq 0,$$

are solutions of the following system of partial-differential equations

$$\frac{\partial F_{01}^{(0)}(x;t)}{\partial t} - \frac{\partial F_{01}^{(0)}(x;t)}{\partial x} + \frac{\partial F_{01}^{(0)}(0;t)}{\partial x}$$
$$= \frac{\partial F_{01}^{(1)}(0;t)}{\partial x} D_1(x) + \nu_2 F_{11}^{(0)}(x;t)$$
$$- (\mu_2 + \theta_2) F_{01}^{(0)}(x;t),$$

$$\frac{\partial F_{11}^{(0)}(x;t)}{\partial t} - \frac{\partial F_{11}^{(0)}(x;t)}{\partial x} + \frac{\partial F_{11}^{(0)}(0;t)}{\partial x}$$
$$= \frac{\partial F_{11}^{(1)}(0;t)}{\partial x} D_1(x) + \theta_2 F_{01}^{(0)}(x;t) - \nu_2 F_{11}^{(0)}(x;t),$$

$$\frac{\partial F_{01}^{(1)}(x;t)}{\partial t} - \frac{\partial F_{01}^{(1)}(x;t)}{\partial x} + \frac{\partial F_{01}^{(1)}(0;t)}{\partial x}$$
$$= \frac{\partial F_{01}^{(0)}(0;t)}{\partial x} R_1(x) + \mu_2 \frac{\partial F_{01}^{(0)}(\infty;t)}{\partial x} G_1(x)$$
$$+ \mu_1 F_{00}^{(1)}(\infty;t) G_1^{(N-1)}(x) - (\mu_2 + \theta_2) F_{01}^{(1)}(x;t)$$
$$+ \mu_2 \int_0^\infty G_1(x - y) \frac{\partial F_{01}^1(y;t)}{\partial y} dy,$$

$$\frac{\partial F_{11}^{(1)}(x;t)}{\partial t} - \frac{\partial F_{11}^{(1)}(x;t)}{\partial x} + \frac{\partial F_{11}^{(1)}(0;t)}{\partial x}$$
$$= \frac{\partial F_{11}^{(0)}(0;t)}{\partial x} R_1(x) + + \theta_2 F_{01}^{(1)}(x;t)$$
$$- \nu_2 F_{11}^{(1)}(x;t) + \mu_1 F_{10}^1(\infty;t) G_1^{(N-1)}(x),$$

$$F_{00}^{(1)}(x;t) = p_1(t) \int_0^\infty G_1^{(N-1)}(x - y) dR_1(y)$$
$$+ p_0(t) G_1^{(N)}(x),$$

$$F_{10}^{(1)}(x;t) = q_1(t) \int_0^\infty G_1^{(N-1)}(x - y) dR_1(y)$$
$$+ q_0(t) G_1^{(N)}(x),$$

where

$$p_i(t) = P\{e(t)=1, \alpha(t)=0, \beta(t)=0, \alpha_1(t)=i\},$$
$$q_i(t) = P\{e(t)=1, \alpha(t)=1, \beta(t)=0, \alpha_1(t)=i\},$$
$$i = 0,1.$$

Here $\alpha_1(t)$ is the number of unavailable servers in the system S_2 and $G_1^k(.)$ is the k-th order convolution of the function $G_1(.)$.

These equations can be obtained by the usual way (Gnedenko & Kovalenko 1989). The idea is to observe the evolution of the basic stochastic process during an infinitesimal small interval of time $(t, t+h)$.

The random event $e(t+h)=0$, $\alpha(t+h)=0$, $\beta(t+h) = 1$ with $\gamma(t+h) < x$ (which probability is $F_{01}^{(0)}(x; t+h)$) occurs if and only if one of the following events holds

- either from the state $e(t)=1, \alpha(t)=0, \beta(t)=1$ with $0 \le \gamma(t) < h$ (the probability of such a random event is $F_{01}^{(1)}(h;t)$) and the duration to the next breakdown of the server in S_1 is less than x (the probability of such an event is $D_1(x)$);
- either $e(t)=0, \alpha(t)=1, \beta(t)=1)$ with $h \le \gamma(t) < x+h$ (with probability $F_{11}^{(0)}(x;t)$ and the breakdown of the server in S_2 has been achieved (with probability $R_2(h)$;
- either $e(t)=1, \alpha(t)=0, \beta(t)=1)$, $h \le \gamma(t) < x+h$ (with probability $F_{01}^{(0)}(x+h;t) - F_{01}^{(0)}(h;t)$) and the service in course in S_2 is not achieved and the server in S_2 don't fails(with probability $1 - H_2(h) - D_2(h)$).

So, we can write:

$$F_{01}^{(0)}(x; t+h) = F_{01}^{(1)}(h;t) D_1(x) + F_{11}^{(0)}(x;t) R_2(h) + [1 - H_2(h) - D_2(h)] \times [F_{01}^{(0)}(x+h;t) - F_{01}^{(0)}(h;t)].$$

This is a simple application of the well known formula of total probabilities.

Now, we can divide both sides of the above equation by h and take the limit when $h \to 0$. So, we obtain the first equation. The other equations can be obtained by the same physical arguments.

4.1 *Remark*

It seems that these equations may hold for arbitrary distributions and/or lifetime and renewal times.

5 RESOLUTION OF THE SYSTEM IN STATIONARY REGIME

Consider the Network in stationary regime, when the following limits exist

$$F_{ij}^k(x) = lim_{t\to\infty} F_{ij}^k(x;t),$$

$$p_k = lim_{t\to\infty} p_k(t),$$

$$q_k = lim_{t\to\infty} q_k(t).$$

Denote

$$\Phi_{ij}^k(s) = \int_0^\infty e^{-st} F_{ij}^k(x)dx,$$

$$a_{ij}^k = \frac{dF_{ij}^k}{dx}(0),$$

$$b_{ij}^k = F_{ij}^k(\infty).$$

Applying the Laplace transform to the system of section 4, we obtain

$$(s - \mu_2 - \theta_2)\Phi_{01}^{(0)}(s) + \nu_2 \Phi_{11}^{(0)}(s) = \frac{1}{s}[a_{01}^{(0)} - a_{01}^{(1)} d_1(s)],$$
$$(s - \nu_2)\Phi_{11}^{(0)}(s) + \theta_2 \Phi_{01}^{(0)}(s) = \frac{1}{s}[a_{11}^{(0)} - a_{11}^{(1)} d_1(s)],$$
$$(s - \theta_2 - \mu_2(1 - h_1(s)))\Phi_{11}^{(1)}(s) = \frac{1}{s}[a_{01}^{(1)} - a_{01}^{(0)} r_1(s) - \mu_2 b_{01}^{(0)} g_1(s) - \mu_1 b_{00}^{(1)} [g_1(s)]^{N-1}],$$
$$(s - \nu_2)\Phi_{11}^{(1)}(s) + \theta_2 \Phi_{01}^{(1)}(s) = \frac{1}{s}[a_{11}^{(1)} - a_{11}^{(0)} r_1(s) - \mu_1 b_{10}^{(1)} [g_1(s)]^{N-1}],$$
$$\Phi_{00}^{(1)}(s) = \frac{1}{s}[p_0 + p_1 r_1(s)[g_1(s)]^N],$$
$$\Phi_{00}^{(1)}(s) = \frac{1}{s}[q_0 + q_1 r_1(s)[g_1(s)]^N].$$

From the two first equations above we get

$$\Phi_{01}^{(0)}(s) = \frac{(a_{01}^{(0)} - a_{01}^{(1)})(s - \nu_2) - \nu_2(a_{11}^{(0)} - a_{11}^{(1)} d_1(s))}{s[(s - \mu_2 - \theta_2)(s - \nu_2) - \nu_2\theta_2]},$$
$$\Phi_{11}^{(0)}(s) = \frac{(a_{01}^{(0)} - a_{01}^{(1)})\theta_2 - (s - \mu_2 - \theta_2)(a_{11}^{(0)} - a_{11}^{(1)} d_1(s))}{s[(s - \mu_2 - \theta_2)(s - \nu_2) - \nu_2\theta_2]}.$$

Next, the third equation give

$$\Phi_{01}^{(1)}(s) = \frac{a_{01}^{(1)} - a_{01}^{(0)} r_1(s) - \mu_2 b_{01}^{(0)} g_1(s) - \mu_1 b_{00}^{(1)} (g_1(s))^{N-1}}{s(s - \theta_2 - \mu_2(1 - g_1(s)))},$$
$$\Phi_{11}^{(1)}(s) = \frac{1}{s - \nu_2}[\frac{1}{s}(a_{11}^{(1)} - a_{11}^{(0)} r_1(s) - \mu_1 b_{10}^{(1)} (g_1(s))^{N-1}) - \theta_2 \Phi_{01}^{(1)}(s)].$$

In these formula, the constants $a_{ij}^{(k)}$ are still unknown. Note first that the equation

$$(s-\mu_2-\theta_2)(s-\nu_2)-\nu_2\theta_2=0$$

has a positive root

$$s_0=\frac{1}{2}[\mu_2+\nu_2+\theta_2 \\ +\sqrt{(\mu_2-\nu_2)^2+\theta_2(\theta_2+2(\nu_2+\mu_2))}].$$

Since the functions $\Phi_{01}^{(0)}(s)$ and $\Phi_{11}^{(0)}(s)$ are analytical functions in the subplane $\Re_e(s)>0$, so when the numerator equal zero, the denominator must also be zero.

So, the unknown constants a_{ij}^{k} are solutions of the following system of equations

$$(a_{01}^{(0)}-a_{01}^{(1})(s_0-\nu_2)-\nu_2(a_{11}^{(0)}-a_{11}^{(1)})d_1(s_0)=0$$

$$(a_{01}^{(0)}-a_{01}^{(1)})\theta_2-(s_0-\mu_2-\theta_2)(a_{11}^{(0)}-a_{11}^{(1)}d_1(s_0))=0.$$

Similarly, since for $s=\nu_2$, the denominator of the function $\Phi_{11}^{(1)}(s)$ equal zero, we have

$$\{[\nu_2-\theta_2-\mu_2(1-g_1(\nu_2))](a_{11}^{(1)}-a_{11}^{(0)}r_1(\nu_2) \\ -\mu_1b_{10}^{(1)}(g_1(\nu_2))^{N})-(a_{01}^{(1)}-a_{01}^{(0)}r_1(\nu_2)- \\ -\mu_2b_{01}^{(0)}g_1(\nu_2)-\mu_1b_{00}^{(1)}(g_1(\nu_2))^{N-1})\}=0$$

Next, consider the function

$$f(s)=s-\theta_2-\mu_2(1-g_1(\nu_2)).$$

It is not difficult to see that for $s=0$, we have $f(0)=-\nu_2$ and $lim_{s\to\infty}f(s)=+\infty$. So, the function $f(s)$ has at least one root in the domain $\Re_e(s)>0$, s_1 say.

Consequently, we have a system of four equations

$$a_{01}^{(1)}-a_{01}^{(0)}r_1(s_1)-\mu_2b_{01}^{(0)}g_1(s_1) \\ -\mu_1b_{00}^{(1)}(g_1(s_1))^{N-1}=0, \\ a_{01}^{(1)}-a_{01}^{(0)}r_1(s_1)-\mu_2b_{01}^{(0)}g_1(s_1) \\ -\mu_1b_{00}^{(1)}(g_1(s_1))^{N-1}=0.$$

Consider now the linear system of algebraic equation which can be written under the form $Ax=b$, where

$$x=(a_{01}^{(0)},a_{11}^{(0)},a_{01}^{(1)},a_{11}^{(1)}) \\ b=(0,0,b_1,b_2) \\ b_1=\mu_1b_{00}^{(1)}(g_1(s_1))^{N-1}+\mu_2b_{01}^{(0)}g_1(s_1) \\ b_2=\mu_1(g_1(s_1))^{N-1}\{b_{10}^{(1)}(\nu_2-\theta_2-\mu_2(1-g_1(\nu_2)) \\ -b_{00}^{1}\}-\mu_2b_{01}^{(0)}g_1(\nu_2)\}$$

and the matrix A has the following form

$$\begin{pmatrix} s_0-\nu_2 & -s_0+\nu_2 & -\nu_2 & \nu_2d_1(s_0) \\ \theta_2 & -\theta_2 & X(\theta_2,\nu_2,\mu_2) & W(\theta_2,\nu_2,\mu_2) \\ r_1(\nu_2) & -1 & U(\theta_2,\nu_2,\mu_2) & V(\theta_2,\nu_2,\mu_2) \\ -r_1(s_1) & 1 & 0 & 0 \end{pmatrix}$$

where

$$X(\theta_2,\nu_2,\mu_2)=\mu_2+\theta_2-s_0$$

$$U(\theta_2,\nu_2,\mu_2)=\theta_2+\mu_2(1-g_1(\nu_2))-\nu_2$$

$$V(\theta_2,\nu_2,\mu_2)=\nu_2-\theta_2)-\mu_2(1-g_1(\nu_2))$$

$$W(\theta_2,\nu_2,\mu_2)=d_1(s_0)(s_0-\mu_2-\theta_2)$$

6 VIRTUAL WAITING TIME

Now, we are able to derive the distribution of the virtual waiting time of an arbitrary request in the station S_1 .

Denote by $\omega(t)$ the virtual waiting time of such a request i.e. waiting time of a request which will be arrived at time t . It's the period between the time t until the departures of all requests arrived before t . If the server is available andfree of requests, then $\omega(t)=0$.

Also denote by $F(x)=lim_{t\to\infty}P\{\omega(t)<x\}$ the limiting probability distribution of the virtual waiting time and

$$\Phi(s)=lim_{t\to\infty}\int_0^\infty e^{-sx}dP\{\omega(t)<x\}$$

is Laplace-Stieltjes transform.

The structure of such a stochastic process is the following. Let $t_1,t_2,t_3....$ the instants of requests of "pure" service (regular one's) and/or "impure service" (renewal of components failures, virus elimination etc.) ($t_1<t_2<t_3<.....$. Then for $t_n<t<t_{n+1}$ the process $\{\omega(t),t\geq 0\}$ can be defined as

$$\begin{cases} \omega(t) & = & 0, \quad \text{if} \quad \omega(t_n)\leq t-t_n,] \\ \omega(t) & = & \omega(t_n)-(t-t_n),\text{if}\,\omega(t_n)\geq t-t_n \end{cases}$$

For $t=t_n$, we have $\omega(t_n+0)=\omega(t_n-0+)+\eta_n$, where η_n is the service time of a regular customer and/or the renewal period of an interruption (due to a physical breakdown or a computer attack) which had occurred at time t_n . Moreover, we assume the initial condition $\omega(0)=0$.

The process $\{\omega(t),t\geq 0\}$ has stepwise linearly decreasing paths as shown in Figure 2.

We have in the previous section derived the joint distribution $\{F_{ij}^{(k)}(x)\}$ of the server state in S_1, S_2 and the variable $\omega(t)$ in stationary regime.

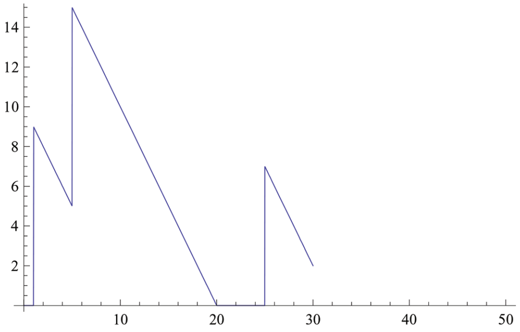


Figure 2. Sample path of the stochastic process $\{\omega(t), t \geq 0\}$.

Now, the Laplace-Stieltjes transform $\Phi(s)$ of the virtual waiting time canbe expressed through these functions as follow:

$$\Phi(s) = s\sum_{i=0}^{1}\sum_{j=0}^{1}\Phi_{ij}^{(1)}(s) + lim_{s\to 0}\Phi_{01}^{(0)}(s) + \Phi_{11}^{(0)}(s).$$

Now, after tedious algebra we found the following explicit expression for the above defined function

$$\begin{aligned}
\Phi(s) = {} & \frac{1}{\mu_2}[a_{01}^{(0)} + a_{11}^{(0)} - a_{01}^{(1)} - a_{11}^{(1)} \\
& + \frac{1}{\nu_2}[\theta_2(a_{01}^{(0)} - a_{01}^{(1)}) - (a_{11}^{(0)} - a_{11}^{(1)})(\mu_2 + \theta_2)]] \\
& + \frac{1}{s-\nu_2}(a_{11}^{(1)} - a_{11}^{(0)}r_1(s) - \mu_1 b_{10}^{(1)}[g_1(s)]^{N-1}) \\
& + [g_1(s)]^N[p_0 + q_0 + (p_1 + q_1)r_1(s)] + \qquad (1) \\
& + \frac{1}{s-\nu_2}(a_{11}^{(1)} - a_{11}^{(0)}r_1(s) - \mu_1 b_{10}^{(1)}[g_1(s)]^{N-1}) \\
& + [g_1(s)]^N[p_0 + q_0 + (p_1 + q_1)r_1(s)] + \\
& + \left(1 - \frac{\theta_2}{s-\nu_2}\right)\left(\frac{1}{s-\theta_2-\mu_2(1-g_1(s)}\right) \\
& \times[a_{01}^{(1)} - a_{01}^{(0)}r_1(s) - \mu_2 b_{01}^{(0)}(s)g_1(s) \\
& - \mu_1 b_{00}^{(1)}[g_1(s)]^{N-1}] \qquad (2)
\end{aligned}$$

We need in the above computations to take into account the normalization condition.

7 APPLICATIONS AND NUMERICAL ILLUSTRATIONS

In this section, we give an application of the above results with some numerical illustrations.

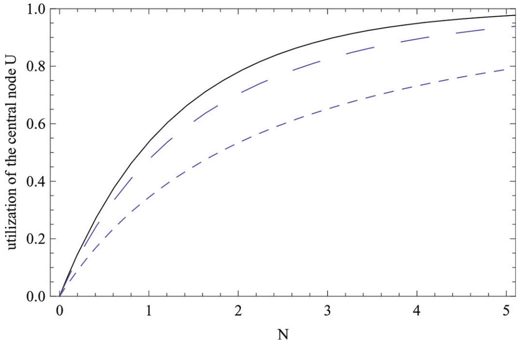


Figure 3. Utilization of the central node.

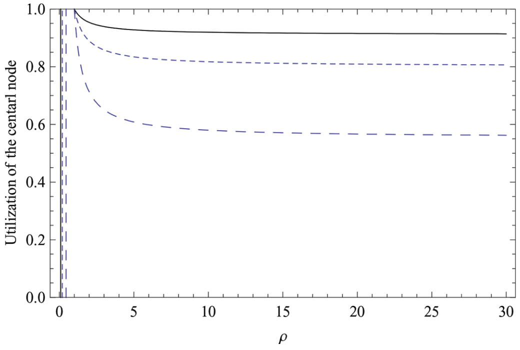


Figure 4. Effect of ϱ on the utilization U .

An interesting performance metric is the utilization of the central node (base station) S_1 against the rest of the network $U = 1 - F(0+) = \Phi(+\infty)$. This metric is plotted as a function of N in Figure 3 for the following cases:

1. $\theta_1 = 0.5$: short dashed line;
2. $\theta_1 = 2$: long dashed line;
3. $\theta_1 = 10$: gray level line;

where θ_1 is the breakdown rate in the central node S_1 . For this experiment, we set $\mu_1 = 1$, $\mu_2 = 2$, $\theta_2 = 1$, $\nu_1 = 1$, $\nu_2 = 2$.

We see how the utilization of the central node increases while N increases and θ_2 decreases.

Denote by m_i the total sojourn time of a customer in the node S_i (i = 1,2). Figure 3 shows the effect of the ratio $\rho = \frac{m_1}{m_2}$ on the utilization U for different values of N :

1. $N = 10$: short dashed line;
2. $N = 5$: long dashed line;
3. $N = 15$: gray level line;

For this experience we take the same numerical values while the breakdown rate in the Central node is fixed as $\theta_1 = 0.5$. Here, we see on Figure 4 that the increasing of ρ decreases the utilization when N decreases.

8 CONCLUSION

In this work, we have provided a method for finding some performance metrics such as the utilization of the central node (base station) in some modern networks such as WSNs or database systems.

ACKNOWLEDGMENT

The authors would like to thanks the anonymous referees for pertinent and helpful comments that allow to improve the quality of the paper.

REFERENCES

Aissani, A. & J. Artalejo (1998). On the single server retrial queue subject to breakdowns. *Queueing systems: theory and applications 30*, 309–321.

Boucherie, R. & N. Dijk (2010). *Queueing networks: A fundamental approach.* Berlin: Springer.

Demirkol, I., C. Ersoy, F. Alagz, & H. Deli (2009). The impact of a realistic packet traffic model on the performance of surveillance wireless sensor networks. *Computer networks 53*, 382–399.

Gaver, D. (1962). A waiting line with interrupted services, including priorities. *J. Roy. Stat. Soc. 69B24(1)*, 73–90.

Gnedenko, B. & I. Kovalenko (1989). *Introduction to queueing theory.* London: Birkhauser.

Medvediev, G. (1978). Closed queueing networks and their optimization. *Cybernetics 6*, 65–73.

Osman, R. & W. Knottenbelt (2012). Database system performance evaluation models: A survey. *Performance Evaluation 69*, 471–493.

Phung-Duc, T. (2012). An explicit solution for a tandem queue with retrials and losses. *Opsearch 12(2)*, 189–207.

Qiu, T., L. Feng, F. Xia, G. Wu, & Y. Zhou (2011). A packet buffer evaluation method exploiting queueing theory for wireless sensor networks. *ComSis 8 (4)*, 1027–1049.

Qiu, T., F. Xia, L. Feng, G. Wu, & B. Jin (2011). Queueing theory-based path delay analysis of wireless sensor networks. *Adv. in Electr. and Comput. Engng. 11(2)*, 3–8.

Senouci, M., A. Mellouk, & A. Aissani (2012). Performance evaluation of network lifetime spatial-temporal distribution for wsn routing protocols. *J. Network and Computer Applications 35(4)*, 1317–1328.

Senouci, M., A. Mellouk, & A. Aissani (2014). Random deployment of wireless sensor networks: A survey and approach. *Internat. J. Ad Hoc and Ubiquitous Computing 15(1–3)*, 133–146.

Senouci, M., A. Mellouk, L. Oukhellou, & A. Aissani (2015). Wsns deployment framework based on the theory of belief functions. *Computer Networks 88(6)*, 12–26.

Taleb., S. & A. Aissani (2010). Unreliable m/g/1 retrial queue: monotonicity and comparability. *Queueing systems: theory and applications 64*, 227–252.

This page intentionally left blank

Risk and hazard analysis

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Risk assessment of biogas plants

K. Derychova & A. Bernatik
Faculty of Safety Engineering, VSB—Technical University of Ostrava, Ostrava-Vyskovice, Czech Republic

ABSTRACT: Biogas represents an alternative source of energy with versatility of utilization. This article summarizes information about production and properties of biogas, storage possibilities and utilization of biogas. For the assessment of the risks of biogas were established major-accident scenarios. These scenarios were complemented by an analysis of sub-expressions of biogas namely fire and explosion, because biogas is formed mainly from methane, which is highly flammable and explosive gas. For analysis of methane were used Fire & Explosion Index and modelling program ALOHA.

1 INTRODUCTION

Due to limited capacity of fossil fuels, it is switching to alternative sources of energy. In these sources can be included biogas, which belongs to the gaseous renewable fuels. In Europe is currently over 14 thousands biogas plants, most of them are in Germany, Italy, Switzerland and France. Czech Republic is in fifth places in the number of biogas plants, as can be seen in Figure 1 (EBA 2015). In the Czech Republic were about 554 biogas plants at the end of 2015.

Biogas is nothing new, its history dates back to the late 19th century. However, its production in the past and nowadays is significantly different. Therefore, it can be alleged that anaerobic fermentation is a new developing and perspective technology. Biogas is basically a mixture of gases, among major components belong methane and carbon dioxide and minor components is formed by: hydrogen sulfide, water, hydrogen, nitrogen, ammonia, oxygen and optionally other substances. Representation of individual components in the mixture and its amount varies depending on the raw materials and technological process (Derychova 2015).

2 BIOGAS—CLEAN ENERGY

Biogas is a widespread term for gas produced by anaerobic fermentation. The anaerobic fermentation takes place in the natural environment (e.g. in

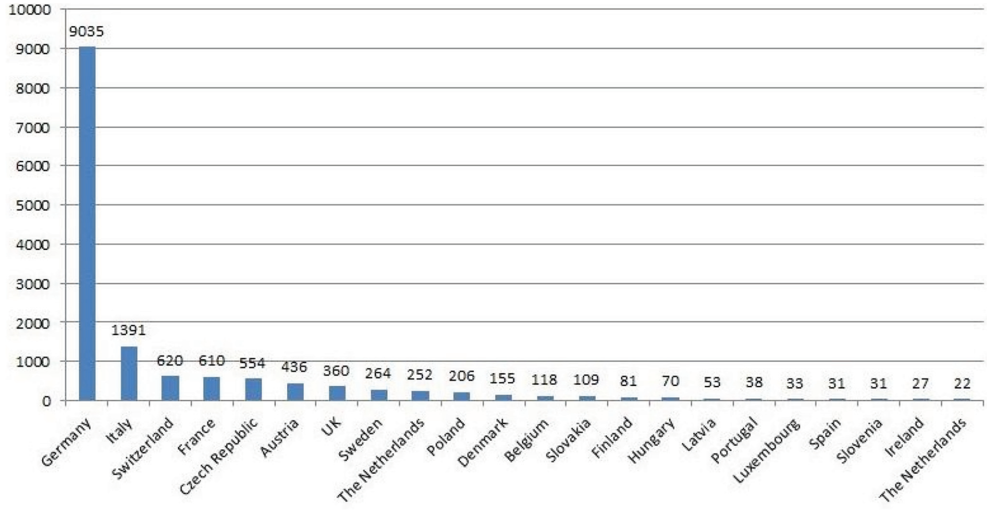


Figure 1. Overview of the number of biogas plants in Europe, according to the EBA (EBA 2015).

wetlands, in the digestive tract of ruminants), in an agricultural environment (rice field, the dunghill), in waste management, further on the landfill sites (landfill gas) at sewage treatment plants and biogas plants (Straka 2010). In the European legislation (2009/28/EC) is biogas formulated as a *fuel gas produced from biomass and/or from the biodegradable fraction of waste, that can be purified to natural gas quality, to be used as biofuel, or wood gas*. And according to (Kara 2007) can be concept of biogas used for all kinds of gas mixtures produced by the operation of micro-organisms, eg. below ground, in the digestive tract of animals, in landfill sites, lagoons or the controlled anaerobic reactors. However, in technical practice, is the biogas presented as a gaseous mixture produced in the anaerobic fermentation of wet organic substances in technical equipment (reactors, digesters, etc.).

Biogas production is a very complicated biochemical process, in which the mixed cultures of organisms decompose organic material in absence of air. Anaerobic fermentation takes place in four consecutive phases—hydrolysis, acidogenesis, acetogenesis and methanogenesis, when the last phase produce methane and carbon dioxide (Juchelkova et al. 2010, Rutz et al. 2012, Schulz et al. 2001). Process of produce biogas takes place at a particular operating temperature (according to the type of bacteria—psychrophilic, mesophilic, thermophilic) at pH from 6.5 to 7.5 and for specific time (according to the type of bacteria from 10 to 120 days). (Derychova 2014; Rutz et al. 2008). The outcome of this process is biogas and a digestate, which is a good quality fertilizer (Schulz et al. 2001). Suitable materials for production of biogas are substances of biological origin, such as plant biomass, animal biomass, organic by-products and organic waste (Kara 2007).

Composition of biogas is variable and dependent on use raw materials supplied to the process; it is confirmed by the authors (Rasi et al. 2007). Ideally biogas contains only two major gases, methane and carbon dioxide. However the raw biogas includes other minor gases, e.g. hydrogen sulfide, nitrogen, oxygen, water vapor, hydrogen, ammonia, and other siloxanes (Jönsson et al. 2003, Kara 2007).

Comparison of the chemical composition of various biogases shows the following Table 1. Proportional representation of the two main components of biogas (methane, carbon dioxide), but also minor components, differ depending on the origin of biogas and on the composition of the starting substrate.

The concentration of methane in the biogas is not permanent, may change the density of the entire gas mixture. If the methane concentration falls below 60%, it becomes biogas heavier than air and may accumulate in depressions at landfills and in reactor vessels. The presence of minor components of biogas can indicate the presence of some chemical elements in the material or malfunction during the fermentation (Kara 2007, Straka 2010).

Storage tanks are built for accumulation of biogas to reduce disparities between production and consumption. The daily cycle of gas consumption can be independently varied. Biogas can be stored for long periods of time and then can be used without the loss. Gas tanks can be divided according to the materials, function and arrangement. Publications (Schulz et al. 2001) divide biogas tanks into tanks designed as a low pressure, medium pressure and high pressure reservoirs. The characteristics of these reservoirs are shown in Table 2. According to the time of storage distributes publications (Krich et al. 2005) gas tanks

Table 1. Approximate composition of the biogas (Jönsson et al. 2003).

Component	Chemical formula	Agricultural biogas plant	Waste water treatment plant	Landfill plant
Methane [vol.%]	CH_4	60–70	55–65	45–55
Carbon dioxide [vol.%]	CO_2	30–40	balance	30–40
Nitrogen [vol.%]	N_2	< 1	< 1	5–15
Hydrogen sulfide [ppm]	H_2S	10–2,000	10–40	50–300

Table 2. Design of biogas tanks (Schulz et al. 2001).

Pressure level	Operation pressure	Bulk	Storage facilities
Low pressure	20–50 mbar	50–200 m^3	gas tank with water seal
	0.05–0.5 mbar	10–2,000 m^3	gas tank with foil cover
Medium pressure	5–20 bar	1–100 m^3	steel storage tank
High pressure	200–300 bar	0.1–0.5 m^3	steel cylinder

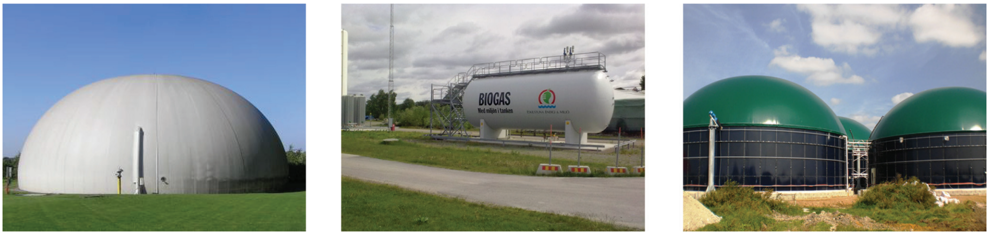


Figure 2. Various types of biogas tanks.

Table 3. Representation of the gases in biogas and biomethane (Derychova 2014, Peterson et al. 2009).

Component	Raw biogas	Upgraded biogas
Methane	40–75 vol.%	95–99 vol.%
Carbon dioxide	25–55 vol.%	≤ 5 vol.%
Water vapor	0–10 vol.%	–
Nitrogen	0–5 vol.%	≤ 2 vol.%
Oxygen	0–2 vol.%	≤ 0.5 vol.%
Hydrogen	0–1 vol.%	≤ 0.1 vol.%
Ammonia	0–1 vol.%	≤ 3 mg/m^3
Hydrogen sulfide	0–1 vol.%	≤ 5 mg/m^3

for short and long term storage. On Figures 2 are shown various types biogas tanks.

Utilization of biogas energy is versatile. Biogas can be used to produce heat, cooling, electricity; further can be used to cogeneration and trigeneration (not often used). For the use of biogas in the transport and distribution to the natural gas grid it is necessary to modify the biogas. Publications (Petersson et al. 2009) present a various types of biogas purification. Biogas plants utilize about 20–40% of produced heat to heat up the digesters (process heat) and other 60–80% of heat is called "waste heat", it is farther used for additional electricity production (Rutz et al. 2012).

The most often is biogas converted directly in cogeneration units into electricity or heat in a biogas plants. Surplus of electricity can be delivered into the electric power grids or distributed through a pipelines as a heat or gas, or can be transported down the road too.

2.1 *Properties of biogas*

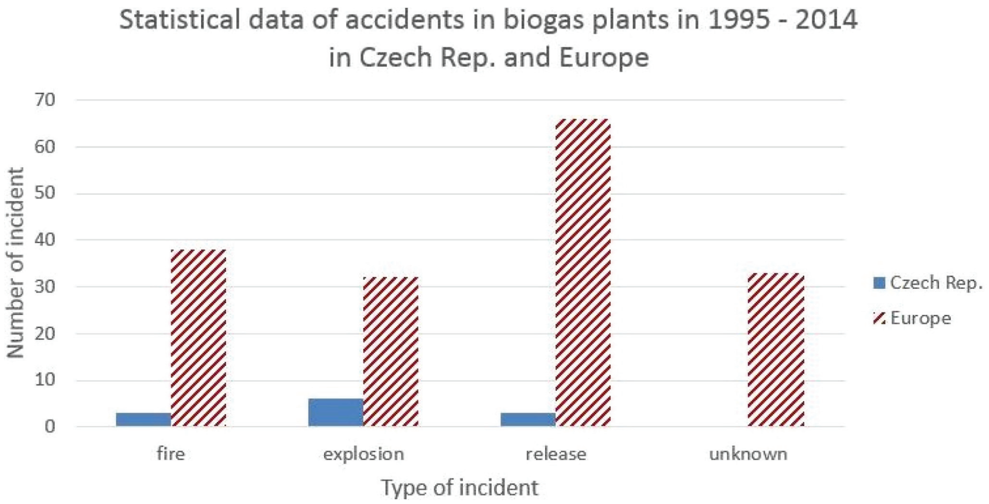
Biogas is composed of majority and minority gases, approximate composition of biogas shown the Table 3. Characteristics of biogas depend on the content of methane. Physical and chemical properties of biogas depend on the material and process parameters (Kara 2007, Straka 2010).

Is generally known that biogas is unbreathable gas. Density of biogas is approximately 1.2 kg/m^3. It is slightly lighter than air, this means that biogas will goes up rapidly and mixed with air (Derychova 2015). Biogas is a flammable gas which is also explosive, under certain conditions. Conditions for an explosion, which must be complied with, are the concentrations of explosive gas between the upper and lower explosive limit (biogas consisting of 65% methane and 35% carbon dioxide has explosive limit 6–12 vol.%), the presence of an explosive mixture in an enclosed space and reaching the ignition temperature (for biogas it is 620–700°C). Under the lower explosion limit does not occur to ignite of the biogas and above the upper explosion limit can biogas only burn a flame. (Schulz et al. 2001) Critical pressure of biogas is in the range of 7.5–8.9 MPa and critical temperature is – 82.5°C. Biogas has a very slow diffusion combustion, the maximum progress speed of the flame in air is 0.25 m/s, because of CO_2 (Rutz et al. 2012).

Biogas has an ability to separate itself into its compounds (thermodiffusion). Therefore, it is appropriate and necessary to know the properties of the individual components of biogas, these gases have its characteristic physical-chemical properties. E.g. carbon dioxide is heavier than air (1.53 kg/m^3), that's why it decreases and adheres to the ground. Methane, which is lighter than air (0.55 kg/m^3) rises into the atmosphere (Delsinne et al. 2010, Rutz et al. 2012).

3 HAZARDOUS MANIFESTATION OF BIOGAS

The increasing number of biogas plants increases the risk that there will be incidents at some of these stations. That prove the events that have occurred in recent times In January 2013 occurred a massive explosion with detonation inside the biogas plant in Chotetov, the station was still under the construction and it was not put into operation. In November 2013 in Chric was found employee in



Graph 1. Statistic of incidents in biogas stations in the Czech Republic and Europe.

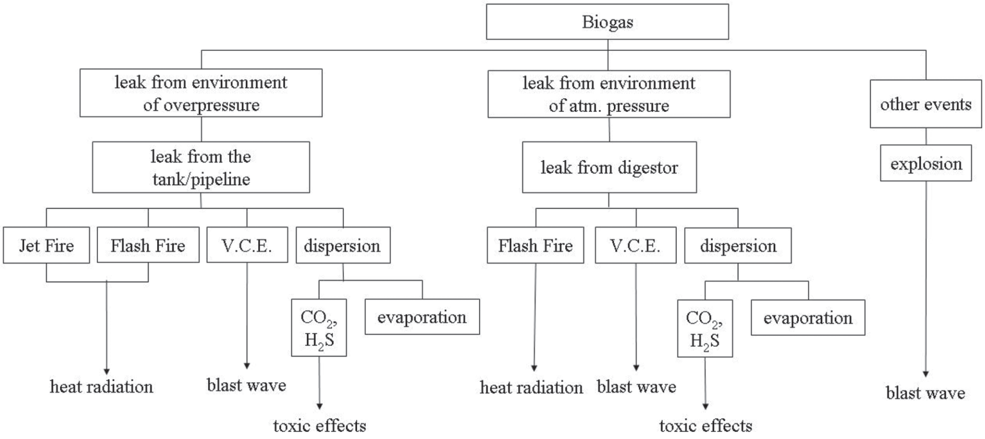


Figure 3. Scheme of scenarios for accidents caused by biogas (Derychova, 2015).

the shaft, most likely was intoxicated by methane. Even abroad, there are accidents which have its casualties. In Germany in 2009 exploded biogas plants, one worker was killed and two others were injured (Derychova 2015). The Graph 1 shows a comparison the number of events in biogas stations in the Czech Republic and in Europe since 1995 till 2013 (Casson et al. 2015, Ministery of the Interior 2014). From the graph can be read events that occurred on biogas stations in a given period, incidents like leaks, fire, explosion and other events (of unknown cause).

The possible scenarios for accidents caused by biogas are identified in the following diagram (Figure 3). (Derychova 2015).

Possible consequences of the accidents caused by biogas are the heat radiation in case of fire, blast (shock) wave with any flying fragments in case of explosion and toxic effects of gases into the atmosphere scattering (Derychova 2015).

4 RISK ANALYSIS OF BIOGAS PLANTS

Safety of biogas plants has to be focused on the most commonly occurring risks which are an explosion (fire), leakage (poisoning, suffocation) and environmental pollution. As was mentioned, composition of biogas depends on the type of biogas

Table 4. Characteristics of methane.

Parameters	Value
CAS	74–82–8
Molecular weight	16 g/mol
Ignition temperature	595°C
Flash point	–188°C
Minimum ignition energy	0.29 mJ
Temperature class	T1
Lower explosion limit	4.4 vol.%
Upper explosion limit	17 vol.%
Explosion group	II A
Classification (1272/2008/ES)	H 280, H 220

plants, on the technological process of production and on input feedstock (Derychova 2014).

Considerable material damage and impact on the lives and health of people have fire and explosion hazards that are related to the production, usage, storage and transport of biogas. Biogas is a mixture of flammable gases, primarily by methane. The risk of fire and explosion is particularly high close to the digesters and gas reservoirs. Methane is extremely flammable and non-toxic gas, which is lighter than air. Its explosion limits are 5–15 vol.% and autoignition temperature is 595 °C. Mixture of methane and air can explode, to ignite of mixture can be set by electric spark or open flame. Methane in high concentration effect on humans in short duration, can lead to asphyxia due to lack of oxygen (Derychova 2015). Fire and explosion characteristics of methane are summarized in Table 4.

4.1 *Fire and Explosion Index (FEI)*

It is an index method for assessing the risk of fire and explosion. This method is a tool for revealing the locations with the greatest potential losses and enables us to predict the extent of damage to equipment. (Dow's F&EI, 1994)

Index was determined for bioreactor, where is biogas produced by microorganisms from varied feedstock. Biogas is formed mainly from methane, that's why is methane used for calculation of FEI. Material factor for methane is given in table in the annex of manual and is determined to 21. Other factors were chosen with considering to physical—chemical parameters of methane and production. Calculated value for FEI was 57.33. According to the table in manual is bioreactor is included to a light degree of hazard with range from 1 to 60. Further calculation for bioreactor was the radius of the affected area. The radius of the affected area R was calculated to 14.68 m. And to this radius has to be added the radius of the considered bioreactor (r = 6.75 m), it leads to actual radius of the affected area, which is 1,442.76 m^2.

The second index was determined for biogas tank. In biogas tank is stored biogas, which mainly substance id methane. So for calculation of FEI was used methane, it is same as previous. Material factor for methane is 21 as in previous case, just value of other factors were different because of various condition during the storage. FEI calculated for biogas tank was 25.2, and is lower than FEI for bioreactor, belongs to range of light degree of hazard too. Calculated radius of the affected area R was 6.45 m. And to this radius was added the radius of the considered biogas tank (r = 7 m), it leads to actual radius of the affected area, which is 568.32 m^2.

4.2 *ALOHA*

ALOHA (Areal Locations of Hazardous Atmospheres) is a software model program for the Windows operating system. It was developed in the eighties of the 20th century by American Organization for Conservation of Nature (US Environmental Protection Agency). This program allows simulation of leakage dangerous substances and subsequent modelling of the effects in graphical form. Simulation involves the following steps when the first is to enter the site of potential accidents (leakage), than select the type of hazardous substances and on the basis of geographical and climatic conditions may be modelled the way and type of release to the environment. Outcome of the program are radiuses of threatening zones in graphic form. (Aloha, 2007)

4.2.1 *Scenario 1*

The first modelled situation was methane leakage from the stationary source (gas tank) by short pipe with diameter 15 cm and pressure in a tank is 2 atm. The parameters of tank and atmospheric conditions are in Table 5. Graphical output for leaking tank, when methane is burning as a jet fire, can be seen in Figure. 4. Potential hazard of this burning methane leakage can be thermal radiation from jet fire or downwind toxic effects of fire by-products.

Burn duration is 8 minutes, maximum flame length is 13 meters, max burning rate is 274 kg/min and total burned amount is 891 kg, which is 47% of total amount of methane in tank.

Out of the resulting graph can be seen that the heat radiation is able to intervene into near area by surface heat radiation about 35 m^2.

The lethal zone is red doted area, the amount of heat radiation is there 10 kW/m^2, and extends to a distance of 15 m from the source. The zone of the 2nd degree burns is indicated by orange thickly

Table 5. Parameters of the first modelled situation.

Bioreactor	
Diameter	14 m
Length	9.78 m
Volume	1,500 m^3
Mass	1,995 kg
Atmospheric Data	
Wind	2.1 m/s, SW
Cloud cover	partly cloudy
Ground roughness	urban
Stability class	E
Air temperature	22°C
Relative humidity	50%

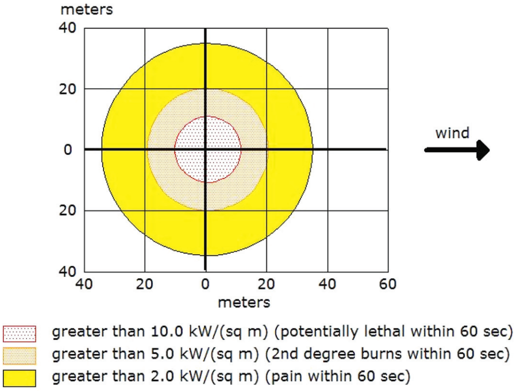


Figure 4. Thermal radiation from jet fire.

dotted area with heat flow 5 kW/m^2, extends to a distance of 22 m. The Pain zone with distance of 35 m from the source is colored by yellow, the intensity of the heat flow there is 2 kW/m^2.

4.2.2 *Scenario 2*

The second modelled situation was methane leakage from the stationary source (bioreactor) by short pipe with diameter 8 inches. The parameters of bioreactor and atmospheric conditions are in Table 6. Graphical output for leaking tank, when methane is not burning as it escapes into the atmosphere can be seen in Figure 5. Potential hazard of this methane leakage it can be downwind toxic effects, vapor cloud flash fire or overpressure from vapor cloud explosion.

Graph shows two regions where is presence of flammable vapor cloud in different concentration. In red dotted area is concentration of methane about 30,000 ppm. In yellow area is concentration of methane at 5,000 ppm, explosion range of mixture methane and carbon dioxide in air.

Table 6. Parameters of the second modelled situation.

Bioreactor	
Diameter	13.5 m
Length	13.9 m
Volume	2,000 m^3
Mass	1,616 kg
Atmospheric Data	
Wind	0.46 m/s, SW
Cloud cover	partly cloudy
Ground roughness	urban
Stability class	F
Air temperature	18°C
Relative humidity	50%

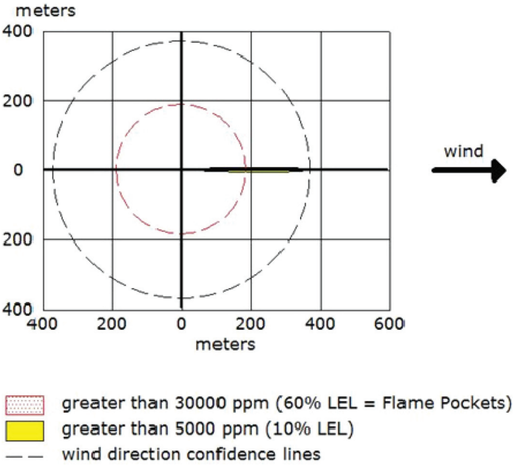


Figure 5. Flammable threat zone.

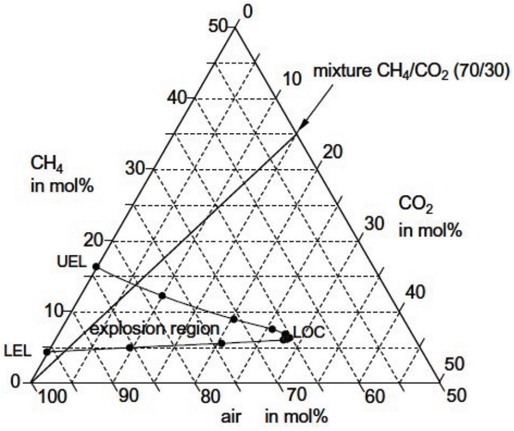


Figure 6. Range of explosion of mixture methane and carbon dioxide in air (Schroeder et al. 2014).

The danger of fire and explosion biogas (flammable methane) examined the authors in the article (Schroeder et al. 2014).The authors point out the necessity to know the explosion limits of gases and gas mixtures mixed with air in order to prevent an explosion when handling the biogas. They concluded that using Le Chatelier's equation for calculating the explosive limits of mixtures, the result especially in the upper explosive limit was wrong. Therefore, has been formed explosion diagrams helped with measured data of explosion limits of biogas to determine exactly the explosion limits of biogas. The explosion diagram is shown in Figure 6.

5 CONCLUSION

Biogas production is a cheap energy from residues and wastes. The benefit of biogas is to reduce the burden to the environment. The advantage of biogas is its versatile utilization depending on its purification. However, biogas plants and biogas production represent a certain hazard. Biogas is a dangerous flammable and toxic gas. Its flammability is due to the presence of methane, toxicity is caused by content of hydrogen sulfide or of carbon dioxide. The source of environmental risk is feedstock placed in reactors or storage feedstock for anaerobic fermentation (liquid manure, slurry).

Evaluation of flammability and explosiveness of biogas plants was carried out by FEI by two chosen scenarios—for production of biogas in the bioreactor and for storage biogas in a tank. Final indexes were quite small, and according to the manual (Dow's F&EI, 1994) were bioreactor and a gas tank included to a light degree of hazard. The radius of the affected area has a value of 1,442.76 m^2 in case of bioreactor, and radius for the gas tank has a value of 568.32 m^2.

The analysis was complemented by modelling program Aloha. In the program was modelling of two scenarios of methane leakage. The first scenario considered pipeline leaking from tank, when methane was burning as a jet fire. Aloha considers for this leakage affected area up to 35 m^2, where there has been lethal zone, zone of 2nd degrees burns and zone of injuries. In case of the second scenario occurred leakage of methane from the bioreactor by the leak pipe. Leaking methane wasn't ignited and therefore it threatened the neighborhood by various concentrations of this gas up to area of 370 m^2 from the bioreactor. These scenarios reveal possible leaks, which may occur in the biogas plants.

With an increasing number of biogas plants is increasing the risk of accidents at these facilities. The impact of incidents will be more of a local character. From the examples above may occur when operating biogas plants to casualties. For this reason, it is necessary to pay special attention to the safety of these facilities and by the appropriate measures (technical and organizational) reduce to rate of risks to avoid damage to life, environment and property.

ACKNOWLEDGEMENT

The article was prepared in the framework of the project SGS SP2016/115.

The article has been done in connection with project Institute of clean technologies for mining and utilization of raw materials for energy use—Sustainability program. Identification code: LO1406 – project is supported by National Program for Sustainability I (2013–2020) financed by the means of state budget of the Czech Republic.

REFERENCES

Aloha: User's Manual. 2007. Washington: US EPA, 195 p. <http://www.epa.gov/osweroe1/docs/cameo/ALOHAManual.pdf>.

Casson Moreno, V., Papasidero, S., Scarponi, G.M., Guglielmi, D. & Cozzani, V. 2015. Analysis of accidents in biogas production and upgrading. *Renewable Energy*, vol.

Delsinne, S. et al. 2010. Biogas Safety and Regulation: Workshop. Paris, 38 p.

Derychova, K. 2015. Biogas and scenarios of its major accidents. *TRANSCOM 2015: 11th European Conference of Young Researches and Scientists. Žilina, Slovak Republic: University of Žilina*.

Derychova, K. 2014. Hazards associated with biogas *Safety, Reliability and Risks 2014: 11th International Conference of Young Researchers. Liberec: Technical University of Liberec*.

Directive 2009/28/EC of 23 April 2009 published in the Official Journal of the European Union, L 140/16 of 5 June 2009, p. 16–62.

Dow´s fire & explosion index hazard classification guide. 1994. Seventh Ed. New York: *American Institute of chemical engineers*.

EBA Biogas Report. 2015. *European Biogas Association*.

Jönsson, O., Polman, E., Jensen, J.K., Eklund, R., Schyl, H. & Ivarsson, S. 2003. Sustainable gas enters the European gas distribution system. *Danish Gas Technology Center*.

Juchelkova, D. & Raclavska, H. 2010. *Biogas*. 9 p.

Kara, J. & et al. 2007. *Production and use of biogas in agriculture.* Ed. 1. Praha: VÚZT. 120 p.

Krich, K. & et al. 2005. Biomethane from Dairy Waste: A Sourcebook for the Production and Use of Renewable Natural Gas in California. 282 p.

Ministery of the Interior—Fire Rescue Service of the Czech Republic, 2004–2014. Statistical Yearbook.

Petersson, A. & Wellinge, A. 2009. Biogas upgrading technologies: Developments and innovations. *IEA Bioenergy.*

Rasi, S., Veijanen, A. & Rintala, A J. 2007. Trace compounds of biogas from different biogas production plants. *Energy*.

Rutz, D., Al Seady, T., Prassl, H., Köttner, M., Finsterwalder, S.V. & Jassen, R. 2008. *Biogas handbook*. Denmark: University of Southrn Denmark Esbjerg.

Rutz, D., Ramanauskalte, R. & Janssen, R. 2012. Handbook on Sustainable Heat Use from Biogas Plants. *Renewable Energies*.

Schroeder, V., Schalau, B. & Molnarne, M. 2014. Explosion Protection in Biogas and Hybrid Power Plants. *Procedia Engineering*. Vol. 84. 259–272 p.

Schulz, H. & Eder, B. 2001. *Biogas—Praxis. Grundlagen—Planung—Anlagenbau—Beispiele.* 1. Ed. Ökobuch Staufen. 167 p.

Straka, F.& et al. 2010. *Biogas: [Guide For Teaching, Design and operation of biogas systems]*. 3. ed. Praha: GAS. 305 p.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Verification of the design for forced smoke and heat removal from a sports hall

P. Kučera
VSB—Technical University of Ostrava, Ostrava, Czech Republic

H. Dvorská
OSH FM, Frydek—Mistek, Czech Republic

ABSTRACT: Many civil facilities commonly include active fire safety systems that help to create favourable conditions in the event of a fire. One of these active fire safety systems is equipment that removes smoke and heat. The article therefore focuses on a variant solution for forced fire ventilation in a concrete sports hall and the use of mathematical modelling of a fire (Fire Dynamics Simulator) to verify the effectiveness of the designed forced fire ventilation system, including simulations of the logical consequences of the system under consideration.

1 INTRODCUTION

Covered sports halls, which are used not only for sports events, but also for other purposes (e.g. concerts, exhibitions), attract a large number of people. For the sake of any eventual emergency, it is necessary to propose technical and organizational measures so that the design of such facilities minimizes the risk of panic situations. In the event of a fire, fire ventilation helps to remove combustible gases (products), smoke and heat, thus prolonging the time interval to ensure the safe escape of persons.

Through mathematical modelling, this article aims to assess forced fire ventilation in sports halls. Regarding the design of a system for removing smoke and heat (hereinafter referred to as ZOKT), two options were chosen. The difference between these variants mainly consists in the number of fire fans, their performance, division of the fire zone of the sports hall into smoke sections, and the decision whether or not to use smoke barriers

2 DESCRIPTION OF THE OBJECT

For the model of the sports halls, based on mathematical modelling in order to verify two different methods of forced fire ventilation, we selected a characteristic sports facility whose parameters were chosen as a representative sample of facilities of a similar nature. These parameters were then used as input for the building's construction and the ZOKT design.

The multifunctional sports hall, which is used not only for sports, but also for cultural events, has two usable aboveground floors and one usable underground floor. The third aboveground floor serves as a technical area and contains HVAC units. The back of house areas of the hall is located on the underground floor. Entrances to the hall for the spectators, restrooms, refreshment seating area and souvenir shops are located on the first aboveground floor. VIP boxes and restaurants are located on the second aboveground floor.

For the purpose of verifying fire ventilation, we chose the main area of the sports hall, i.e. the playing surface with walkways and seating area for spectators, which constitutes a separate fire compartment (Figure 1).

The vertical clearance of the assessed hall space is 22.5 m. The dimensions of the ice surface are

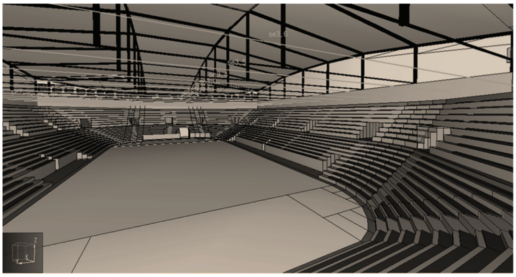


Figure 1. A view of the indoor sports hall facility (PyroSim program).

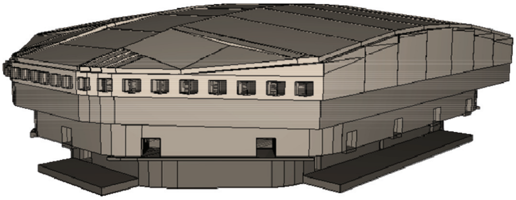


Figure 2. External layout of the sports hall (PyroSim program).

60 m × 30 m. The seating area reaches a height of 9.2 m above the ice surface. The maximum dimensions of the skin circumference above the seating area are 95 m × 65 m. The building is roofed with a steel framed structure. A rendering of the roof cladding and the hall's exterior is shown in Figure 2.

3 DESCRIPTION OF THE EQUIPMENT FOR FORCED SMOKE AND HEAT REMOVAL

According to the requirements for fire safety in construction, the sports hall must be fitted with equipment for removing smoke and heat. The structural system of the building is non-combustible; the structural components are of the DP1 type. The entire building is equipped with an electrical fire alarm system. Permanent fire extinguishing equipment is not considered.

The equipment for forced smoke and heat removal is designed with a forced gas outlet and a natural air intake.

Fire ventilation uses fire wall fans that correspond to operating temperatures of 200°C for 120 min with fire resistance class F_{200} 120.

The forced fire ventilation is designed as a forced self-acting ventilation system according to the requirements of (ČSN 73 0802, 2009), (ČSN 73 0831, 2011) in connection with (ČSN EN 12101–5, 2008).

3.1 *Option 1*

For the purpose of fire ventilation, the hall's fire zone areas are divided into four smoke sections. Smoke barriers (partitions) separating different smoke sections are building structures that meet the requirement of E 15 DP1, i.e. the criterion for properties of D_{600}.smoke barriers Wall fans for forced fire ventilation are installed on the third aboveground floor. Details of the location of the fans and smoke barriers are shown in Figure 3. The supply of fresh air is provided through inlets from the outside on the first underground floor. In the event of a fire, the air supply and wall fans are activated automatically by the electrical fire alarm system. For the needs of energy balance, we consider the possible simultaneous operation of two ventilated sections.

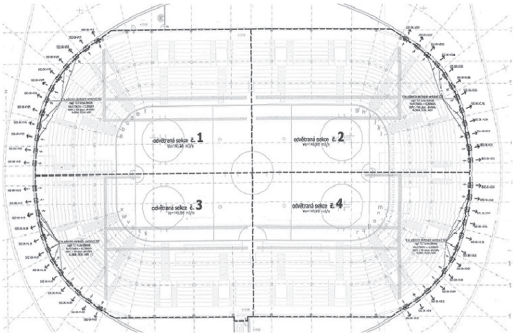


Figure 3. Floorplan drawing of the ZOKT design, option 1.

The occurrence of a fire is expected only in one smoke section; therefore, the calculation is performed for representing smoke section no. 1.

In each smoke section, the suction power will be provided by twelve wall fans with class F_{200} 120 (200°C/120 min.), $V_{o,l}$ = 11.75 $m^3 \cdot s^{-1}$ = 42.30 $m^3 \cdot hour^{-1}$ fire resistance. Δp = 200 Pa. The E 15 DP1 smoke wall has a D_{600} 30 fire resistance.

3.2 *Option 2*

For the purpose of smoke and heat removal, the hall areas consist of five smoke sections. The first four sections are designed identically to option number one. The fifth section is situated in the middle of the hall and forms a ring with a radius of about 13 m.

Assuming that the electrical fire alarm system activates the relevant group of ZOKT fans, the ZOKT design does not contain smoke barriers in the space under the roof structure.

Fire ventilation of sections nos. 1–4 is provided by fire wall fans installed on the third aboveground floor. Section No. 5 uses fire roof radial fans installed at the roof of the hall. Details of the location of the fans and smoke barriers are shown in Figure 4.

The supply of fresh air is provided through inlets from the outside on the first underground floor. In the event of a fire, the air supply and wall, as well as the roof fans, are activated automatically by the electrical fire alarm system. For the needs of energy balance, we consider the possible simultaneous operation of two ventilated sections.

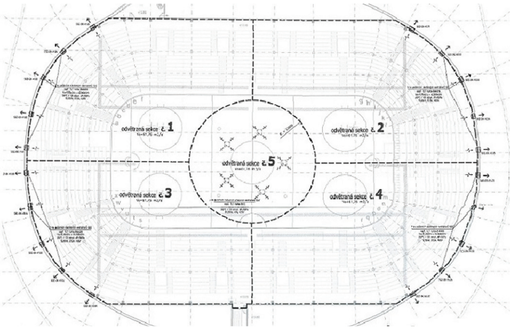


Figure 4. Floorplan drawing of the ZOKT design, option 2.

The occurrence of a fire is expected only in one smoke section; therefore, the calculation is performed for representing smoke section no. 1.

In smoke sections nos. 1–4, the suction power will be provided by five fire wall fans, each with fire resistance class F_{200} 120 (200°C/120 min.), $V_{o,1} = 12.35\ m^3 \cdot s^{-1} = 44.46\ m^3 \cdot hour^{-1}$. $\Delta p = 200$ Pa. The E 15 DP1 smoke wall has D_{600} 30 fire resistance.

In smoke section no. 5, the suction power will be provided by five roof radial fire fans with fire resistance class F_{200} 120 (200°C/120 min.), $V_{o,1} = 12.35\ m^3 \cdot s^{-1} = 44.46\ m^3 \cdot hour^{-1}$. $\Delta p = 200$ Pa.

4 BUILDING A MODEL OF MULTIPURPOSE SPORTS HALL FOR PERFORMING A MATHEMATICAL SIMULATION

For the mathematical simulation of the sports hall, we used the Fire Dynamics Simulator (FDS) program.

4.1 *Model geometry*

When compiling the geometry of the main hall space with a playing surface, seating area and walkways, constituting a separate fire zone, we made substantial modifications and simplifications regarding the choice of building materials. At the same time, we defined a large amount of materials that mutually differ in their chemical composition, thermo-physical properties and spatial arrangement (Kučera, 2010). Another input variable that has a major impact on the fire simulation is the method of modelling material pyrolysis.

4.2 *Definition of the design fire*

Because the sports hall should also serve for cultural and similar purposes, the fire was designed for the worst case scenario when the playing surface may contain stalls with a various assortment of goods, 60 kg·m^{-2} (ČSN 73 0802, 2009), Annex A, Tab. A.1, paragraph 6.2.1b)).

In order to verify the fire ventilation and represent the amount of combustible gases, the fire was simplified and simulated using the model of free fire development, presented in the methodological guideline for preparing fire-fighting documentation (Hanuška, 1996) and the subsequent action taken by fire brigades. The main time-dependent fire design parameters were the fire area (m^2) and heat release rate (kW·m^{-2}) determined in accordance with (ČSN EN 1991-1-2, 2004). The following graph in Figure 5 shows the heat release rate per square meter (RHR_f).

4.3 *Fire detection and activation of the ZOKT*

Although both fire ventilation variants were designed differently, the activation of fire fans and logical consequences of fire safety equipment are identical.

The sports hall is guarded by an electric fire alarm system. The system reports fire using optical and acoustic alarms. It consists of automatic optical-smoke fire detectors with individual addressing, and call button points. These elements are connected to the electrical fire alarm system centre through circular lines. It is a two-stage alarm signalling system with day and night functions. Under the roof of the hall are suction devices consisting of a network of pipes, sucking in air samples from the protected space and bringing them to the laser detector.

For the purposes of the model, the function of the electrical fire alarm system was significantly reduced and the logical consequences of the system were simplified. Fire alarms work only using ceiling detectors which, in the model, are placed at a height of about 12 m above the floor of the

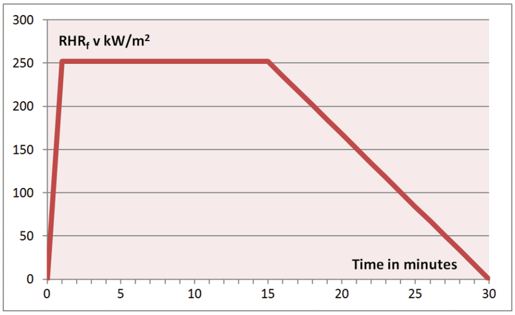


Figure 5. Graph of the heat release rate per square meter (RHR_f).

guarded area. In the FDS program, the optical-smoke fire detectors consisting of the pipe system are simulated using a *beamdetector* (McGrattan, 2010). The doors in the escape routes are open and the air supply is assured throughout the fire simulation.

The model does not include the effects of HVAC equipment. For the most realistic and reliable simulation, maximally two sections are activated upon detection.

5 THE RESULTS OF THE EFFECTIVENESS OF THE FIRE VENTILATION DESIGNS

In both simulations, a crucial factor for making a comparison of the fire ventilation design was the time required to evacuate persons from the hall area to walkways (t_u = 5.2 min ≈ 323 sec) and the time of filling the hall area with smoke (t_e = 11.49 min ≈ 690 sec), pursuant to (ČSN 73 0802, 2009). These values were determined from the starting project and were fixed for both options.

5.1 *The results of the first ZOKT design option*

When simulating a fire on the playing surface, the detectors were first activated in section no. 1 and then in section no. 3. Additional triggering of fire fans was not possible due to the operational limits of the reserve source. The graph in Figure 6 shows individual time intervals between the time of evacuation, activation of individual sections and the time of filling the space with smoke.

The time interval between the activation in section no. 1 and section no. 3 is very narrow. Fire fans in section no. 1 and section no. 3 were activated after 457 seconds and 496 seconds, respectively. Therefore, this demonstrates an early and intensive outflow of combustible gases outside the area under consideration.

After 60 seconds, the fire was in its beginning stages; after 323 seconds, the combustible gases accumulated at a height of about 13 m above the floor of the hall, i.e. about 3.8 m above the floor of the highest walkway (the highest level of the seating area). In terms of fire safety of construction and assessing the evacuation conditions, this height difference is safe. After 690 seconds, the smoke already reached the lower border of smoke barriers and gradually spread into adjacent sections. This fact correlates with the calculation and confirms its accuracy. At the time of the expected intervention by fire brigades, the hall was considerably filled with smoke; however, the performance and number of fire fans is dimensioned to meet the conditions for evacuating persons, conditions for active fire-fighting intervention and reducing thermal stress on the building structures.

5.2 *The results of the second ZOKT design option*

When simulating a fire on the playing surface, the detectors were first activated in section no. 5 and then in section no. 1. Additional triggering of fire fans was not possible due to the operational limits of the reserve source. The graph in Figure 7 again shows individual time intervals between the time of evacuation, activation of individual sections and the time of filling the space with smoke.

The time interval between the activation of central section no. 5 and section no. 1 is fairly broad. Fire fans in section no. 5 and section no. 1 were started after 459 seconds and 690 seconds, which corresponds to the time of filling the considered space with smoke. It can be therefore concluded that the initial outflow of combustible gases

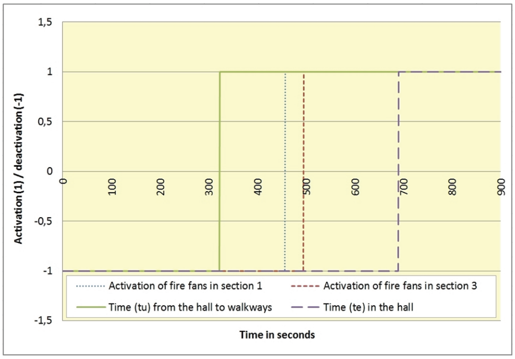


Figure 6. Graph of the time intervals in the first ZOKT design option.

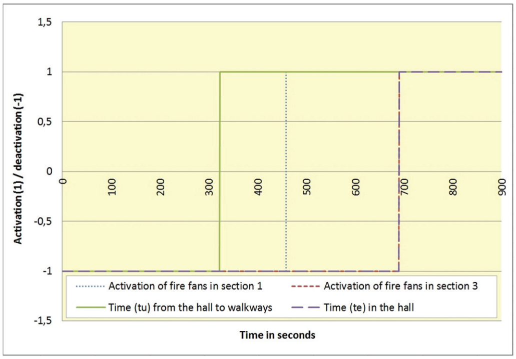


Figure 7. Graph of the time intervals in the second ZOKT design option.

through fire fans in section no. 5, located directly in the dome of the hall, was timely; however, the prolonged time interval between starting the other section indicates the possibility of increased speed of creating an accumulated layer of smoke in the area under consideration.

After 60 seconds, the fire was in its beginning stages; after 323 seconds, the combustible gases were present over the entire surface of the assessed section due to their spreading. Smoke movement was not prevented by any barrier; this resulted in smoke accumulation just under the roof structure at a height of 16 m above the hall floor, which is about 6.8 m above the floor of the highest walkway (the highest level of the seating area). In terms of fire safety of construction and assessing evacuation conditions, this height difference is safe. After 690 seconds, the lower smoke layer reached the imaginary boundary of the original smoke barriers (12 m above the playing surface). By this time, however, the space was ventilated only by 5 ceiling fans. Due to the low intensity of the outflow of combustible gases, the space was considerably filled with smoke and especially combustible gases spread into other smoke sections. The layer of clean air and its height under the smoke boundary above the playing surface correlates with the calculation and confirms its accuracy, but the density of smoke, its accumulation, spreading and thermal stresses on the building structures at that time were very troublesome. At the time of intervention by fire brigades, the space was already filled with so much smoke that the lower boundary of the smoke accumulation layer reached a height of about 5 m above the playing surface and reduced visibility. This confirmed the previous assumption of a very rapid process of filling the space with smoke due to the low power of fire fans. This lack resulted in a considerable restriction of active fire-fighting intervention and higher thermal stress on the building structures.

6 A COMPARISON OF THE ZOKT OPTIONS

6.1 *Activation of fire fans*

The results of the mathematical model clearly show that fire fans in the first sections are in both cases activated roughly at the same time interval, approximately after 458 seconds (after 457 seconds regarding section no. 1 in the first option and after 459 seconds regarding section no. 5 in the second option). In the first ZOKT design option, the fire fans of the other section (no. 3) are activated after 496 seconds, which is approximately at the same time. Fire fans in both sections are started simultaneously due to the division of smoke sections in the guarded area using the smoke barriers. Combustible gases thus accumulate in activated sections identically and there is no reason for a longer time delay. In the second ZOKT design option, the fire fans are activated with a significant time difference, after 690 seconds. This delay is due to the different way of dividing the space into smoke sections and the imaginary borders of these sections. The non-use of smoke barriers leads to the accumulation of combustible gasses under the highest part of the roof dome (section no. 5) and smoke spread throughout the area of the sports hall results in late activation of the fire fans in the other section.

6.2 *The formation of smoke*

The decisive factor in assessing the two different ZOKT projects was primarily the evacuation time (t_u = 323 sec) and the accumulation of combustible gases within this time interval. Figure 8 compares both options exactly after 323 seconds. With regard to the evacuation conditions, both ZOKT designs can be regarded as satisfactory, because the smoke is not so intense and does not endanger people escaping from the sports hall.

Such positive results are not achieved by fire ventilation designs at time intervals moving beyond the time of evacuation. The most noticeable difference in the function of fire ventilation consists in the time of intervention by fire brigades, after 900 seconds (15 minutes). The difference is illustrated in Figure 9. In case of the first option, the ZOKT design fulfils its function smoothly. The output from the fire fans and their activation is

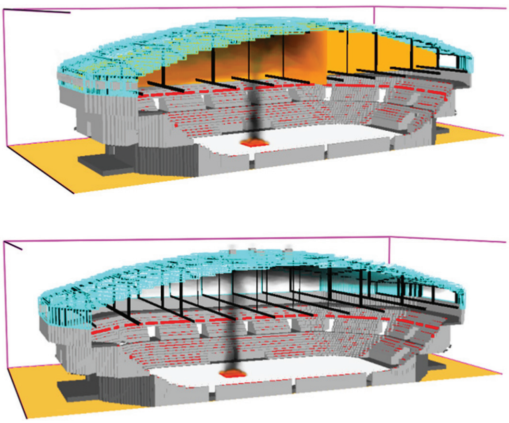


Figure 8. The accumulation of combustible gases after 323 seconds (option 1 on the top, option 2 on the bottom).

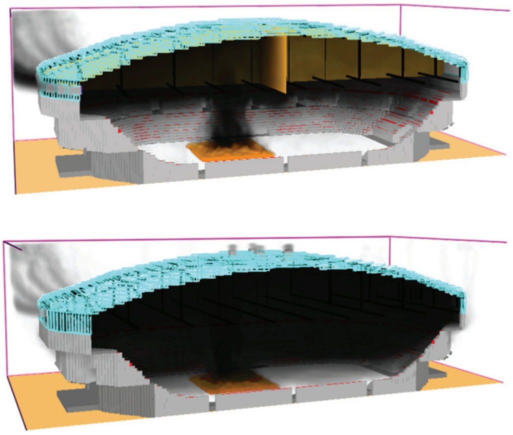


Figure 9. The accumulation of combustible gases after 900 seconds (option 1 on the top, option 2 on the bottom).

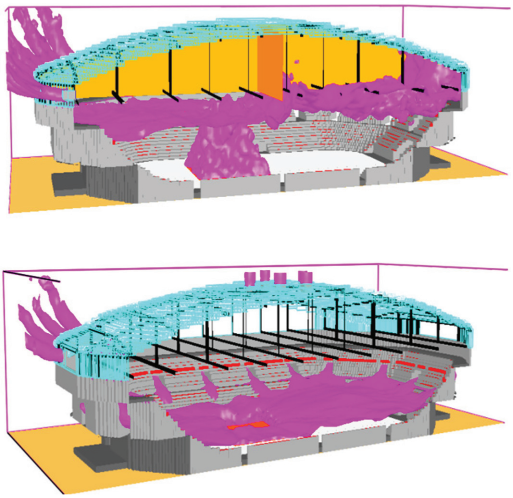


Figure 10. Distribution of the representative temperature of combustible gases 40°C (option 1 on the top, option 2 on the bottom).

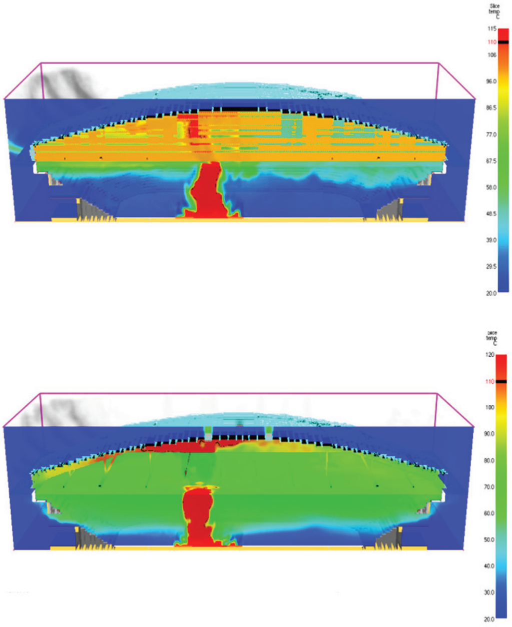


Figure 11. Distribution of the representative temperature of combustible gases 40°C (option 1 on the top, option 2 on the bottom).

sufficient to ventilate the sports hall and allows the fire brigades to perform their fire-fighting activities. Regarding the second option, the functioning of the fire ventilation cannot be considered as satisfactory. Smoke in the space and strongly reduced visibility may limit fire-fighting activities.

This negative function of fire ventilation is caused by the missing installation of smoke barriers, the unsatisfactory calculated performance of fire fans and their number in each given smoke section. Finally, the area of the sports hall is also thickly filled with smoke due to the wide time interval between starting individual sections in the second ZOKT design option.

6.3 *Representative temperatures of the combustible gases in the area*

In assessing temperatures, we created a simulation representing the lowest temperature of the combustible gases at the lower end of the smoke layer (40°C). This representative temperature was chosen primarily because it is the temperature that negatively affects the human body. The simulation was assessed at the time after 900 seconds when the accumulation of combustible gases was most intense. Figure 10 shows the level of this representative temperature. In the first ZOKT design option, the neutral plane occurs at a sufficient height. Temperatures below the neutral plane are lower and provide a sufficient area of clean air. In the second ZOKT design option, the neutral plane occurs nearly at the floor of the fire section. Above this plane, we can assume temperatures much higher than in the previous ZOKT design option; the thermal load acting on the building structures is therefore much stronger.

In order to have the clearest possible idea of achieving temperatures in the area, particularly above the level of the representative temperature

Table 1. The results of assessing the two different ZOKT designs for the sports hall.

Practical requirements	ZOKT design option 1	ZOKT design option 2
Early activation of fire fans in the primary section	YES	YES
Early activation of fire fans in the secondary section	YES	NO
Optimum smoke density and smoke accumulation	YES	NO
Ensuring conditions for the evacuation of people	YES	YES
Ensuring conditions for fire-fighting activities	YES	NO
Sufficient dimensioning of the performance of fire fans	YES	NO
Quality location of fire fans	NO	YES

of 40°C, we provided the following comparison images—see Figure 11. They demonstrate the distribution of temperatures in the area (i.e. temperatures of gases), in the middle of the sports hall and the height (12 m above the playing surface) where the smoke barriers start to occur.

Figure 11 illustrates temperatures in the area along a longitudinal plane. Temperature distribution in other planes is comparable; therefore, only one demonstration was selected. Nevertheless, despite this one-sided view, the temperatures are very distinct. In the first ZOKT design option, the space temperature reaches 115°C (in red). In the second ZOKT design option, the reached temperature is slightly higher, i.e. 120°C.

7 DISCUSSION

Using the Fire Dynamics Simulator simulation model, we looked at two different ZOKT design options for the protected area of the sports hall. Table 1 below presents the basic questions that were asked during the assessment of the two fire ventilation options.

The first ZOKT design option, where the guarded area is divided into four smoke sections using smoke barriers and where each section is ventilated by 12 fire fans, meets all the requirements of the proper project of equipment for smoke and heat removal. The only unsatisfactory aspect was the location of 12 fire fans in the wall of the roof structure. The ZOKT system could achieve higher efficiency if the fire fans were installed in the ceiling of the roof structure where the building structures are most thermally stressed. Such placement of fire fans is advisable to consult with experts through static drawings.

In the second ZOKT design option, the guarded area was divided into five smoke sections without using any smoke barriers; each section was ventilated by 5 fire fans; section 5 had a circular shape and its fire fans were installed in the ceiling of the roof structure, in the central highest point. Although this project meets the conditions of evacuation, it does not meet the conditions for effective fire-fighting activities. Thick smoke in the hall area at the time of the arrival of fire brigades would result in the slowing and hindering of such fire-fighting activities. It could also lead to changes in the material properties of steel structures due to their higher thermal stress.

8 CONCLUSION

Verification of the effectiveness of forced fire ventilation of a sports hall through mathematical modelling shows the first ZOKT design option to be optimistic, more practical, safer and more effective than the second ZOKT option.

Fire modelling is certainly a promising area, which will find its application in many practical situations and may also explain many ambiguities, especially in the case of interactions between fire safety systems. A combination of standard computational techniques and modelling seems to be the optimal approach that ultimately leads to financial savings and optimizations of project designs.

REFERENCES

ČSN 73 0802. 2009. *Fire protection of buildings – Non-industrial buildings*. Prague: Czech Office for Standards, Metrology and Testing, 122 p. (in Czech).

ČSN 73 0831. 2011. *Fire protection of buildings – Assembly rooms*. Prague: Czech Office for Standards, Metrology and Testing, 36 p. (in Czech).

ČSN EN 12101-1. 2006. *Smoke and heat control systems - Part 1: Specification for smoke barriers*. Prague: Czech Office for Standards, Metrology and Testing, 44 p. (in Czech).

ČSN EN 12101-3. 2003. *Smoke and heat control systems - Part 3: Specification for powered smoke and heat exhaust ventilators*. Prague: Czech Office for Standards, Metrology and Testing, 32 p. (in Czech).

ČSN EN 1991-1-2. 2004. *Eurocode 1: Actions on structures - Part 1-2: General actions - Actions on structures* *exposed to fire*. Prague: Czech Office for Standards, Metrology and Testing, 56 p. (in Czech).

ČSN P CEN/TR 12101-5. 2008. *Smoke and heat control systems—Part 5: Guidelines on functional recommendations and calculation methods for smoke and heat exhaust ventilation systems*. Prague: Czech Office for Standards, Metrology and Testing, 100 p. (in Czech).

Hanuška, Z. 1996. *Methodical instructions for for preparing fire-fighting documentation (2nd Edition)*, Prague: Fire Rescue Service of the Czech Republic. FACOM, 78 p. (in Czech). ISBN 80-902121-0-7.

Kučera, P., Pezdová, Z. *Základy matematického modelování požáru*. Association of Fire and Safety Engineering, (2010), 111 p. (in Czech). ISBN 978-80-7385-095-1.

McGrattan, K. et al. 2010. *Fire dynamics Simulator (Version 5) – User´s Guide*. NIST Special Publication 1019-5, National Institute of Standards and Technology, Building and Fire Research Laboratory, Maryland, USA.

Stochastic reliability modelling, applications of stochastic processes

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

The methods of parametric synthesis on the basis of acceptability region discrete approximation

Y. Katueva & D. Nazarov
Laboratory of Complex Systems Reliability Control, Institute of Automation and Control Processes FEB RAS, Vladivostok, Russia

ABSTRACT: The methods of parametric synthesis (parameter sizing) for providing parametric reliability using acceptability region discrete approximation with a regular grid are discussed. These methods are based on parametric optimization using deterministic criterion for the case of lack of information on parametric deviation trends and their distribution laws. A volume of a convex symmetrical figure inscribed into acceptability region is considered as an objective function of this optimization task. The methods of inscribing these figures into discrete approximation of acceptability region based on multidimensional implementation of Moore and von Neumann neighbourhoods are proposed in this work.

1 INTRODUCTION

The stage of parametric synthesis at analog engineering system design consists in determination of nominal parameter values which yield system performances within their requirements. Optimal parametric synthesis procedure requires the account of parameter deviations under influence of various factors including manufacturing ones to provide parametric reliability of an engineering system.

Unfortunately, the information on probabilistic characteristics of parametric deviations, and consequent system parametric faults are often unknown. Therefore, rational solutions on parameter sizing are required under the conditions of a lack of this information. For this purpose, the system should be designed robust to maximum set of parametric deviations, focusing on the worst case. Thus, as an objective function of optimal parametric synthesis, it is proposed to use system performance reserve esing, (Abramov et al. 2007), (Abramov and Nazarov 2012). Geometric methods for finding nominal parameter values, which provide maximum of system's performance reserve, based on discrete representation of an Acceptability Region (AR) with the use of a regular grid are proposed.

2 THE PARAMETRIC SYNTHESIS PROBLEM

Suppose that we have a system which depends on a set of n parameters $\mathbf{x} = (x_1,...,x_n)^T$. We will say that system is acceptable if $\mathbf{y}(\mathbf{x})$ satisfy the conditions (1):

$$\mathbf{y}_{\min} \le \mathbf{y}(\mathbf{x}) \le \mathbf{y}_{\max}, \tag{1}$$

where $\mathbf{y}(\mathbf{x})$, $\boldsymbol{y}_{\min}$ and $\boldsymbol{y}_{\max}$ are m-vectors of system responses (output parameters) and their specifications, e.g. $\mathbf{y}_1(\mathbf{x})$ - average power, $\mathbf{y}_2(\mathbf{x})$ - delay, $\mathbf{y}_3(\mathbf{x})$ -gain.

Possible variations of internal parameters are defined by the conditions of components' physical realizability, and their specifications. The constraints for these parameters variations represent a bounding box:

$$B_T = \{\mathbf{x} \in \mathbb{R}^n : x_{i\min} \le x_i \le x_{i\max}, \forall i = 1,2,\ldots,n\}. \tag{2}$$

The inequalities (1) define a region D_x in the space of design parameters

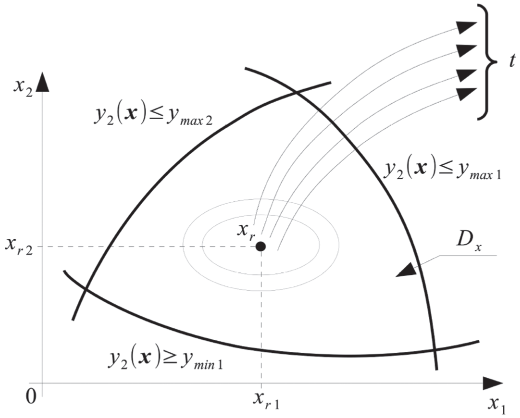
$$D_x = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{y}_{\min} \le \mathbf{y}(\mathbf{x}) \le \mathbf{y}_{\max}\}. \tag{3}$$

D_x is called the Region of Acceptability (RA) for the system. Figure 1 illustrates such a region.

The engineering system parameters are subject to random variations (aging, wear, temperature variations) and the variations may be considered as stochastic processes:

$$\mathbf{X} = \mathbf{X}(\mathbf{x}, t). \tag{4}$$

The stochastic processes of parameter variations $\mathbf{X}$ mean random manufacturing realization of system's components and thier upcoming degradation. Therefore, the conditions (1) can be met only with a certain probability



t
x_2
$y_2(\boldsymbol{x}) \le y_{max\,2}$
$y_2(\boldsymbol{x}) \le y_{max\,1}$
x_r
D_x
$x_{r\,2}$
$y_2(\boldsymbol{x}) \ge y_{min\,1}$
0
$x_{r\,1}$
x_1

Figure 1. Region of acceptability D_x defined by system response functions.

$$P(\mathbf{x}) = P(\mathbf{y}_{\min} \le \mathbf{y}(\mathbf{X}(\mathbf{x},t)) \le \mathbf{y}_{\max}, \forall t \in [0,T]) \qquad (5)$$

or

$$P(\mathbf{x}) = P(\mathbf{X}(\mathbf{x},t) \in D_x, \forall t \in [0,T]). \qquad (6)$$

The probability (5), (6) is an estimation of the designed system reliability.

In general, the parametric optimization (optimal parametric synthesis) problem can be stated as follows.

Given the characteristics of random processes $\mathbf{X}(t)$ of system parameters variations, a region of admissible deviation B_T and a service time T, find such a deterministic vector of parameter ratings (nominals) $\mathbf{x}^0 = (x_1^0,\dots,x_n^0)^T$ that the reliability (5), (6) be maximized (Abramov, Katueva, & Nazarov 2007).

Unfortunately, both the degradation model (4), and the region D_x are unknown. The practical way in uncertainty conditions consists in replacing the original stochastic criterion with a certain deterministic one. One of them is a so-called a “minimal serviceability reserve”.

System performance reserve allows to estimate the distance from system’s parameters vector $\mathbf{x}$ from AR boundary, ∂D_x, and, consequently, parameters’ variation margins which keep the performances within their specifications (1). In this case, the optimal parametric synthesis problem is reduced to finding of a point $\mathbf{x}$ which has maximum distance from the AR boundary (Abramov et al. 2007), (Abramov et al. 2008), (Abramov and Nazarov 2012), (Abramov and Nazarov 2015):

$$\mathbf{x} = \arg\max_{\mathbf{x}\in D_x} dist(\mathbf{x},\partial D_x). \qquad (7)$$

The distance $dist(\mathbf{x},\partial D_x)$ may be estimated as the shortest distance from point $\mathbf{x}$ to AR boundary ∂D_x in every coordinate direction. The next practical algorithm of the criterion (7) calculation is based on the approximation of region D_x.

Let’s note that the information about D_x configuration and its boundary ∂D_x is unknown. Among various methods of AR approximation proposed (Xu et al. 2015), (Director et al. 1978), (Krishna and Director 1995), (Bernacki et al. 1989), (Grasso et al. 2009), in this paper, we will use a discrete approximation, using a set of elementary hypercubes, based on multidimensional probing method on a regular grid (Abramov and Nazarov 2012).

In this work, the methods for solving optimal parametric synthesis problem (7) on the basis of discrete representation of AR using various objective functions.

3 THE ACCEPTABILITY REGION DISCRETE APPROXIMATION BASED ON A REGULAR GRID

A circumscribed box B_0

$$\begin{aligned} &B_0 = \{\mathbf{x} \in B_T : a_i^0 \le x_i \le b_i^0, \forall i = 1,2,\dots,n\},\\ &a_i^0 = \min_{\mathbf{x}\in D_x} x_i, b_i^0 = \max_{\mathbf{x}\in D_x} x_i \end{aligned} \qquad (8)$$

is usually constructed prior to AR construction in order to narrow search area and can be considered as zero-order estimation of AR configuration. Usually Monte-Carlo method is used to determine its borders $\mathbf{a}^0$, $\mathbf{b}^0$ (Abramov et al. 2007).

The discrete approximation of AR is constructed on the basis of B_0 with equidistant splitting of every i-th parameter’s axis within $\left[a_i^0, b_i^0\right], i = \overline{1,n}$ into $l_i, i = \overline{1,n}$ ranges which forms a regular grid inside this circumscribed box B_0.

The grid nodes define a vertices of Elementary Boxes (EB) which are used for AR approximation. Every single EB is identified with a set of indices $(k_1 k_2 \dots k_n)$, where $1 \le k_i \le l_i, i = \overline{1,n}$. All these EB comprise a set $B_g = \left\{e_{k_1 k_2 \dots k_n} : 1 \le k_i \le l_i \forall i = \overline{1,n}\right\}$. The amount of EB can be calculated using (9):

$$L = \left|B_g\right| = \prod_{i=1}^{n} l_i. \qquad (9)$$

The grid step $h_i, i = \overline{1,n}$ for every parameter is obtained using

$$h_i = (b_i^0 - a_i^0)/l_i, i = \overline{1,n}. \qquad (10)$$

The bounds of an EB can be obtained using the following expressions:

$$a_i^{k_i} = a_i^0 + (k_i + 1)h_i,$$
$$b_i^{k_i} = a_i^{k_i} + h_i, k_i = \overline{1, l_i} i = \overline{1, n}. \tag{11}$$

An EB consists of points inside parameter space within its bounds:

$$e_{k_1 k_2 \ldots k_n} = \left\{ \mathbf{x} \in B_0 : a_i^{k_i} \le x_i \le b_i^{k_i} \ \forall i = \overline{1, n} \right\}. \tag{12}$$

Thus, the circumscribed box B_0 can be considered as a union of EB as it is expressed in (13):

$$B_0 = B_0^g = \bigcup_{k_1=1}^{l_1} \bigcup_{k_2=1}^{l_2} \ldots \bigcup_{k_n=1}^{l_n} e_{k_1 k_2 \& k_n}. \tag{13}$$

Although every EB is uniquely identified with a set of indices $(k_1 k_2 \ldots k_n)$, the most appropriate way of their enumeration is using scalar index $p = 1, 2, \ldots L$ which is univalently associated with the EB indices, and can be calculated using (14):

$$p(k_1 k_2 \ldots k_n) = k_1 + (k_2 - 1)l_1 +$$
$$+ (k_3 - 1)l_1 l_2 + \ldots + (k_n - 1)l_1 l_2 \ldots l_{n-1}. \tag{14}$$

Inverse indices transformation $k_i(j)$ can be obtained with sequential calculations using (15):

$$k_n = \frac{p - 1}{l_1 l_2 \ldots l_{n-1}} + 1$$
$$k_{n-1} = \frac{p - k_n l_1 l_2 \ldots l_n}{l_1 l_2 \ldots l_{n-2}} + 1$$
$$\ldots \tag{15}$$
$$k_1 = p - \sum_{i=2}^{n} \left((k_i - 1) l_1 \ldots l_{i-1} \right), n > 1.$$

Thus, every EB can be enumerated with scalar index $e_p \in B_g, p = 1, \ldots L$.

Every EB is assigned with a representative point $\mathbf{x}_r^{k_1 k_2 \ldots k_n} = (r_1^{k_1}, r_2^{k_2}, \ldots r_n^{k_n})^T$ usually located in its geometric centre:

$$r_i^{k_i} = a_i^0 + h_i k_i - \frac{h_i}{2}, i = \overline{1, n}. \tag{16}$$

The Figure 2 illustrates the AR discrete approximations for 2-dimension input parameter space.

According to univalent relation between indices $(k_1 k_2 \ldots k_n)$ and p given in (14) and (15), lets denote $\mathbf{x}_r^p = \mathbf{x}_r^{k_1 k_2 \ldots k_n}$. A representative point can be referenced both by indices $(k_1 k_2 \ldots k_n)$ set and corresponding scalar index p.

Membership characteristic function for AR discrete approximation is defined in (17):

$$\chi(\mathbf{x}_r^{k_1 k_2 \ldots k_n}) = \begin{cases} 1, if \ \mathbf{y}_{\min} \le \mathbf{y}(\mathbf{x}_r^{k_1 k_2 \ldots k_n}) \le \mathbf{y}_{\max}, \\ \qquad 0 \ otherwise. \end{cases} \tag{17}$$

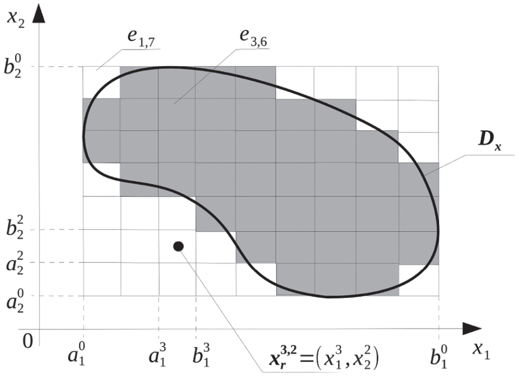


Figure 2. Region of acceptability D_x discrete approximations.

RA discrete approximation is comprised of EB which representative points fulfilling performances' design specifications (1):

$$D_x^g = \left\{ e_{k_1 k_2 \ldots k_n} \in B_0 : \chi(\mathbf{x}_r^{k_1 k_2 \ldots k_n} = 1 \right\}. \tag{18}$$
$$\forall k_1 = \overline{1, l_1}, \ldots, k_n = \overline{1, l_n}.$$

If an index set for all EB comprising AR approximation is defined as:

$$I_{D_x} = \left\{ p \in 1, 2, \ldots L : \chi(\mathbf{x}_r^{k_1 k_2 \ldots k_n} = 1 \right\}, \tag{19}$$

where $\mathbf{x}_r^p = \mathbf{x}_r^{k_1 k_2 \ldots k_n}$ and $k_i = k_i(p)$ according to (15), we can reformulate (18) into the following form (20):

$$D_x^g = \left| e_p \in B_0 : p \in I_{D_x} \right|. \tag{20}$$

The amount of elementary boxes which approximate AR is denoted by $M = \left| D_x^g \right|$, and it is true also that $M = \left| I_{D_x} \right|$.

Data structures for storing information on grid and set D_x^g are described in (Abramov et al. 2008), and (Abramov and Nazarov 2012). Essential difficulty of AR approximation consists in high dimension of parameter space, incomplete prior information and only pointwise exploration of parameter space with system performances (1) calculation. The application of parallel computing significantly facilitates AR approximation (Abramov et al. 2009), (Abramov and Nazarov 2015).

The AR data structures can be also exploited for calculation of a centroid $\mathbf{x}^c = (x_1^c, x_2^c, \ldots x_n^c)^T$ during determination of the set D_x^g which approximates AR. The centroid coordinates are calculated with the following expression:

$$x_i^c = \frac{\sum_{p=1}^{M} x_r^p}{M} p \in I_{D_x}, i = \overline{1,n}, \tag{21}$$

where representative points are taken only from EB indexed in the AR approximation index set (19). It was proven that for convex AR this centroid $\mathbf{x}^c$ is a solution to optimal parametric synthesis problem (Abramov et al. 2007).

4 THE CRITERION OF SYSTEM PERFORMANCE RESERVE

The Criterion of System Performance Reserve (7) implies calculation of the minimal distance r to AR boundary for every EB which belongs to AR approximation, excluding boundary EB. Every EB $e_{k_1k_2\ldots k_n} \in B_g$ is assigned a weight $w(e_{k_1k_2\ldots k_n}) = r$ of distance to the nearest boundary EB. After assigning weights, the EB with maximal weight are emphasized as optimal ones.

For a discrete approximation of AR, the problem (7) of finding optimal parameters vector consists in finding EB which acts as a centre of a convex symmetrical figure comprised of EB from AR approximation. The volume of this figure is calculated as the amount of EB it consists of. The measurement unit of distances between EB (in particular from any EB to AR boundary) is an EB rib.

Lets consider various methods of measurement of a distance from any EB $e_{k_1k_2\ldots k_n} \in D_x^g$ to AR boundary ∂D_x^g. As the minimal distance to AR boundary can not exceed a half of bounding box range (for the most optimistic case when AR fully fills the bounding box), it is correct to suppose that

$$r \le \min_{\forall i=\overline{1,n}} (k_i^c - 1, l_i - k_i^c), \tag{22}$$

where $l_i, i = 1,2,\ldots n$ is the amount of grid steps for i-th parameter. It is reasonable to say that $r = 0$ for every EB which belongs to AR boundary.

One of the methods of determining of the shortest distance from an EB to AR boundary consists in constructing of a maximal cube comprised of EB from AR approximation with the centre in this EB. This multidimensional cube with the range r is called r-cube. The method of r-cube construction is based on Moore neighborhood (Kier et al. 2005), (Schiff 2008) idea applied to multidimensional space as illustrated in Figure (3).

In addition to the expression (22) we can say that the distance r for every EB $e_{k_1k_2\ldots k_n}$ which acts as a center of r-cube has the following limitations:

$$0 \le r \le \min_{\forall i=\overline{1,n}} (k_i - 1, l_i - k_i). \tag{23}$$

The example of application the method of optimal EB selection via inscribing r-cube with maximal r for 2-dimensional parameter space is illustrated in Figure (4).

Lets define r-neighborhood of an EB $e_{k_1k_2\ldots k_n} \in B_g$ with indices $(k_1k_2\ldots k_n)$ as a set $E_{k_1k_2\ldots k_n}^r \in B_g$ comprised of EB with indices $(m_1m_2\ldots m_n)$ which fulfill (24)

$$\sum_{i=1}^{n} |m_i - k_i| \le r. \tag{24}$$

Thus, the set $E_{k_1k_2\ldots k_n}^r$ is a figure defined with (25):

$$E_{k_1k_2\ldots k_n}^r = \left\{ e_{m_1m_2\ldots m_n} \in B_g : \sum_{i=1}^{n} |m_i - k_i| \le r \right\}. \tag{25}$$

The optimal parametric synthesis problem for providing maximal system performance reserve consists in searching of an EB $e_{k_1k_2\ldots k_n}^r \in D_x^g$ which acts as a center of maximal r-neighborhood $E_{k_1k_2\ldots k_n}^r \in D_x^g$ comprised of EB which belong to AR approximation D_x^g

The value r for r-neighborhood of $e_{k_1k_2\ldots k_n}^r \in D_x^g$ an EB $E_{k_1k_2\ldots k_n}^r \in D_x^g$ expresses the minimal Manhattan distance (Kier et al. 2005), (Schiff 2008) from this EB to AR approximation boundary. A set of EB which have maximal Manhattan distance to AR boundary and act as centers of maximal r-neighborhoods in two-dimensional space is illustrated in Figure (5).

The determination of r-neighborhoods implements the method of narrowing areas which requires analytical description of AR boundary with a finite set of hypersurfaces and consists in iterative narrowing of the region towards the optimal center. As for the discrete AR representation, the step unit per one iteration of the region narrowing is one EB (one grid step (10). The procedure of

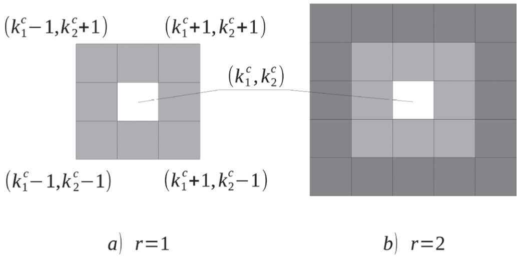


Figure 3. Two-dimensional r-cubes. (Left) $r = 1$; (right) $r = 2$.

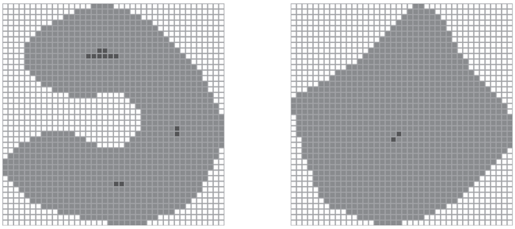


Figure 4. The centers of maximal r-cube inscribed.

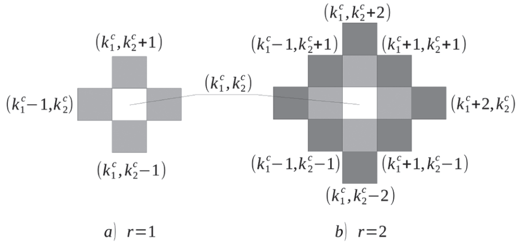


Figure 5. Two-dimensional example of r-neighborhood with (left) $r = 1$; (right) $r = 2$.

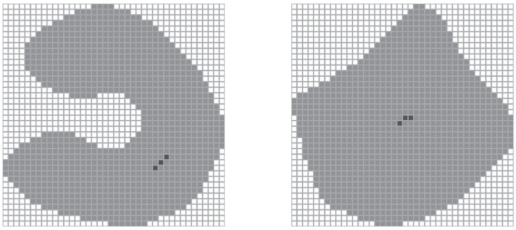


Figure 6. The centers of maximal r-neighborhoods inscribed.

narrowing of AR discrete approximation boundary performs moving neighborhood boundary to the adjacent EB backward from nearest boundary B_g^B of EB set (25).

A zero-order narrowing boundary is represented by the AR discrete approximation boundary ∂D_x^g. A first-order narrowing border consists of EB with Manhattan distance to AR boundary equal to 1, i.e. first-order narrowing border is comprised of EB which have maximal r-neighborhood with $r = 1$. The next level of narrowing boundary is comprised of EB with r-neighborhood for $r = 2$, and so on until narrowing boundary is merged into separate EB. These EB are considered as optimal ones in the sense of system performance reserve (7).

For every single EB, a maximal narrowing border level coefficient can be calculated. This coefficient equals to minimal Manhattan distance from this EB to the boundary ∂D_x^g of discrete AR approximation. This maximal narrowing border level coefficient can be evaluated via constructing maximal r-neighborhood around this EB. Thus, the method of narrowing boundaries in the case of discrete AR approximation is reduced to determination of EB with maximal r-neighborhood (see Figure 6).

Despite high computational cost of EB enumeration, using discrete RA approximation in the task of narrowing regions has the following advantages:

- There is not need to determine new points at every iteration of region narrowing;
- There is no need to approximate the boundary with hyper-surfaces (polynomial or hyperspherical);
- There is no need to determine equidistant surfaces and sets of their contact points at every single iteration;
- No checks of the narrowing boundary degeneration into a point are required;
- The step of boundary narrowing is fixed and equal to the length EB rib (10).

The algorithm of determination of the maximal r-cube allows to obtain global solution of the parametric synthesis problem (7) for enough large and convex AR, the algorithm of determination of maximal r-neighborhood yields a set of local optimal solutions of the problem (7).

5 CONCLUSIONS

It is evident that discrete acceptability region approximation described in this work consumes computational resources during its construction and requires much resources for storing its data, and powerful computing facilities with parallel computing technologies should be widely involved during solving of this task. Nevertheless such approximation provides the most complete and detailed description of multidimensional acceptability region configuration a priori unknown, which is typical for actual complex systems with plenty of varying internal parameters. The methods proposed in this work are based on the criterion of maximal performance reserve, and aimed to facilitation of parametric synthesis task solution with the account of parametric reliability requirements, and lack of information on parametric deviation trends. Locating system internal parameters into the center of a figure inscribed in an acceptability region implements worst-case strategy of parameter sizing task. Inscribing of a figure of maximal volume in acceptability region maximizes system performance reserve within the scope of this worst-case strategy. Acceptability region can also be used in the task of evaluation of variation ranges within the region which allows to estimate system sensitivity and reveal its vulnerability to deviations of a particular parameter.

The algorithms proposed in this work were tested on models of various analog circuits (transistor-transistor logic, multivibrators, amplifiers).

ACKNOWLEDGEMENT

This work is partially supported with the grant 14-08-00149 of the Russian Foundation for Basic Research.

REFERENCES

Abramov, O., Y. Katueva, & D. Nazarov (2007). Reliabilitydirected distributed computer-aided design system. *Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management, Singapore.*, 1171–1175.

Abramov, O., Y. Katueva, & D. Nazarov (2008). Construction of acceptability region for parametric reliability optimization. *Reliability & Risk Analysis: Theory & Applications. 1 No.3*, 20–28.

Abramov, O., Y. Katueva, & D. Nazarov (2009). Distributed computing environment for reliability-oriented design. *Reliability & Risk Analysis: Theory & Applications. 2, No.1(12)*, 39–46.

Abramov, O. & D. Nazarov (2012). Regions of acceptability in reliability design. *Reliability: Theory & Applications. 7, No.3(26)*, 43–49.

Abramov, O. & D. Nazarov (2015). Regions of acceptability using reliability-oriented design. *Recent Developments on Reliability, Maintenance and Safety. WIT Transaction on Engineering Sciences 108*, 376–387, doi:10.2495/QR2MSE140431.

Bernacki, R., J. Bandler, J. Song, & Q.J. Zhang (1989). Efficient quadratic approximation for statistical design. *IEEE Transactions on Circuits and Systems. 36, no. 11*, 1449–1454.

Director, S., G. Hatchel, & L. Vidigal (1978). Computationally efficient yield estimation procedures based on simplicial approximation. *IEEE Trtansactions on Circuits and Systems. 25, no. 3*, 121–130.

Grasso, F., S. Manetti, & M. Piccirilli (2009). A method for acceptability region representation in analogue linear networks. *International Journal of Circuit Theory and Applications. 37*, 1051–1061, doi:10.1002/cta.518.

Kier, L.B., P.G. Seybold, & C.K. Chao-Kun Cheng (2005). *Modeling Chemical Systems using Cellular Automata*. Netherlands: Springer.

Krishna, K. & S. Director (1995). The linearized performance penalty (lpp) method for optimization of parametric yield and its reliability. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. 14, no. 12*, 1557–1568.

Schiff, J.L. (2008). *Cellular automata: a discrete view of the world*. University of Auckland: John Wiley & Sons inc.

Xu, X.B., Z. Liu, Y.W. Chen, D.L. Xu, & C.L. Wen (2015). Circuit tolerance design using belief rule base. *Mathematical Problems in Engineering 2015*, 12 p. doi: 10.1155/2015/908027.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Inference for a one-memory Self-Exciting Point Process

E. Gouno & R. Damaj
Laboratoire de Mathématiques de Bretagne Atlantique, Université de Bretagne Sud, Vannes, France

ABSTRACT: We consider a one-memory self-exciting point process with a given intensity function. Properties of this point process are studied and the Mino distribution is introduced as the interarrival times distribution of the process. This distribution has a hazard function which decreases or increases over a period of time before becoming constant. A maximum likelihood procedure is driven to obtain the MLE of the Mino distribution parameters. Quality of the estimates is investigated through generated data.

1 INTRODUCTION

Introduced by Hawkes (1971), Self-Exciting Point Processes (SEPP) are counting processes which intensity depends on all or part of the history of the process itself. These models find applications in many fields: seismology (Ogata 1999), neurophysiology (Johnson 1996), genetics, epidemiology, reliability (Ruggeri & Soyer 2008), economy (Bowsher 2007), economy (Bowsher 2007). The intensity of SEPP is not only a function of time but it is also a function of the number of jumps occurring in the process. In some situations, only a small number of more recent events will influence the evolution of the process; then the process is a self-exciting processes with limited memory. In this work, we consider the case where the intensity process of the self-exciting point process depends only on the latest occurrence, i.e. is a one-memory self-exciting point processes. The motivation to investigate such a one-memory SEPP is a reliability study of electrical equipments exposed to thunderstorm. In this study, this process appears as a good candidate to describe the effect of lightning strike on the reliability of the equipments. Therefore a method is required to make inference on one-memory SEPP. We assume that the impulse response function characterizing the intensity process is modeled as an exponential function having a constant coefficient that takes positive or negative values. This model has been considered by Mino (2001) who suggested a method using an EM algorithm, to obtain the maximum likelihood estimates of the parameters without solving the non linear optimization problems usually involved. We introduce and define the Mino distribution to describe the distribution of the interarrival times. Then we show that the SEPP considered by Mino is a renewal process where the interarrival times has a Mino distribution. Maximum likelihood estimation of the process intensity parameters is considered. The method is applied on simulated data. The results are compared with those obtained by Mino (2001).

2 ONE-MEMORY SELF-EXCITING POINT PROCESSES

We focus our attention on the one-memory self-exciting point processes $\{N(t), t \geq 0\}$ with intensity process:

$$\lambda(t) = \mu(1 + \alpha e^{-\beta(t - w_{N(t)})}) \tag{1}$$

where $w_{N(t)}$ is the occurrence time of the $N(t)^{th}$ jump, $\mu > 0$, $\alpha \geq -1$ et $\beta > 0$.

If $\alpha = 0$ or if β goes to $+\infty$, $\lambda(t) = \mu$ and the process is a standard homogeneous Poisson process.

If $\alpha > 0$, $\lambda(t)$ increases after each jump of the process; the process is said to be *excited*.

If $-1 \leq \alpha < 0$, $\lambda(t)$ decreases after each jump of the process; the process is said to be *inhibited*. The Figures 1 and 2 show representations of intensity for different values of μ, α, β. These representations are obtained from simulations of occurrence dates of the process $w_1, \ldots, w_{N(T)}$ (see appendix A).

Proposition 1. – *Let $\{N(t), t \geq 0\}$ be a one-memory self-exciting point process with intensity process $\{\lambda(t), t \geq 0\}$ defined as in (1). Then the interarrival times $T_i, i = 1, 2, \ldots$ form a sequence of statistically independent random variables with cumulative distribution function:*

$$Pr(T_i \leq t) = 1 - \exp\left\{-\mu\left[t + \frac{\alpha}{\beta}\left(1 - e^{-\beta t}\right)\right]\right\} \tag{2}$$

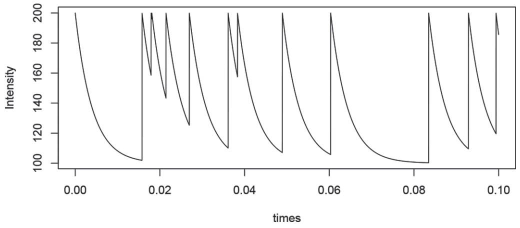


Figure 1. Intensity of an excited one-memory self-exciting point process with: $\alpha=1$, $\mu=100$, $\beta=250$ et $T=0.1$ ms.

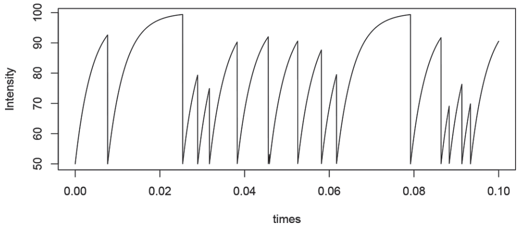


Figure 2. Intensity of an inhibited one-memory self-exciting point process with: $\alpha=-0.5$, $\mu=100$, $\beta=250$ et $T=0.1$ ms.

and the probability density function is:

$$f(t)=\mu(1+\alpha e^{-\beta t})\exp\left\{-\mu[t+\frac{\alpha}{\beta}(1-e^{-\beta t})]\right\} \tag{3}$$

Proof: The proof follows from Snyder and Miller (1991), theorem 6.3.4 p.314 and its corollary p.316. We have:

$$\begin{aligned} P(T_i>t) &= P[N(w_{i-1},w_{i-1}+t)=0] \\ &= \exp\left\{-\int_{w_i}^{w_i+t}\lambda(s)ds\right\} \\ &= \exp\left\{-\int_{w_{i-1}}^{w_{i-1}+t}\mu(1+\alpha e^{-\beta(s-w_{i-1})})ds\right\} \\ &= \exp\left\{-\mu[t+\frac{\alpha}{\beta}(1-e^{-\beta t})]\right\} \end{aligned}$$

from which (2) and (3) are easily deduced.

In the sequel, we say that a random variable X with a probability distribution of the form (1) follows a *Mino distribution* with parameters (μ,α,β). We denote: $X \sim Mino(\mu,\alpha,\beta)$.

It follows from proposition 1 that a one-memory self-exciting point process with intensity (1), is equivalent to a renewal process with independent interarrivals times having a Mino distribution. In the next section, we investigate some properties of the Mino distribution.

3 THE MINO DISTRIBUTION

Let X be a r.v. following a Mino distribution (3) with parameters (μ,α,β). When $\alpha=0$ or when β goes to $+\infty$, X follows an exponential distribution with parameter μ. The Figures 3 and 4 display the density for different parameters values. The hazard function for the Mino distribution is displayed in Figure 5. One can see that this distribution is convenient to model random variables with decreasing or increasing hazard rate in early life that becomes constant in useful life. To express the expectation of a Mino r.v., we introduce the function Γ^* defined by:

$$\Gamma^*(x;a)=\int_0^x z^{a-1}\,e^z dz. \tag{4}$$

Setting $\eta=\frac{\mu}{\beta}$ and assuming $\alpha>0$, the mean of X can be computed as:

$$\mathrm{E}\{X\}=\frac{\Gamma^*(\alpha\eta;\eta)}{\beta(\alpha\eta)^\eta}\,e^{-\eta\alpha}. \tag{5}$$

If $\alpha<0$, the mean can be expressed as:

$$\mathrm{E}\{X\}=\frac{\Gamma(|\alpha|\eta;\eta)}{\beta(|\alpha|\eta)^\eta}\,e^{-\eta\alpha}. \tag{6}$$

where $\Gamma(.,.)$ is the lower incomplete gamma function that is:

$$\Gamma(x;a)=\int_0^x u^{a-1}e^{-u}du.$$

Some examples of expectation values are given in Table 1.

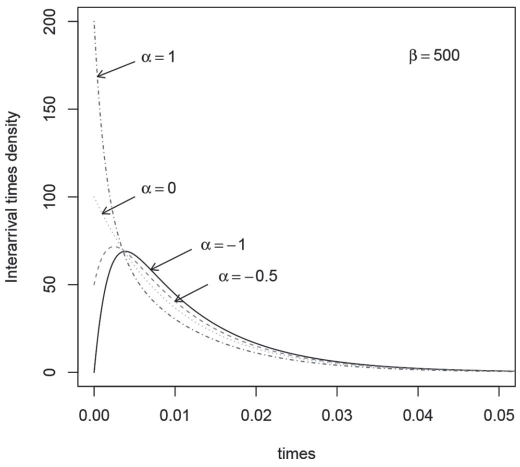


Figure 3. Probability density function of a Mino distribution for $\mu=100$, $\beta=50$ and different values of α.

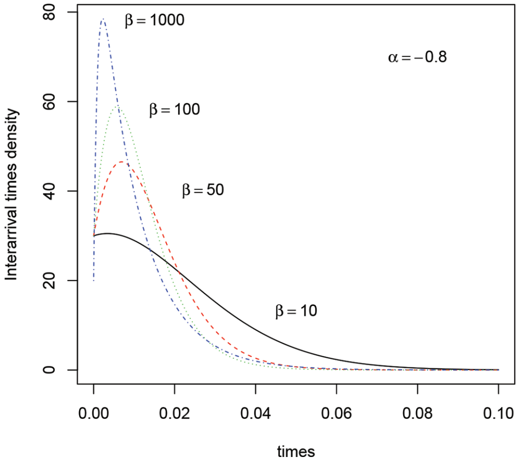


Figure 4. Probability density function of a Mino distribution with $\mu = 100$, $\alpha = -0.8$ and different values of β.

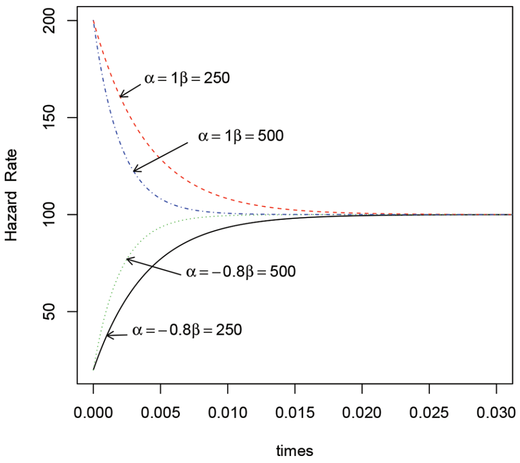


Figure 5. Hazard rate function of a Mino Distribution with $\mu = 100$.

Table 1. Examples of expected-value for β = 500.

	α	$E(X)$
	–1	0.011828
μ = 100	–0.5	0.010872
	1	0.007937
	–1	0.006697
μ = 200	–0.5	0.005777
	1	0.003770

4 PARAMETERS ESTIMATION

Mino (2001) proposes to use an EM algorithm to estimate the parameters. He introduces artificially two data models; one representing point event occurrences (observable) and the other one representing no point events occurrences (unobservable). The author claims that this approach provides simpler equations to realize the maximum likelihood estimates avoiding the classical nonlinear optimisation problem. We suggest to consider the classical maximum likelihood method using the Mino distribution that we have defined previously. This approach is easy to implement and allows to check the existence and uniqueness of the MLE.

Let $(t_1,\ldots,t_n)$ be a n-sample of independent random variables from a Mino distribution with parameters (μ,α,β). The log-likelihood is:

$$\log L(\mu,\alpha,\beta) = n\log\mu + \sum_{i=1}^{n}\log\left(1+\alpha e^{-\beta t_i}\right) \\ -\mu\sum_{i=}^{n} t_i - \frac{\mu\alpha}{\beta}\sum_{i=1}^{n}\left(1-e^{-\beta t_i}\right) \tag{7}$$

Remark that this expression is similar to the expression obtained considering the logarithm of the sample function for a self-exciting point process given by theorem 6.2.2, p. 302. from Snyder and Miller (1991).

From (7), the likelihood equations are:

$$\begin{cases} -\sum_{i=1}^{n} t_i - \frac{\alpha}{\beta}\sum_{i=1}^{n}(1-e^{-\beta t_i}) + \frac{n}{\mu} = 0 & (8) \\ -\frac{\mu}{\beta}\sum_{i=1}^{n}(1-e^{-\beta t_i}) + \sum_{i=1}^{n}\frac{e^{-\beta t_i}}{1+\alpha e^{-\beta t_i}} = 0 & (9) \\ \frac{\mu\alpha}{\beta^2}\sum_{i=1}^{n}(1-e^{-\beta t_i}) - \frac{\mu\alpha}{\beta}\sum_{i=1}^{n} t_i e^{-\beta t_i} & (10) \\ -\alpha\sum_{i=1}^{n}\frac{t_i e^{-\beta t_i}}{1+\alpha e^{-\beta t_i}} = 0 & (11) \end{cases}$$

A Newton-Raphson algorithm is used to solve this system of equations which does not admit explicit solution. Existence and uniqueness of the MLE can be proven applying theorem 2.6 p.761 from Mäkeläinen et al. (1981). One needs to prove that the gradient vector vanishes in at least one point and that the Hessian matrix is negative definite at every point where the gradient vanishes. The Hessian matrix is:

$$H = \begin{bmatrix} \frac{\partial^2 \log L}{\partial\mu^2} & \frac{\partial^2 \log L}{\partial\mu\partial\alpha} & \frac{\partial^2 \log L}{\partial\mu\partial\beta} \\ \frac{\partial^2 \log L}{\partial\mu\partial\alpha} & \frac{\partial^2 \log L}{\partial\alpha^2} & \frac{\partial^2 \log L}{\partial\alpha\partial\beta} \\ \frac{\partial^2 \log L}{\partial\mu\partial\beta} & \frac{\partial^2 \log L}{\partial\alpha\partial\beta} & \frac{\partial^2 \log L}{\partial\beta^2} \end{bmatrix}$$

where L stands for $L\,(\mu,\alpha,\beta)$. We have:

$$\frac{\partial^2 \log L}{\partial \mu^2} = -\frac{n}{\mu^2}$$

$$\frac{\partial^2 \log L}{\partial \alpha^2} = -\sum_{i=1}^{n} \frac{e^{-2\beta t_i}}{(1+\alpha e^{-\beta t_i})^2}$$

$$\frac{\partial^2 \log L}{\partial \beta^2} = -\frac{2\mu\alpha}{\beta^3}\sum_{i=1}^{n}(1-e^{-\beta t_i}) + \frac{2\mu\alpha}{\beta^2}\sum_{i=0}^{n} t_i e^{-\beta t_i}$$
$$+\frac{\mu\alpha}{\beta}\sum_{i=1}^{n} t_i^2 e^{-\beta t_i} + \alpha\sum_{i=1}^{n}\frac{t_i{}^2 e^{-\beta t_i}}{(1+\alpha e^{-\beta t_i})^2}$$

$$\frac{\partial^2 \log L}{\partial \mu \partial \alpha} = -\frac{1}{\beta}\sum_{i=1}^{n}(1-e^{-\beta t_i})$$

$$\frac{\partial^2 \log L}{\partial \mu \partial \beta} = \frac{\alpha}{\beta^2}\sum_{i=1}^{n}(1-e^{-\beta t_i}) - \frac{\alpha}{\beta}\sum_{i=1}^{n} t_i e^{-\beta t_i}$$

$$\frac{\partial^2 \log L}{\partial \alpha \partial \beta} = \frac{\mu}{\beta^2}\sum_{i=1}^{n}(1-e^{-\beta t_i}) - \frac{\mu}{\beta}\sum_{i=1}^{n} t_i e^{-\beta t_i}$$
$$-\sum_{i=1}^{n}\frac{t_i e^{-\beta t_i}}{1+\alpha e^{-\beta t_i}} + \alpha\sum_{i=1}^{n}\frac{t_i e^{-2\beta t_i}}{(1+\alpha e^{-\beta t_i})^2}$$

The Hessian is negative definite if the first upper-left minor is negative, the second upper-left minor is positive and the principal minor of order 3 is negative. The first upper-left minor is obviously negative.

The second upper-left minor D_2 is:

$$D_2 = \frac{n}{\mu^2}\sum_{i=1}^{n} A_i^2 - \frac{1}{\beta^2}\left[\sum_{i=1}^{n}(1-e^{-\beta t_i})\right]^2$$

where $A_i = \frac{e^{-\beta t_i}}{1+\alpha e^{-\beta t_i}}$.

Equation (9) gives $\sum_{i=1}^{n}(1-e^{-\beta t_i}) = \frac{\beta}{\mu}\sum_{i=1}^{n} A_i$. Thus

$$D_2 = \frac{1}{\mu^2}\left[n\sum_{i=1}^{n} A_i^2 - \left(\sum_{i=1}^{n} A_i\right)^2\right]$$
$$\geq \frac{1}{\mu^2}\left[\frac{1}{n}\sum_{i=1}^{n} A_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} A_i\right)^2\right]$$
$$= \frac{1}{\mu^2 n}\sum_{i=1}^{n}\left(A_i - \frac{1}{n}\sum_{i=1}^{n} A_i\right)^2 \geq 0$$

For the principal minor of ordre 3, the following expression can be obtained:

$$D_3 = \frac{\alpha}{\mu^2}\left\{\left[\frac{\mu}{\beta}\sum_{i=1}^{n} t_i^2 e^{-\beta t_i} + \sum_{i=1}^{n} t_i^2 e^{\beta t_i} A_i^2 - \frac{2}{\beta}\sum_{i=1}^{n} t_i A_i\right]\mu^2 D_2 + \alpha\sum_{j=1}^{n}\left[\left(\sum_{i=1}^{n} t_i A_i^2\right) - A_j\sum_{i=1}^{n} t_i A_i\right]^2\right\}$$

The sign of D_3 can be studied considering conditions on the parameters.

5 APPLICATIONS

The MLE is computed for different sets of parameters values and samples sizes identical to those chosen by Mino (2001). 100 samples of inter-arrival times are generated using the inversion method described in appendix A, for each setting. The means and the standard deviation of the MLE are displayed in Tables 2 and 3. One can see that the MLE are very close to the input parameters for μ and α. The results for β are slightly worse. The standard deviation is rather small in all cases and decreases as the sample size increases as expected.

Table 2. Mean and S.D. of the MLE calculated from 100 Monte Carlo runs.

Sample size	True Parameters	Estimates means	S.D.
20000	$\mu = 100$	99.753	0.099
	$\alpha = -1$	−1.002	0.001
	$\beta = 500$	521.867	2.309
10000	$\mu = 100$	99.820	0.137
	$\alpha = -1$	−1.020	0.001
	$\beta = 500$	517.082	3.047
5000	$\mu = 100$	99.595	1.790
	$\alpha = -1$	−1.020	0.002
	$\beta = 500$	520.823	4.462
20000	$\mu = 100$	100.261	0.202
	$\alpha = -0.5$	−0.502	0.003
	$\beta = 500$	494.628	14.255
10000	$\mu = 100$	100.170	0.241
	$\alpha = -0.5$	−0.5003	0.002
	$\beta = 500$	488.221	17.603
5000	$\mu = 100$	100.481	0.342
	$\alpha = -0.5$	−0.498	0.001
	$\beta = 500$	486.176	22.988
20000	$\mu = 100$	100.076	0.449
	$\alpha = 1$	1.009	0.013
	$\beta = 500$	498.702	13.193
10000	$\mu = 100$	100.033	0.337
	$\alpha = 1$	1.035	0.014
	$\beta = 500$	523.564	14.525
5000	$\mu = 100$	99.808	0.892
	$\alpha = 1$	1.010	0.021
	$\beta = 500$	500.517	18.356

Table 3. Mean and S.D. of the MLE calculated from 100 Monte Carlo runs.

Sample	True	Estimates	
size	parameters	mean	SD
20000	$\mu = 200$	198.649	0.206
	$\alpha = -1$	−1.021	0.0008
	$\beta = 500$	527.448	1.983
10000	$\mu = 200$	199.108	0.279
	$\alpha = -1$	−1.022	0.001
	$\beta = 500$	525.699	2.460
5000	$\mu = 200$	198.979	0.464
	$\alpha = -1$	−1.021	0.001
	$\beta = 500$	524.578	3.626
20000	$\mu = 200$	200.347	0.271
	$\alpha = -0.5$	−0.501	0.001
	$\beta = 500$	501.277	4.894
10000	$\mu = 200$	200.524	0.436
	$\alpha = -0.5$	−0.502	0.002
	$\beta = 500$	496.981	8.153
5000	$\mu = 200$	200.202	0.515
	$\alpha = -0.5$	−0.499	0.003
	$\beta = 500$	502.108	10.306
20000	$\mu = 200$	200.065	0.328
	$\alpha = 1$	0.999	0.004
	$\beta = 500$	500.495	3.636
10000	$\mu = 200$	199.868	0.527
	$\alpha = 1$	1.007	0.006
	$\beta = 500$	500.486	5.493
5000	$\mu = 200$	199.868	0.650
	$\alpha = 1$	1.007	0.008
	$\beta = 500$	503.351	7.966

6 CONCLUSION

In this work we have investigated a particular self-exciting point process. We suggest to consider this process as a renewal process and we define the interarrival times distribution that we dename *Mino distribution*. Some properties of this distribution are explored. Statistical inference is driven via the maximum likelihood approach. Results are obtained on simulation data. Further work will be conducted to develop a Bayesian approach and to consider goodness of fit test.

Appendix A: Simulation of a Mino distribution

To obtain realisation of a r.v. having a Mino distribution, we use the following well-known result: *Let F be a cumulative distribution function. Then the cdf of the r.v.* $F^{-1}(U)$ *where U is a uniform r.v. on* [0,1], *is F*. For the Mino distribution the inverse of the cdf cannot be obtained in closed-form since it is supposed to be deduced expressing x with $F(x)$ from equation (12).

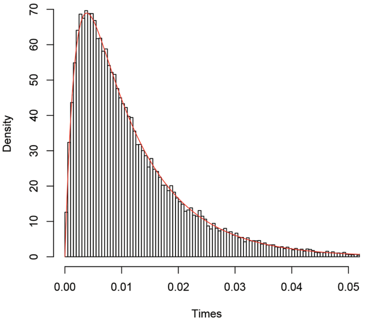


Figure 6. Histogram obtained with sample size equal to 50000 for a Mino distribution with parameters $\mu = 100$, $\alpha = -1$ and $\beta = 500$.

$$x + \frac{\alpha}{\beta}(1 - e^{-\beta x}) = \frac{1}{\mu \log(1 - F(x))} \tag{12}$$

This problem can be solved using an iterative scheme. One can suggests the following algorithm to generate realisations of a Mino distribution:

1. generate a uniform number u on [0,1]
2. Starting from an initial value $t^{(0)}$,

   while $|t^{(p+1)} - t^{(p)}| > \varepsilon(\varepsilon \to 0)$,
   compute $t^{(p+1)} = t^{(p)} - \phi(t^{(p)}) / \phi'(t^{(p)})$
   where $\phi(x) = x + \frac{\alpha}{\beta}(1 - e^{-\beta x}) - \frac{1}{\mu \log u}$
   and $\phi'(x) = 1 + \alpha e^{-\beta x}$.

The Figure 6 displays the exact pdf and the histogram of 50000 realisations obtained with the algorithm previously described, for a Mino distribution with parameters $\mu = 100$, $\alpha = -1$ and $\beta = 500$.

REFERENCES

Bowsher, C. (2007). Modeling security market events in continuous time: intensity based, multivariate point process model. *J. Econometrics 141*, 876–912.

Hawkes, A. (1971). Spectra of some self-exciting and mutually exciting point process. *Biometrika 58*, 83–90.

Johnson, D. H. (1996). Point process models of single-neuron discharges. *J. Comput. Neurosci. 3*, 275–299.

Mäkeläinen, T., K. Schimdt, & G. Styan (1981). On the existence and uniqueness of the maximum likelihood estimate of vector-valued parameter in fixed-size samples. *Ann. Statist. 9*, 758–767.

Mino, H. (2001). Parameter estimation of the intensity process of self-exciting point processes using the EMalgorithm. *IEEE Trans. Instrum. Meas. 50*(3), 658–664.

Ogata, Y. (1999). Seismicity analysis through pointprocess modelling: a review. *Pure Appl. Geophys. 155*, 471–507.

Ruggeri, F. & R. Soyer (2008). Advances in Bayesian Software Reliability Modeling. *Advances in Mathematical Modeling for Reliability, T. Bedford et al. (Eds), IOS Press*, 165–176.

Snyder, L. & I. Miller (1991). *Random Point Processes in Time and Space*. Springer.

This page intentionally left blank

System reliability analysis

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

House events matrix for application in shutdown probabilistic safety assessment

M. Čepin
Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia

ABSTRACT: The fault tree is a method for identification and assessment of combinations of the undesired events that can lead to the undesired state of the system. The objective of the work is to present a mathematical model of extended fault tree method with the house events matrix to enable integration of several system models into one model. The mathematical model of the fault tree and of its extension with the house events matrix is presented. The theory is supported by simple examples, which facilitates the understanding of the approach.

1 INTRODUCTION

The fault tree is a widely used method for evaluation of reliability and safety (Vesely, 2002, Kumamoto, Henley, 1996). It is applied in various industries and fields of application (Čepin, 1997, Čepin 2005). Its repute is gained primarily when integrated with the event tree analysis as a part of the probabilistic safety assessment related to nuclear safety (Čepin, 2002, Martorel et al. 2006) and related to aircraft safety.

1.1 *Objective*

The objective of the work is to present a mathematical model of extended fault tree method with the house events matrix to enable integration of several system models into one model.

Integration of several system models into one model may facilitate the extension of probabilistic safety assessment, which was originally made for power plant full power operation, to consideration of other modes of operation (Kiper 2002, Čepin, Mavko, 2002, Čepin, Prosen, 2008).

2 METHOD

The fault tree is a method for identification and assessment of combinations of the undesired events that can lead to the undesired state of the system, which is either a system fault or a failure of specific function of the system (Kumamoto, Henley, 1996, PRA Guide, 1982). The undesired state of the system is represented by a top event. The fault tree can be represented in graphical form or in the form of Boolean equations, which integrate the top event with logical gates and which integrate the logical gates with other logical gates and with the primary events (Ren, Dugan, 1998, Čepin, Mavko, 2002).

The primary events are the events, which are not further developed and represent the components and their failure modes or human error events. Not all the component failure modes are considered. The analysis is focused to those, which can cause failure of the system. In other words, the analysis is focused to those which can cause the top event.

The primary events can be basic events or house events. The basic events are the ultimate parts of the fault tree, which represent the undesired events on the level of the components, e.g. the component failures, the missed actuation signals, the human errors, effects of the test and maintenance activities, the common cause analysis contributions.

The house events represent the logic switches. They represent conditions set either to true or to false, which support the modelling of connections between the gates and the basic events and enable that the fault tree better represents the system operation and its environment.

The fault tree is mathematically represented by a set of boolean equations.

$$G_i = f(G_p, B_j, H_s);\ i, p \in \{1..P\}, j \in \{1..J\},\ s \in \{1..S\} \quad (1)$$

G_i – logical gate i
G_p – logical gate p
Bj – basic event j
H_s – house event s
p – number of gates
j – number of basic events
s – number of house events

The qualitative analysis is a process of Boolean reduction, where all Boolean equations are inserted one into another. Then they are rearranged into a sum of products considering the laws of Boolean algebra.

$$MCSi = \prod_{j=1}^{m} Bj \tag{2}$$

$MCSi$ – minimal cut set i
B_j – basic event j
m – number of basic events in minimal cut set i

The sum of products represents the minimal cut sets. Minimal cut sets are combinations of the smallest number of basic events, which if occur simultaneously, may lead to the top event. In other words, the minimal cut set represents the smallest set of component failures, which can cause failure of the system.

$$TOP = \sum_{z=1}^{Z} MCS_i \tag{3}$$

Z – number of minimal cut sets

House events disappear from the results equation, because their values such as 0 or 1 are used in the Boolean reduction of equations.

In theory, different house events values in a set of house events may change the model significantly.

This is the key of the idea behind the house events matrix. Namely, in probabilistic safety assessment as it was initially used, it is possible that for a single safety system, several fault trees are needed. They may differ because of different success criteria. For example, in some configuration we can rely only to one out of two system trains, if the other is in maintenance. In the other configuration, we have two system trains available. Fault trees may differ due to different boundary conditions, as they are linked to different scenarios with different requirements.

For example: auxiliary feedwater system with two motor driven pumps and one motor driven pump is available in nuclear power plant with pressurized reactor. The complete system can be considered in majority of conditions. In the conditions of station blackout, electrical power is not available and motor driven pumps are not applicable, but turbine pump and related equipment is applicable. So, the model is much more applicable, if the motor driven pumps and their related equipment are cut off from the model for the station blackout condition.

The house events matrix is introduced to list the house events and their respective values for the complete set of conditions that may appear in the analysis.

The house events matrix represents the values of house events for all conditions of the analysis in one dimension (all columns of the matrix) and for all house events in the other dimension (all rows of the matrix).

The house events matrix identifies values of house events for all house event names related to the system analysis in its rows and for all the conditions in its columns.

Values, which are either 0 or 1 (or false or true) are assigned to each house event for all conditions in the matrix.

The quantification of a fault tree equipped with house events matrix mathematically is similar to the fault tree without house events. The difference is in a number of results sets. One fault tree without house events has one set of results. Fault tree with house events matrix has so many sets of results as it is the number of different house events matrix columns.

Fault tree quantification equation is performed using the following equation.

$$\begin{aligned} P_{TOP} = \sum_{z=1}^{Z} P_{MCSi} - \sum_{i<j} P_{MCSi \cap MCSj} + \\ + \sum_{i<j<k} P_{MCSi \cap MCSj \cap MCSk} \\ - \ldots + (-1)^{n-1} P \bigcap_{i=1}^{n} MCSi \end{aligned} \tag{4}$$

P_{TOP} – probability of a top event (failure probability of a system or function, which is defined in a top event)
P_{MCSi} – probability of minimal cut set i
Z – number of minimal cut sets
n – number of basic events in the largest minimal cut set (related to number of basic events representing minimal cut set)

The probabilities of minimal cut sets should be calculated considering their possible dependency.

$$\begin{aligned} P_{MCSi} = \\ P_{B1} \cdot P_{B2} | P_{B1} \cdot P_{B3} | P_{B1} \cap P_{B2} \cdot \ldots \\ \cdot P_{Bm} | P_{B1} \cap P_{B2} \cap \ldots \cap P_{Bm-1} \end{aligned} \tag{5}$$

Under assumption that the basic events sets are mutually independent, the equation is simplified.

$$P_{MCSi} = \prod_{j=1}^{m} P_{Bj} \tag{6}$$

P_{MCSi} – probability of minimal cut set i
P_{Bj} – probability of basic event B_j

m – number of basic events in minimal cut set i

Probability of basic event depends on the nature of functioning of component modeled in the basic event and its respective failure mode.

Probability model is selected in relation to the failure mode of the component in the basic event. Parameters of probabilistic model are obtained from data base.

Generic data base can be used, but site specific data base for particular equipment is much more suitable.

2.1 *Application of house events matrix*

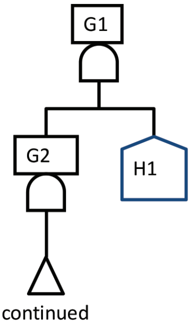
Figure 1 shows the simplest application of the house event under AND gate G1. Logical gates are represented by rectangles. House events are represented by pentagons. Logical gate defines the relation between input events to the upper event and the upper event. Triangle is used for continuation.

If the value of the house event H1 is 0, then equipment modeled in gate G2 does not apply in the model. The result of G1 is empty set, because G1 happens if H1 and G2 both happen.

If the value of the house event H1 is 1, then equipment modeled in gate G2 represents the complete event G1.

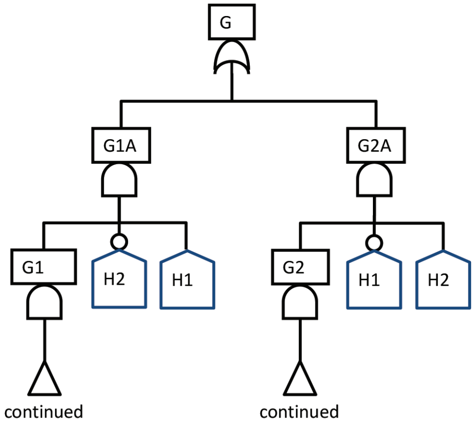
Figure 2 shows example fault tree G. It integrates two variants G1 and G2 of continued fault tree with house events H1 and H2. Gates G1 A and G2 A are in between. Circle above house event represent negation of this house event.

If house event H1 is switched on, which means its value is 1, and house event H2 is switched off, which means its value is 0 (and value of its negation is 1), the gate G1 propagates to G1 A and consequently to gate G. At the same time the gate G2 does not propagate up, because it is switched off by house event H2 set to 0 and by negation of house event H1 (which is 0).



G1 = G2 AND H1
if H1 = 0 then
G1 = 0
if H1 =1 then
G1 = G2

Figure 1. Example of house event under AND gate.



G = G1A OR G2A
G1A = G1 AND NOT H2 AND H1
G2A = G2 AND NOT H1 AND H2

Figure 2. Example of house event under AND gate.

If house event H2 is switched on, which means its value is 1, and house event H1 is switched off, which means its value is 0 (and value of its negation is 1), the gate G2 propagates to G2 A and consequently to gate G. At the same time the gate G1 does not propagate up, because it is switched off by house event H1 set to 0 and by negation of house event H2 (which is 0).

Figure 2 shows one fault tree G, where two variants G1 and G2 are represented in one model. At the same time it is assured that wrong combination of values of house events results in empty set of fault tree analysis which can represent an alarm that the model should be checked. Namely, both house events set to 1 or both set to 0 would give empty set of results.

Figure 3 shows a fault tree example for auxiliary feedwater system, where 5 system configurations are joined in a single fault tree.

2.2 *Application of house events matrix for shutdown probabilistic safety assessment*

House events matrix can be increasingly important in shutdown probabilistic safety assessment, which is an important issue considered in the last years (Čepin, Prosen, 2008, NUREG/CR-6144, 1995, Papazouglou, 1998, Swaminathan, Smidts, 1999).

Probabilistic safety assessment of a nuclear power plant deals with a number of safety systems, large number of components and is complex. Its complexity increases significantly when other than power operation modes are considered.

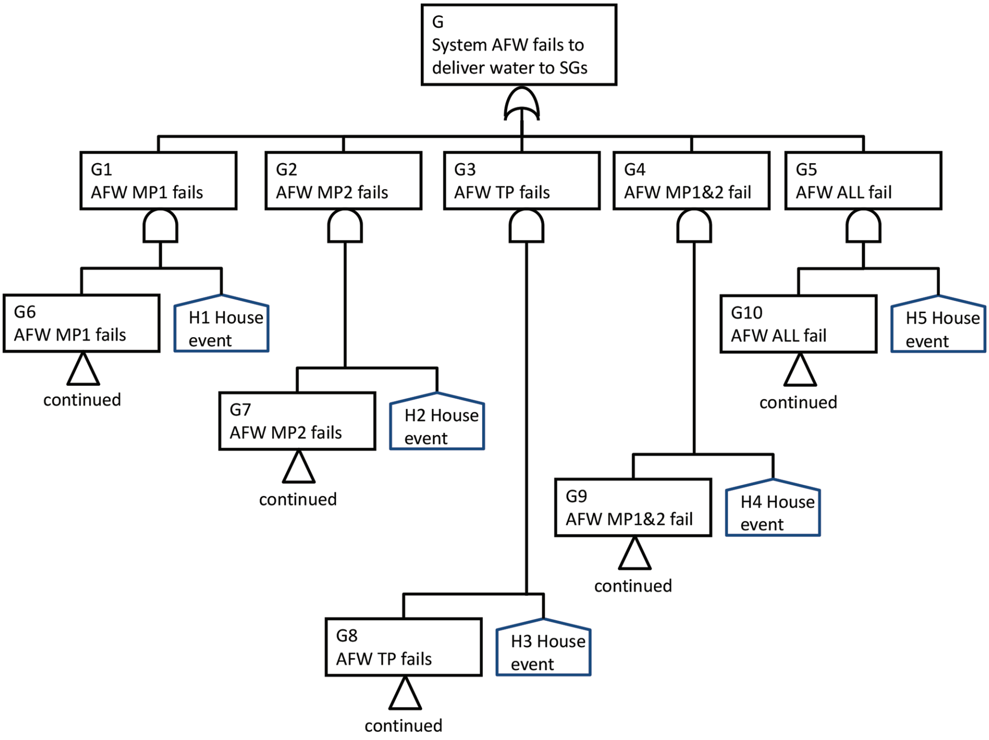


Figure 3. Fault tree example for auxiliary feedwater system.

Table 1. House event matrix for 4 modes of operation of a nuclear power plant for human failure events included in the fault trees of safety systems.

		Mode 1	Mode 2	Mode 3	Mode 4
		Power Operation	Startup	Hot Standby	Hot Shutdown
Event Group	Event Identification	POS-M1-G1	POS-M2-G1	POS-M3-G1	POS-M4-G1
Human Failure Events	HFE01	1	1	0	0
	HFE01A	0	0	1	1
	HFE02	1	1	0	0
	HFE02A	0	0	1	1
	HFE03	1	1	0	0
	HFE03A	0	0	1	1
	HFE04	1	1	0	0
	HFE04A	0	0	1	1
	HFE05	1	1	0	0
	HFE05A	0	0	1	1
	HFE06	1	1	0	0
	HFE06A	0	0	1	1
	HFE07	1	1	0	0
	HFE07A	0	0	1	1
	HFE08	1	1	0	0
	HFE08A	0	0	1	1
	HFE09	1	1	0	0
	HFE09A	0	0	1	1

Table 1 shows house event matrix for 4 modes of operation of a nuclear power plant for human failure events included in the fault trees of safety systems.

It includes 9 human failure events, which appear in different fault trees or event trees and which need a change on their respective human error probabilities due to different plant conditions. The human error probability should change in the model due to several reasons. One of them is the time available for action, which is larger in shutdown, so probability is smaller in such conditions.

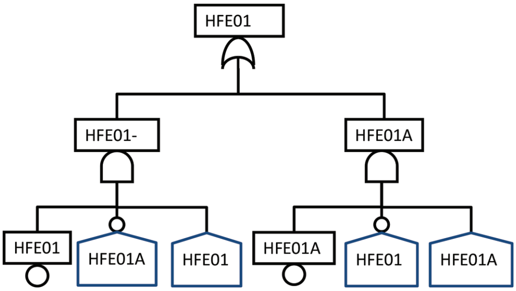


Figure 4. Example of fault tree replacing basic event.

Figure 4 shows one of fault tree portions, which were introduced to replace a single event HFE01 in original model for power operation with a fault tree HFE01 including 2 house events representing 2 basic events, which replace basic event HFE01.

Basic event HFE01 remains in the new fault tree for the power operation, where house event with the same event is needed for its activation.

Basic event HFE01 A represents an addition to the model together with both house events, and is applicable for plant shutdown mode and includes smaller human error probability then HFE01.

Gate HFE01 is split to gate HFE01- and gate HFE01 A in order to keep the naming scheme to keep the transparency.

Name HFE01- is selected because gates have to be uniquely named and name HFE01 exists in the fault tree as an upper event. HFE01 represents operator failure to establish emergency boration if automatic boration with normally considered path fails.

Table 2 shows house events matrix for 4 modes of operation of a nuclear power plant for initiating events only.

Table 2. House events matrix for 4 modes of operation of a nuclear power plant for initiating events only.

		Mode 1	Mode 2	Mode 3	Mode 4
		Power Operation	Startup	Hot Standby	Hot Shutdown
Event Group	Event Identification	POS-M1-G1	POS-M2-G1	POS-M3-G1	POS-M4-G1
Initiating Events	ATWS-	1	1	0	0
	CCWS-	1	1	1	1
	ESWS-	1	1	1	1
	IAIR-	1	1	1	1
	ISLO-	1	1	1	0
	ISLO1	0	0	0	1
	LDC--	1	1	1	1
	LLOC-	1	1	1	0
	LLOC1	0	0	0	1
	LOSP-	1	1	1	1
	MLOC-	1	1	1	0
	MLOC1	0	0	0	1
	SBO--	1	1	1	1
	SGTR-	1	1	0	0
	SGTR1	0	0	1	0
	SLB--	1	1	0	1
	SLOC-	1	1	1	0
	SLOC1	0	0	0	1
	TRMF-	1	1	0	1
	TR---	1	1	0	1
	VESF-	1	1	1	0
	VESF1	0	0	0	1

3 CONCLUSIONS

The objective of the work was to present a mathematical model of extended fault tree method with the house events matrix to enable integration of several system models into one model, which is done.

The mathematical model of the fault tree and of its extension with the house events matrix is presented. The theory is supported by simple examples, which facilitates the understanding of the approach.

Furthermore, the theory is supported by realistic examples from probabilistic safety assessment. The deficiency of the approach is the software support, because existing probabilistic safety assessment models are extremely complex and if software platform for evaluation does not support consideration of house events, the approach is not practical.

If the house events are supported, their use and the application of the house events matrix may significantly contribute to reduce complexity of the models in case that they are expanded with consideration of other modes than full power operation.

REFERENCES

Čepin M., B. Mavko, 1997, Probabilistic Safety Assessment Improves Surveillance Requirements in Technical Specifications, *Reliability Engineering and Systems Safety,* 56, 69–77.

Čepin M., B. Mavko, 2002, A Dynamic Fault Tree, *Reliability Engineering and System Safety*, Vol. 75, No. 1, pp. 83–91.

Čepin M., 2002, Optimization of Safety Equipment Outages Improves Safety, *Reliability Engineering and System Safety*, 77,71–80.

Čepin M., 2005, Analysis of Truncation Limit in Probabilistic Safety Assessment, *Reliability Engineering and System Safety, 87, 395–403.*

Čepin M., R. Prosen, 2008, Probabilistic Safety Assessment for Hot Standby and Hot Shutdown, Proceedings of Nuclear Energy for New Europe.

Kiper K., 2002, Insights from an All-Modes PSA at Seabrook Station, Proceedings of PSA 2002, p. 429–434.

Kumamoto H., E.J. Henley, 1996, *Probabilistic Risk Assessment and Management for Engineers and Scientists*, IEEE Press, New York.

Martorell, S., Carlos, S., Villanueva, J.F., Sánchez, A.I., Galvan, B., Salazar, D., Čepin, M., 2006, Use of Multiple Objective Evolutionary Algorithms in Optimizing Surveillance Requirements, *Reliability Engineering and System Safety,* 91 (9), 1027–1038.

NUREG/CR-6144, 1995, Evaluation of Potential Severe Accidents During Low Power and Shutdown Operations at Surry Unit 1, US NRC.

Papazoglou I.A., 1998, Mathematical Foundations of Event Trees, *Reliability Engineering and System Safety*, 61, 169–183.

PRA Guide, 1982, *Probabilistic Risk Assessment Procedures Guide*, NUREG/CR-2300, Vol. 1,2, US NRC, Washington DC.

Ren Y., J.B. Dugan, 1998, Optimal Design of Reliable Systems Using Static and Dynamic Fault Trees, IEEE *Transactions on Reliability*, 234–244.

Swaminathan S., C. Smidts, 1999, The Mathematical Formulation for the Event Sequence Diagram Framework, *Reliability Engineering and System Safety*, 65, 103–118.

Vesely W., J. Dugan, J. Fragola, J. Minarick, J. Railsback, 2002, *Fault Tree Handbook with Aerospace Applications, National Aeronautics and Space Administration*, NASA.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Imprecise system reliability using the survival signature

Frank P.A. Coolen
Department of Mathematical Sciences, Durham University, Durham, UK

Tahani Coolen-Maturi
Durham University Business School, Durham University, Durham, UK

Louis J.M. Aslett
Department of Statistics, Oxford University, Oxford, UK

Gero Walter
School of Industrial Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

ABSTRACT: The survival signature has been introduced to simplify quantification of reliability of systems which consist of components of different types, with multiple components of at least one of these types. The survival signature generalizes the system signature, which has attracted much interest in the theoretical reliability literature but has limited practical value as it can only be used for systems with a single type of components. The key property for uncertainty quantification of the survival signature, in line with the signature, is full separation of aspects of the system structure and failure times of the system components. This is particularly useful for statistical inference on the system reliability based on component failure times.

This paper provides a brief overview of the survival signature and its use for statistical inference for system reliability. We show the application of generalized Bayesian methods and nonparametric predictive inference, both these inference methods use imprecise probabilities to quantify uncertainty, where imprecision reflects the amount of information available. The paper ends with a discussion of related research challenges.

1 INTRODUCTION

In mathematical theory of reliability the main focus is on the functioning of a system given the functioning, or not, of its components and the structure of the system. The mathematical concept which is central to this theory is the *structure function* (Barlow & Proschan 1975). For a system with m components, the state vector is $\underline{x} = (x_1, x_2, \ldots, x_m) \in \{0,1\}^m$, with $x_i = 1$ if the i th component functions and $x_i = 0$ if not. The labelling of the components is arbitrary but must be fixed to define $\underline{x}$. The structure function $\phi: \{0,1\}^m \rightarrow \{0,1\}$, defined for all possible $\underline{x}$, takes the value 1 if the system functions and 0 if the system does not function for state vector $\underline{x}$. Most practical systems are coherent, which means that $\phi(\underline{x})$ is not decreasing in any of the components of $\underline{x}$, so system functioning cannot be improved by worse performance of one or more of its components. The assumption of coherent systems is made throughout this paper and is convenient from the perspective of uncertainty quantification for system reliability. It is further logical to assume that $\phi(\underline{0}) = 0$ and $\phi(\underline{1}) = 1$, so the system fails if all its components fail and it functions if all its components function.

For larger systems, working with the full structure function may be complicated, and one may particularly only need a summary of the structure function in case the system has exchangeable components of one or more types. We use the term ‘exchangeable components’ to indicate that the failure times of the components in the system are exchangeable (De Finetti 1974). Recently, we introduced such a summary, called the *survival signature*, to facilitate reliability analyses for systems with multiple types of components (Coolen & Coolen-Maturi 2012). In case of just a single type of components, the survival signature is closely related to the system signature (Samaniego 2007), which is well-established and the topic of many research papers during the last decade. However, generalization of the signature to systems with multiple types of components is extremely complicated (as it involves ordering order statistics of different distributions), so much so that it cannot be applied to most practical systems. In addition to the possible use for such systems, where the benefit only occurs if there are multiple components of the same types, the survival signature is arguably also easier to interpret than the signature.

Consider a system with $K \geq 1$ types of components, with m_k components of type $k \in \{1,\ldots,K\}$ and $\sum_{k=1}^{K} m_k = m$. Assume that the random failure times of components of the same type are exchangeable (De Finetti 1974). Due to the arbitrary ordering of the components in the state vector, components of the same type can be grouped together, leading to a state vector that can be written as $\underline{x} = (\underline{x}^1, \underline{x}^2, \ldots, \underline{x}^K)$, with $\underline{x}^k = (x_1^k, x_2^k, \ldots, x_{m_k}^k)$ the sub-vector representing the states of the components of type *k*.

The *survival signature* for such a system, denoted by $\Phi(l_1,\ldots,l_K)$, with $l_k = 0,1,\ldots,m_k$ for $k = 1,\ldots,K$, is defined as the probability for the event that the system functions given that *precisely* l_k of its m_k components of type *k* function, for each $k \in \{1,\ldots,K\}$ (Coolen & Coolen-Maturi 2012). There are $\binom{m_k}{l_k}$ state vectors $\underline{x}^k$ with $\sum_{i=1}^{m_k} x_i^k = l_k$. Let $S_{l_k}^k$ denote the set of these state vectors for components of type *k* and let $S_{l_1,\ldots,l_K}$ denote the set of all state vectors for the whole system for which $\sum_{i=1}^{m_k} x_i^k = l_k, k = 1,\ldots,K$. We also introduce the notation $\underline{l} = (l_1,\ldots,l_K)$. Due to the exchangeability assumption for the failure times of the m_k components of type *k*, all the state vectors $\underline{x}^k \in S_{l_k}^k$ are equally likely to occur, hence (Coolen & Coolen-Maturi 2012)

$$\Phi(\underline{l}) = \left[\prod_{k=1}^{K} \binom{m_k}{l_k}\right]^{-1} \times \sum_{\underline{x} \in S_{l_1,\ldots,l_K}} \phi(\underline{x})$$

Let $C_t^k \in \{0,1,\ldots,m_k\}$ denote the number of components of type *k* in the system that function at time $t > 0$. Then, for system failure time T_S ,

$$P(T_S > t) = \sum_{l_1=0}^{m_1} \cdots \sum_{l_K=0}^{m_K} \Phi(\underline{l}) P\left(\bigcap_{k=1}^{K} \{C_t^k = l_k\}\right)$$

There are no restrictions on dependence of the failure times of components of different types, as the probability $P(\bigcap_{k=1}^{K} \{C_t^k = l_k\})$ can take any form of dependence into account, for example one can include common-cause failures quite straightforwardly into this approach (Coolen & Coolen-Maturi 2015). However, there is a substantial simplification if one assumes that the failure times of components of different types are independent, and even more so if one assumes that the failure times of components of type *k* are conditionally independent and identically distributed with CDF $F_k(t)$. With these assumptions, we get

$$P(T_S > t) = \sum_{l_1=0}^{m_1} \cdots \sum_{l_K=0}^{m_K} \Phi(\underline{l}) \times \prod_{k=1}^{K} \binom{m_k}{l_k} [F_k(t)]^{m_k - l_k} [1 - F_k(t)]^{l_k} \tag{1}$$

The main advantage of the survival signature, in line with this property of the signature for systems with a single type of components (Samaniego 2007), is that the information about the system structure is fully separated from the information about functioning of the components, which simplifies related statistical inference as well as considerations of optimal system design. In particular for study of system reliability over time, with the structure of the system, and hence the survival signature, not changing, this separation also enables relatively straightforward statistical inferences where even the use of imprecise probabilistic methods (Augustin, Coolen, de Cooman, & Troffaes 2014, Coolen & Utkin 2011) is quite straightforward. Such methods have the advantage that imprecision for the system survival function reflects the amount of information available. The next two sections briefly discuss such methods of statistical inference for the system failure time. First we show an application of generalized Bayesian methods, with a set of prior distributions instead of a single prior distribution. This is followed by a brief discussion and application of nonparametric predictive inference (Coolen 2011), a frequentist statistical method which is based on relatively few assumptions, enabled through the use of imprecise probabilities, and which does not require the use of prior distributions. The paper ends with a brief discussion of research challenges, particularly with regard to upscaling the survival signature methodology for application to large-scale real-world systems and networks.

2 IMPRECISE BAYESIAN INFERENCE

The reliability of a system, for which the survival signature is available, is quite straightforwardly quantified through its survival function, as shown in the previous section. We briefly consider a scenario where we have test data that enable learning about the reliability of the components of different types in the system, where we assume independence of the failure times of components of different types. The numbers of components in the system, of each type, that are functioning at time *t*, denoted by C_t^k for $k = 1,\ldots,K$, are the random quantities of main interest. One attractive statistical method to learn about these random quantities from test data is provided by the Bayesian framework of statistics, which can be applied with the assumption of a parametric distribution for the component failure times (Walter, Graham, & Coolen 2015) or in a nonparametric manner (Aslett, Coolen, & Wilson 2015). We briefly illustrate the latter approach.

Assume that there are m_k components of type k in the system, and we are interested in the probability distribution of C_t^k. Suppose that n_k components of the same type k were tested, these are not the components that are in the system but their failure times are assumed to be exchangeable with those in the system. We assume that for all tested components the failure time has been observed, let s_t^k denote the number of these components that still functioned at time t. A convenient and basic model for C_t^k is the Binomial distribution, where the probability of 'success', that is a component still to be functioning at time t, can, in the Bayesian framework, be conveniently modelled as a random quantity with a Beta prior distribution. Different to the standard parameterization for the Beta distribution, we define a Beta prior distribution through parameters $n^{(0)}_{k,t}$ and $y^{(0)}_{k,t}$ with as interpretations a pseudocount of components and the expected value of the success probability, respectively. Hence, these parameters can be interpreted in the sense that the prior distribution represents beliefs reflecting the same information as would result from observing $n^{(0)}_{k,t}$ components of which $n^{(0)}_{k,t} y^{(0)}_{k,t}$ still function at time t (Walter 2013). Doing this leads to straightforward updating, using the test information consisting of observations of n_k components of which $s_{k,t}$ were still functioning at time t. The updating results in a similar Beta distribution as the prior, but now with parameter values $n^{(n)}_{k,t} = n^{(0)}_{k,t} + n_k$ and $y^{(n)}_{k,t}$ the weighted average of $y^{(0)}_{k,t}$ and $s_{k,t}/n_k$, with weights proportional to $n^{(0)}_{k,t}$ and n_k, respectively. This leads to the posterior predictive distribution (Walter, Aslett, & Coolen 2016)

$$P(C_t^k = l_k \mid s_t^k) = \binom{m_k}{l_k} \times \frac{B(l_k + n^{(n)}_{k,t} y^{(n)}_{k,t}, m_k - l_k + n^{(n)}_{k,t}(1 - y^{(n)}_{k,t}))}{B(n^{(n)}_{k,t} y^{(n)}_{k,t}, n^{(n)}_{k,t}(1 - y^{(n)}_{k,t}))}$$

This model can also relatively straightforwardly be used with a set of Beta prior distributions rather than a single one, a generalization fitting in the theory of imprecise probability (Augustin, Coolen, de Cooman, & Troffaes 2014). At each value of t one calculates the infimum and supremum of the probability $P(C_t^k = l_k \mid s_t^k)$ over the set of prior parameters, with $n^{(0)}_{k,t} \in \left[\underline{n}^{(0)}_{k,t}, \overline{n}^{(0)}_{k,t}\right]$ and $y^{(0)}_{k,t} \in \left[\underline{y}^{(0)}_{k,t}, \overline{y}^{(0)}_{k,t}\right]$, with the bounds of these intervals chosen to reflect a priori available knowledge and its limitations. The use of such prior sets, with only an interval of possible values specified for each parameter, provides much flexibility for modelling prior beliefs and indeterminacy, together with interesting ways in which the corresponding sets of posterior (predictive) distributions and related inferences can vary. Most noticeably, this model enables conflict between prior beliefs and data to be shown through increased imprecision, that is difference between upper and lower probabilities for an event of interest (Walter 2013). We illustrate the use of this model, together with the survival signature, for a small system in Example 1, without attention to such prior-data conflict, further details on this will be presented elsewhere (Walter, Aslett, & Coolen 2016).

Example 1

As a small example, consider the system with three types of components presented in Figure 1. The survival signature of this system is given in Table 1, where all cases with $l_3 = 0$ have been omitted as the system cannot function if the component of Type 3 does not function, hence $\Phi(l_1, l_2, 0) = 0$ for all (l_1, l_2).

For component types 1 and 2, we consider a near-noninformative set of prior survival functions. For components of type 3, we consider an informative set of prior survival functions as given in Table 2. This set could result from eliciting prior survival probabilities at times $t = 0,1,2,3,4,5$ only, and using those values to deduce such prior probabilities for all other values of t without further assumptions. These prior assumptions, together with sets of posterior survival functions, are illustrated in Figure 3 (presented at the end of the paper); test data for components of type 1 and 2 are taken as $\{2.2, 2.4, 2.6, 2.8\}$ and $\{3.2, 3.4, 3.6, 3.8\}$, respectively. For components of type 3 test data are taken as $\{0.5, 1.5, 2.5, 3.5\}$, which are well in line with expectations according to the set of prior distributions. The posterior sets of survival functions for each component type and for the whole system show considerably smaller imprecision than the corresponding prior sets, which is mainly due to the low prior

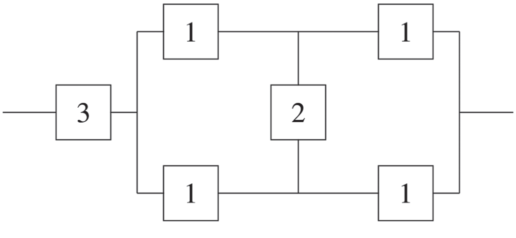


Figure 1. System with 3 types of components.

Table 1. Survival signature for the system in Figure 1 for cases with $l_3 = 1$.

l_1	l_2	$\Phi(l_1, l_2, 1)$	l_1	l_2	$\Phi(l_1, l_2, 1)$
0	0	0	0	1	0
1	0	0	1	1	0
2	0	1/3	2	1	2/3
3	0	1	3	1	1
4	0	1	4	1	1

Table 2. Lower and upper prior functioning probability bounds for component type 3 in the system of Figure 1.

t	(0, 1)	(1, 2)	(2, 3)	(3, 4)	(4, 5)
$\underline{y}_{3,t}^{(0)}$	0.625	0.375	0.250	0.125	0.010
$\overline{y}_{3,t}^{(0)}$	0.999	0.875	0.500	0.375	0.250

strength intervals we chose for this example, namely $\left[\underline{n}_{1,t}^{(0)}, \overline{n}_{1,t}^{(0)}\right] = \left[\underline{n}_{2,t}^{(0)}, \overline{n}_{2,t}^{(0)}\right] = [1,2], \left[\underline{n}_{3,t}^{(0)}, \overline{n}_{3,t}^{(0)}\right] = [1,4]$ for all t. We see that posterior lower and upper survival functions drop at those times t when there is a failure time in the test data, or a drop in the prior survival probability bounds. Note that the lower bound for prior system survival function is zero for all t due to the prior lower bound of zero for type 1 components, and for the system to function at least two components of type 1 must function. A further reason why the imprecision reduces substantially in this example is that the data do not conflict with the prior beliefs. With these sets of prior distributions such prior-data conflict can only really occur for components of type 3, as such conflict logically requires at least reasonably strong prior beliefs to be taken into account through the set of prior distributions. If test failure times for the components of type 3 were unexpectedly small or large, the imprecision in the lower and upper posterior survival functions for this component would increase, with a similar effect on the corresponding overall lower and upper system survival functions. A detailed analysis illustrating this effect will be presented elsewhere (Walter, Aslett, & Coolen 2016).

3 NONPARAMETRIC PREDICTIVE INFERENCE

Nonparametric Predictive Inference (NPI) (Coolen 2011) is a frequentist statistical framework based on relatively few assumptions and considering events of interest which are explicitly in terms of one or more future observations. NPI can be considered suitable if there is hardly any knowledge about the random quantity of interest, other than the data which we assume consist of n observations, or if one does not want to use such further information, e.g. to study effects of additional assumptions underlying other statistical methods. NPI uses lower and upper probabilities, also known as imprecise probabilities, to quantify uncertainty (Augustin, Coolen, de Cooman, & Troffaes 2014) and has strong consistency properties from frequentist statistics perspective (Augustin & Coolen 2004, Coolen 2011). NPI provides a solution to some explicit goals formulated for objective (Bayesian) inference, which cannot be obtained when using precise probabilities (Coolen 2006), and it never leads to results that are in conflict with inferences based on empirical probabilities. NPI for system survival functions, using the survival signature, was recently presented (Coolen, Coolen-Maturi, & Al-nefaiee 2014) and is briefly summarized here.

We now present NPI lower and upper survival functions for the failure time T_S of a system consisting of multiple types of components, using the system signature combined with NPI for Bernoulli data (Coolen 1998). This enables the NPI method to be applied to, in principle, all systems. The failure times of components of different types are assumed to be independent. It must be emphasized that the NPI framework does not assume an underlying population distribution in relation to random quantities, and therefore also not that these are conditionally independent given some probability distribution. In fact, NPI explicitly takes the inter-dependence of multiple future observations into account. This requires a somewhat different approach for dealing with imprecise probabilities to that presented for the imprecise Bayesian approach in the previous section.

NPI will be used for learning about the components of a specific type in the system, from data consisting of failure times for components that are exchangeable with these. We assume therefore that such data are available, for example resulting from testing or previous use of such components. It is assumed that failure times are available for all tested components. As in the previous section, let n_k, for $k \in \{1,\ldots,K\}$, denote the number of components of type k for which test failure data are available, and let s_t^k denote the number of these components which still function at time t.

The NPI lower survival function is derived as follows. Remember that C_t^k denotes the number of components of type k in the system which function at time t, where it is assumed that failure ends the functioning of a component. Under the assumptions for the NPI approach (Coolen 1998), we derive the following lower bound for the survival function

$$P(T_S > t) \ge \sum_{l_1=0}^{m_1} \cdots \sum_{l_K=0}^{m_K} \Phi(\underline{l}) \prod_{k=1}^{K} \overline{D}(C_t^k = l_k)$$

where

$$\overline{D}(C_t^k = l_k) = \overline{P}(C_t^k \le l_k) - \overline{P}(C_t^k \le l_k - 1) = \binom{n_k + m_k}{n_k}^{-1} \binom{s_t^k - 1 + l_k}{s_t^k - 1} \times \binom{n_k - s_t^k + m_k - l_k}{n_k - s_t^k}$$

In this expression, $\overline{P}$ denotes the NPI upper probability for Bernoulli data (Coolen 1998). For each component type *k*, the function $\overline{D}$ ensures that maximum possible probability, corresponding to NPI for Bernoulli data (Coolen 1998), is assigned to the event $C_t^k = 0$, so $\overline{D}(C_t^k = 0) = \overline{P}(C_t^k = 0)$. Then, $\overline{D}(C_t^k = 1)$ is defined by putting the maximum possible remaining probability mass, from the total probability mass available for the event $C_t^k \le 1$, to the event $C_t^k = 1$. This is achieved by $\overline{D}(C_t^k = 1) = \overline{P}(C_t^k \le 1) - \overline{P}(C_t^k = 0)$. This argument is continued, by assigning for increasing l_k the maximum possible remaining probability mass $\overline{D}(C_t^k = l_k)$. As the survival signature is increasing in l_k for coherent systems, as assumed in this paper, and the resulting $\overline{D}$ is a precise probability distribution, the right-hand side of the inequality above is indeed a lower bound and it is the maximum possible lower bound. As such, it is the NPI lower probability for the event $T_S > t$, giving the NPI lower survival function for the system failure time (for $t > 0$)

$$\underline{P}(T_S > t) = \sum_{l_1=0}^{m_1} \cdots \sum_{l_K=0}^{m_K} \Phi(\underline{l}) \prod_{k=1}^{K} \overline{D}(C_t^k = l_k)$$

The corresponding NPI upper survival function for T_S is similarly derived, using the upper bound

$$P(T_S > t) \le \sum_{l_1=0}^{m_1} \cdots \sum_{l_K=0}^{m_K} \Phi(\underline{l}) \prod_{k=1}^{K} \underline{D}(C_t^k = l_k)$$

where

$$\underline{D}(C_t^k = l_k) = \underline{P}(C_t^k \le l_k) - \underline{P}(C_t^k \le l_k - 1) = \binom{n_k + m_k}{n_k}^{-1} \binom{s_t^k + l_k}{s_t^k} \times \binom{n_k - s_t^k + m_k - l_k - 1}{n_k - s_t^k}$$

In this expression, $\underline{P}$ denotes the NPI lower probability for Bernoulli data (Coolen 1998). This construction ensures that minimum possible weight is given to small values of C_t^k, resulting in the NPI upper survival function for the system failure time

$$\overline{P}(T_S > t) = \sum_{l_1=0}^{m_1} \cdots \sum_{l_K=0}^{m_K} \Phi(\underline{l}) \prod_{k=1}^{K} \underline{D}(C_t^k = l_k)$$

We illustrate this NPI method for system reliability using the survival signature in Example 2 (Coolen, Coolen-Maturi, & Al-nefaiee 2014).

Example 2

Consider the system with $K = 2$ types of components as presented in Figure 2. The survival signature for this system is presented in Table 3, it is easily verified by checking all possible combinations of the specific components of each type which function or not.

To illustrate NPI for the system survival time, suppose that $n_1 = 2$ components exchangeable with those of type 1 and $n_2 = 2$ components exchangeable with those of type 2 were tested. First suppose that failure times $t_1^2 < t_1^1 < t_2^2 < t_2^1$ were observed, with t_j^k the *j*-th ordered failure time of a component of type *k*. The resulting NPI lower and upper survival functions for the system failure time T_S are specified in Table 4, together with the results

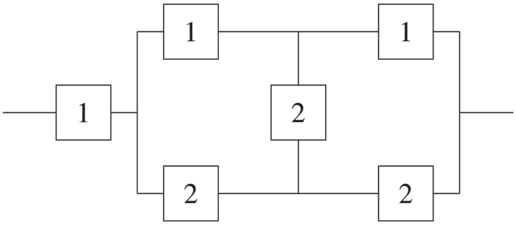


Figure 2. System with 2 types of components.

Table 3. Survival signature of the system in Figure 2.

l_1	l_2	$\Phi(l_1, l_2)$	l_1	l_2	$\Phi(l_1, l_2)$
0	0	0	2	0	0
0	1	0	2	1	0
0	2	0	2	2	4/9
0	3	0	2	3	6/9
1	0	0	3	0	1
1	1	0	3	1	1
1	2	1/9	3	2	1
1	3	3/9	3	3	1

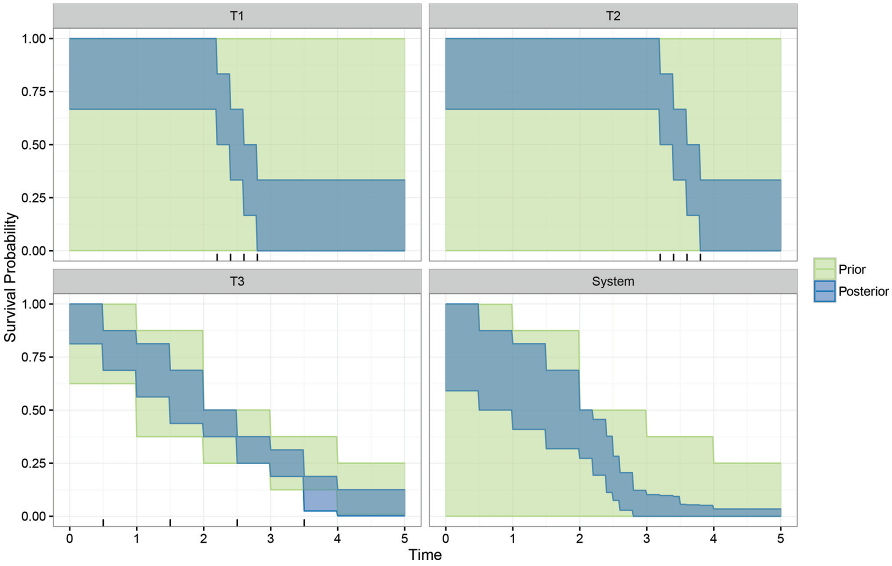


Figure 3. Prior and posterior sets of survival functions for the system in Figure 1 and its three component types. The component failure times, that form the test data, are denoted with tick marks near the time axis.

Table 4. Lower and upper survival functions for the system in Figure 2 and two data orderings.

	$t_1^2 < t_1^1 < t_2^2 < t_2^1$	
$\underline{P}(T_s > t)$	$\underline{P}(T_s > t)$	$\overline{P}(T_S > t)$
$(0, t_1^2)$	0.553	1
(t_1^2, t_1^1)	0.458	1
(t_1^1, t_2^2)	0.148	0.553
(t_2^2, t_2^1)	0.100	0.458
(t_2^1, ∞)	0	0.148
	$t_1^1 < t_1^2 < t_2^1 < t_2^2$	
$t \in$	$\underline{P}(T_S > t)$	$\overline{P}(T_S > t)$
$(0, t_1^1)$	0.553	1
(t_1^1, t_1^2)	0.230	0.667
(t_1^2, t_2^1)	0.148	0.553
(t_2^1, t_2^2)	0	0.230
(t_2^2, ∞)	0	0.148

for the case with the test failure times ordered as $t_1^1 < t_1^2 < t_2^1 < t_2^2$.

For the ordering $t_1^2 < t_1^1 < t_2^2 < t_2^1$, in the first interval in Table 4 we have not yet seen a failure in the test data, so the NPI upper probability that the system will function is equal to one, which is logical as we base the inferences on the data with few additional assumptions. In the second interval, one failure of type 2 has occurred but we do not have any evidence from the data against the possibility that a component of type 1 will certainly function at times in this interval, so the NPI upper survival function remains one. In the fourth interval, both type 2 components have failed but only one component of type 1 has failed. In this interval, to consider the lower survival function the system is effectively reduced to a series system consisting of three components of type 1, with one 'success' and one 'failure' as data, denoted by (2, 1). As such a series system only functions if all three components function, the NPI lower survival function within this fourth interval is equal to $\underline{S}_{T_S}(t) = \frac{1}{3} \times \frac{2}{4} \times \frac{3}{5} = 0.100$, which follows by sequential reasoning, using that, based on n observations consisting of s successes and $n - s$ failures, denoted as data (n, s), the NPI lower probability for the next observation to be a success is equal to $s/(n+1)$ (Coolen 1998). The NPI lower probability for the first component to function, given test data (2,1), is equal to 1/3. Then the second component is considered, conditional on the first component functioning, which combines with the test data to two out of three components observed (or assumed) to be functioning, so combined data (3,2), hence this second component will also function with NPI lower probability 2/4. Similarly, the NPI lower probability for the third component to function, conditional on functioning of the first two components in the system, so with combined data (4,3), is equal to 3/5. In the final interval, we are beyond the failure times of all the tested components, so we no longer have evidence in favour of the system to function, so $\underline{S}_{T_S}(t) = 0$, but the system might of course still function as reflected by $\overline{S}_{T_S}(t) = 0.148$.

For the second case in Table 4, with data ordering $t_1^1 < t_1^2 < t_2^1 < t_2^2$, we have $\overline{S}_{T_S}(t) = 0.667$ in the second interval, where one failure of type 1 has occurred in the test data. In the fourth interval, both tested components of type 1 have failed, leading to $\underline{S}_{T_S}(t) = 0$. Both of these values are directly related to the required functioning of the left-most component in Figure 2.

4 DISCUSSION

The survival signature is a powerful and quite basic concept. As such, further generalizations are conceptually easy, for example one can straightforwardly generalize the survival signature to multi-state systems such that it again summarizes the structure function in a manner that is sufficient for a range of uncertainty quantifications for the system reliability. The survival signature can also be used with a generalization of the system structure function where the latter is a probability instead of a binary function, or even an imprecise probability. This enables uncertainty of system functioning for given states of its components to be taken into account, which may be convenient, for example, to take uncertain demands or environments for the system into consideration. In this paper, we only considered test data with observed failure times for all tested components. If test data also contain right-censored observations, this can also be dealt with, both in the imprecise Bayesian and NPI approaches (Walter, Graham, & Coolen 2015, Coolen & Yan 2004, Maturi 2010) (more information about NPI is available from www.npi-statistics.com). This generalization is further relevant as, instead of assuming availability of test data, it allows us to take process data for the actual components in a system into account while this system is operating, hence enabling inference on the remaining time until system failure.

Upscaling the survival signature to large real-world systems and networks, consisting of thousands of components, is a major challenge. However, even for such systems the fact that one only needs to derive the survival signature once for a system is an advantage, and also the monotonicity of the survival signature for coherent systems is very useful if one can only derive it partially. For small to medium-sized systems and networks, the survival signature is particularly easy to compute using the Reliability Theory R package (Aslett 2016b), available from www.louisaslett.com. Using this package it is straightforward to express your system in terms of an undirected graphical structure, after which a single call to the function compute System Survival Signature suffices. The function will compute all of the cut sets of the system and perform the combinatorial analysis, returning a table which contains the survival signature just as in Tables 2 and 3. For example, computation of the survival signature for the system in Figure 1 is achieved with 3 simple commands

```
s <- graph.formula(s-1-2-3-t,
                   s-1-4-5-t,
                   2:4-6-3:5)
setCompTypes(s,
        list("T1"=c(2,4,3,5),
             "T2"=6,
             "T3"=1))
computeSystemSurvivalSignature(s)
```

Full instructions and some worked examples are available within the package. There are numerous other functions in the package, enabling computation of: the legacy system signature (Samaniego 2007); the continuous-time Markov chain representation of repairable systems; as well as numerous inference algorithms for Bayesian inference on the system signature using only system-level data (Aslett 2013).

Full instructions and some worked examples are available within the package. There are numerous other functions in the package, enabling computation of: the legacy system signature (Samaniego 2007); the continuous-time Markov chain representation of repairable systems; as well as numerous inference algorithms for Bayesian inference on the system signature using only system-level data (Aslett 2013).

The survival signature enables some interesting applications which would otherwise be intractably difficult. For example, often a system designer may consider the design (structure) of their system to be a trade secret and so be unwilling to release it to component manufacturers, while at the same time component manufacturers are frequently unwilling to release anything more than summary figures for components, e.g. mean-time-between-failures. These two opposing goals lead to a situation in which it would seem unrealistic to achieve a full probabilistic reliability assessment and to honour the privacy requirements of all parties. However, recent work (Aslett 2016a) makes use of the survival signature to allow cryptographically secure evaluation of the system reliability function, where the functional form resulting from the survival signature decomposition in Equation (1) is crucial to enabling encrypted computation using so-called homomorphic encryption schemes (Aslett, Esperança, & Holmes 2015). The equivalent decomposition in terms of the structure function leads to difficulties in encrypted computation, so that this application may be intractable without use of the survival signature.

ACKNOWLEDGEMENTS

The authors wish to thank Professor Radim Bris for his kind invitation to present this work at the ICAMER 2016 conference. Louis Aslett was supported by the i-like project (EPSRC grant reference number EP/K014463/1). Gero Walter was supported by the DINALOG project 'Coordinated Advanced Maintenance and Logistics Planning for the Process Industries' (CAMPI).

REFERENCES

Aslett, L. (2013). *MCMC for Inference on Phase-type and Masked System Lifetime Models*. Ph. D. thesis, Trinity College Dublin.

Aslett, L. (2016a). Cryptographically secure multiparty evaluation of system reliability. *Pending journal submission*.

Aslett, L. (2016b). *Reliability Theory: Tools for structural reliability analysis*. R package.

Aslett, L., F. Coolen, & S. Wilson (2015). Bayesian inference for reliability of systems and networks using the survival signature. *Risk Analysis 35*, 1640–1651.

Aslett, L., P. Esperança, & C. Holmes (2015). A review of homomorphic encryption and software tools for encrypted statistical machine learning. Technical report, University of Oxford.

Augustin, T. & F. Coolen (2004). Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference 124*, 251–272.

Augustin, T., F. Coolen, G. de Cooman, & M. Troffaes (2014). *Introduction to Imprecise Probabilities.* Chichester: Wiley.

Barlow, R. & F. Proschan (1975). *Statistical Theory of Reliability and Life Testing.* New York: Holt, Rinehart and Winston.

Coolen, F. (1998). Low structure imprecise predictive inference for Bayes' problem. *Statistics & Probability Letters 36*, 349–357.

Coolen, F. (2006). On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information 15*, 21–47.

Coolen, F. (2011). Nonparametric predictive inference. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science*, pp. 968–970. Springer.

Coolen, F. & T. Coolen-Maturi (2012). On generalizing the signature to systems with multiple types of components. In W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, and J. Kacprzyk (Eds.), *Complex Systems and Dependability*, pp. 115–130. Springer.

Coolen, F. & T. Coolen-Maturi (2015). Predictive inference for system reliability after common-cause component failures. *Reliability Engineering and System Safety 135*, 27–33.

Coolen, F., T. Coolen-Maturi, & A. Al-nefaiee (2014). Nonparametric predictive inference for system reliability using the survival signature. *Journal of Risk and Reliability 228*, 437–448.

Coolen, F. & L. Utkin (2011). Imprecise reliability. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science*, pp. 649–650. Springer.

Coolen, F. & K. Yan (2004). Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference 126*, 25–54.

De Finetti, B. (1974). *Theory of Probability*. Chichester: Wiley.

Maturi, T. (2010). *Nonparametric Predictive Inference for Multiple Comparisons*. Ph. D. thesis, Durham University.

Samaniego, F. (2007). *System Signatures and their Applications in Engineering Reliability*. New York: Springer.

Walter, G. (2013). *Generalized Bayesian Inference under Prior-Data Conflict*. Ph. D. thesis, Ludwig Maximilian University of Munich.

Walter, G., L. Aslett, & F. Coolen (2016). Bayesian nonparametric system reliability with sets of priors. *In submission.*

Walter, G., A. Graham, & F. Coolen (2015). Robust Bayesian estimation of system reliability for scarce and surprising data. In L. Podofillini, B. Sudret, B. Stojadinović, E. Zio, and W. Kröger (Eds.), *Safety and Reliability of Complex Engineered Systems: ESREL 2015*, pp. 1991–1998. CRC Press.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Parallel algorithms of system availability evaluation

M. Kvassay, V. Levashenko & E. Zaitseva
University of Zilina, Zilina, Slovakia

ABSTRACT: There are different methods for the calculation of indices and measures in reliability analysis. Some of most used indices are system availability/reliability and importance measures. In this paper new algorithms for the calculation of system availability and some of importance measures are developed based on the parallel procedures. The principal step in the development of these algorithms are construction of matrix procedures for the calculation these indices and measures.

1 INTRODUCTION

The estimation of a system reliability is provided based on different indices and measures. As a rule the computational complexity of the calculation of these indices and measures depends on the system dimension. One of way this computational complexity decreasing is the use of parallel procedure (Green et al. 2011, Lingfeng & Singh 2009).

Kucharev et al. (1990) shown that the parallel procedure can be designed based on matrix interpretation of computational procedure. Therefore the transformation of traditional computational procedures for the calculation of indices and measures in matrix form is important step in the design of parallel algorithms. In this paper we consider such transformation for calculation of system availability and some of *Importance Measures* (IMs). The initial representation and mathematical description of investigated system in this case must be defined in matrix or vector form. There are some typical form of investigated system representation in reliability analysis: structure function; Markovian model; Mote-Carlo model etc. The structure function can be considered as Boolean function (Barlow & Proschan 1975). This interpretations allows using of vector representation of Boolean function for a structure function too. Therefore in this paper the parallel algorithm for the calculation of system availability is developed based on structure function of system. The structure based algorithms are used in the development of parallel algorithms for calculation of IMs too.

The importance analysis enables one to estimate the impact of a system element on the system failure or functioning. Consideration is given at that to the structural distinctions of the system and the failure/operability probabilities of its elements. By the system operability is meant its ability to function at a fixed time instant (Barlow & Proschan 1975). Analysis of element importance is used in the system design, diagnosis, and optimization. Many IMs are used today to allow for various aspects of the impact of system elements on its failure or operability.

2 A SYSTEM REPRESENTATION BY STRUCTURE FUNCTION

2.1 *The structure function*

Consider a system of n components. Every component state is designated as x_i ($i = 1,...,n$) where $x_i = 1$ is working state of the component and $x_i = 0$ indicates the system failure. The probability of failure is defined for every system component as $q_i = \Pr\{x_i = 0\}$. Therefore the probability of the i-th component is $p_i = \Pr\{x_i = 1\} = 1 - q_i$.

The structure function of the system defines correlation of system state depend on system component states unambiguously (Zaitseva 2012):

$$\phi(x) = \phi(x_1, \ldots, x_n): \{0, 1\}^n \rightarrow \{0, 1\}. \quad (1)$$

There are two groups of system type that are coherent and non-coherent system. The coherent system has assumption (Beeson & Andrews 2003, Fricks & Trivedi 2003):

a. The system and its components have two states: up (working) and down (failed);
b. All system components are relevant to system;
c. The system structure function is monotone non-decreasing: $\phi(x_1, \ldots, 1, \ldots, x_n) \neq \phi(x_1, \ldots, 1, \ldots, x_n)$;
d. The failure and repair rate of the components are constant;
e. Repaired components are as good as new.

The system is non-coherent if one or more of these assumption are not thru.

Consider typical form of the structure function (1) representation. The function (1) is Boolean function. It is permits to use mathematical approach of Boolean algebra for this function representation and investigation its properties.

There are some representations of the structure function (1) in point of view Boolean algebra. Truth table, Binary Decision Diagram (BDD) and analytical representation (formula) can be used for the structure function initial description according to (Brown & Vranesic 2000).

2.2 *Table and matric representation of the structure function*

A truth table includes a list of combinations of 1's and 0's assigned to the binary variables, and column that shows the value of the function each binary combination (Fig. 1). The number of rows in the truth table is 2^n, where n is the number of variables.

The binary combination of variables in the truth table can be ordered from 0 to 2^n-1 according to coding (lexicographical order). The fixed order of variable allows consider the column of function values only (Fig. 1). Such representation of Boolean function is named as truth table column vector or truth vector.

Therefore the structure function of system (1) can be represented by the truth table or truth vector unambiguously. For example, consider the trivial system of three components (n = 3) in Fig. 2. The truth table is shown in Table 1 and the truth vector of this system is $\mathbf{x} = [0\ 0\ 0\ 0\ 0\ 1\ 1\ 1]^T$. Consider the truth vector element $x^{(5)} = 1$. The state vector for this function value is defined by the transformation of the parameter $i = 5 \Rightarrow (i_1, i_2, i_3) = (1, 0, 1)$. Therefore, the truth vector element $x^{(5)} = 1$ agrees with the function value $\phi(1, 0, 1) = 1$.

2.3 *Binary decision diagram*

A BDD is a directed acyclic graph of a Boolean function representation. This graph has two terminal nodes, labelled 0 and 1. Each non-terminal node is labelled with a function variable x_i and has two outgoing edges. The left edge is labelled "0" and the other outgoing edge is labelled "1".

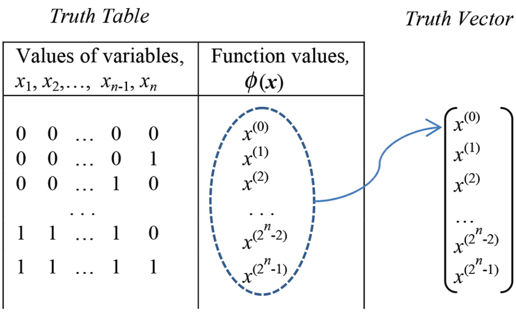


Figure 1. Truth vector of the structure function.

A BDD is a widely used tool reliability analysis. Some methods for reliability analysis based on this tool are discussed in papers (Zaitseva et al. 2015, Chang et al. 2004).

Terminal nodes of the BDD correspond to the system state. Non-terminal node outgoing edges are interpreted as component states.

For example, the BDD of the structure function of the series parallel system in Fig. 2 is shown in Fig. 3.

2.4 *Analytical representation of the structure function*

The analytical representation has different form for the function. As a rule this function are define by the formula with operators AND, OR and NOT. For example, the structure function of the system in Fig. 2 can be presented by the formula:

$$\phi(\mathbf{x}) = AND(x_1, OR(x_2, x_3)). \tag{2}$$

But there are other analytical representations for Boolean function and one of them is arithmetic polynomial form $A(\mathbf{x})$ (Kucharev et al. 1990):

$$\begin{aligned}\phi(\mathbf{x}) = A(\mathbf{x}) &= \sum_{k=0}^{2^n-1} a^{(k)} x_1^{k_1} x_2^{k_2} \dots x_n^{k_n} = \\ &= a^{(0)} + a^{(1)} x_n + a^{(2)} x_{n-1} + a^{(3)} x_{n-1} x_n + \dots + \\ &\quad + a^{(2^n-2)} x_1 \dots x_{n-1} + a^{(2^n-1)} x_1 \dots x_n,\end{aligned} \tag{3}$$

Table 1. Truth table of the structure function.

Values of variables, x_1, x_2, x_3	Function values, $\phi(\mathbf{x})$
0 0 0	0
0 0 1	0
0 1 0	0
0 1 1	0
1 0 0	0
1 0 1	1
1 1 0	1
1 1 1	1

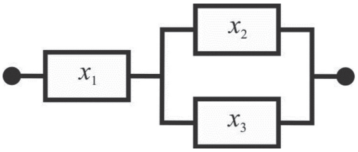


Figure 2. A simple series-parallel system.

where $a^{(k)}$ is polynomial coefficients; $k_1 \ldots k_i \ldots k_n$ is binary description of parameter k ($k = 0, 1, \ldots, 2^n\text{-}1$); $x_i^{k_i} = x_i$ if $k_i = 1$ and $x_i^{k_i} = 1$ if $k_i = 0$.

This arithmetic polynomial form (3) in matrix form is (Kucharev et al. 1990):

$$\mathbf{x} = \mathrm{A}_n \cdot \mathbf{a} \tag{4}$$

where $\mathbf{x} = [x^{(0)}\ x^{(1)}\ \ldots\ x^{(2n\text{-}1)}]^{\mathrm{T}}$ is the truth vector of function (1); $\mathbf{a} = [a^{(0)}\ a^{(1)}\ \ldots\ a^{(2n\text{-}1)}]^{\mathrm{T}}$ is vector of coefficients $a^{(k)}$ for polynomial (3); A_n is matrix that is calculated by recurrent equation:

$$\mathrm{A}_n = \mathrm{A}_1 \otimes \mathrm{A}_{n-1}, \quad \mathrm{A}_1 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \tag{5}$$

where ⊗ is symbol of Kronecker product and the element of the matrix A_1 is calculated as:

$$a_{st} = s^t, \tag{6}$$

for $s, t \in \{0, 1\}$ and $0^0 \equiv 1$.

The polynomial coefficients $a^{(k)}$ can be calculate based on inverse matrix procedure (Kucharev et al. 1990):

$$\mathbf{a} = \tilde{\mathrm{A}}_n \cdot \mathbf{x} \tag{7}$$

where $\tilde{\mathrm{A}}_n$ is inverse matrix for A_n and:

$$\tilde{\mathrm{A}}_n = \tilde{\mathrm{A}}_1 \otimes \tilde{\mathrm{A}}_{n-1}, \quad \tilde{\mathrm{A}}_1 = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}. \tag{8}$$

For example, consider the trivial system of three components ($n = 3$) in Fig. 2. The truth vector of this function is $\mathbf{x} = [0\ 0\ 0\ 0\ 0\ 1\ 1\ 1]^{\mathrm{T}}$. Computer the coefficient of polynomial (3) according equation (8):

$$\mathbf{a} = \mathbf{K}_2 \cdot \mathbf{x}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\ -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ -1 \end{bmatrix}. \tag{9}$$

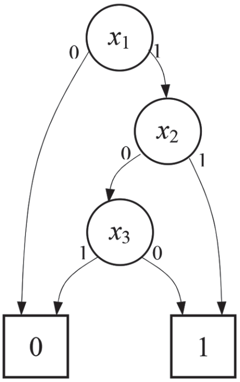
Describe the arithmetic polynomial form $A(\boldsymbol{x})$ for the structure function of the system in Fig. 1. The arithmetic polynomial form for $n = 3$ is:

$$\begin{aligned} \phi(\boldsymbol{x}) = A(\boldsymbol{x}) &= \sum_{k=0}^{7} a^{(k)} x_1^{k_1} x_2^{k_2} x_3^{k_3} \\ &= a^{(0)} x_1^0 x_2^0 x_3^0 + a^{(1)} x_1^0 x_2^0 x_3^1 + a^{(2)} x_1^0 x_2^1 x_3^0 \\ &\quad + a^{(3)} x_1^0 x_2^1 x_3^1 + a^{(4)} x_1^1 x_2^0 x_3^0 + a^{(5)} x_1^1 x_2^0 x_3^1 \\ &\quad + a^{(6)} x_1^1 x_2^1 x_3^0 + a^{(7)} x_1^1 x_2^1 x_3^1 \\ &= a^{(0)} + a^{(1)} x_3 + a^{(2)} x_2 + a^{(3)} x_2 x_3 + a^{(4)} x_1^1 \\ &\qquad + a^{(5)} x_1 x_3 + a^{(6)} x_1 x_2 + a^{(7)} x_1 x_2 x_3 \end{aligned} \tag{10}$$

Use the vector coefficients of the structure function (9) $\mathbf{a} = [0\ 0\ 0\ 0\ 0\ 1\ 1\ \text{-}1]^{\mathrm{T}}$ for this form:

$$\phi(\boldsymbol{x}) = A(\boldsymbol{x}) = x_1 x_3 + x_1 x_2 - x_1 x_2 x_3 \tag{11}$$

BDD of the system



The pseudo-code of the BDD

```
ϕ(x) = ite(x₁, 0, G)
   G = ite(x₂, K, 1)
   K = ite(x₃, 0, 1)
or
ϕ(x)=ite(x₁, 0, ite(x₂, ite(x₃,0,1),1))
```

Figure 3. BDD of structure function of simple series-parallel system in Fig. 2.

3 THE SYSTEM AVAILABILITY

Every system component is characterized by probability p_i (represents the availability of component i) and probability q_i (defines its unavailability):

$$p_i = \Pr\{x_i = 1\},\ q_i = \Pr\{x_i = 0\},\ p_i + q_i = 1. \qquad (12)$$

When the system structure function and availabilities of all system components are known, then system availability/unavailability can be computed as follows (Barlow & Proschan 1975, Schneeweiss 2009):

$$A = \Pr\{\phi(\boldsymbol{x}) = 1\},\ U = \Pr\{\phi(\boldsymbol{x}) = 0\},\ A + U = 1. \qquad (13)$$

The availability is one of the most important characteristics of any system. It can also be used to compute other reliability characteristics, e.g. mean time to failure, mean time to repair, etc. (Beeson & Andrews 2003, Schneeweiss 2009).

There is interesting property of the arithmetic polynomial form. The replacement of the variables x_i by the probabilities of component working p_i allows to obtain the probabilistic form of the structure function that is system availability.

Theorem 1. The system availability (probability of the working state) is calculate by arithmetical polynomial form (3) in which the Boolean variables x_i are changed by relevant probability of component state working:

$$A = \sum_{k=0}^{2^n-1} a^{(k)} \cdot p_1^{k_1} \cdot p_2^{k_2} \cdot \ldots \cdot p_n^{k_n} \qquad (14)$$

where $a^{(k)}$ is coefficients polynomial (3); p_i ($i = 1, \ldots, n$) is probability of the i-th component working state, and $p_i^{k_i} = 1$ if $k_i = 0$ and $p_i^{k_i} = p_i$ if $k_i = 1$.

Proof. According to (Kucharev et al. 1990) the arithmetical polynomial form is canonical form of Boolean function representation. Therefore all elements of the polynomial form (3) $a^{(k)} \cdot x_1^{k_1} \cdot x_2^{k_2} \cdot \ldots \cdot x_n^{k_n}$ are mutually independent events. And variables of Boolean function are interpreted as independent events according to Kumar & Breuer 1981. Therefore in case of probabilistic analysis the Boolean function variables can be replaced by probabilities of this events.

For example, compute the availability of the system in Fig. 1 based on the arithmetical polynomial form of this system structure function (11). According to the Theorem 1 the Boolean variables of this form are replaced by the probabilities p_i of the system components functioning:

$$A = p_1 p_3 + p_1 p_2 - p_1 p_2 p_3. \qquad (15)$$

In comparison, compute this system availability according to traditional way based on the structure function AND-OR-representation (2):

$$\begin{aligned} A &= \Pr\{AND(x_1, OR(x_2, x_3))\} = p_1(p_2 + p_3 - p_2 p_3) \\ &= p_1 p_2 + p_1 p_3 - p_1 p_2 p_3. \end{aligned} \qquad (16)$$

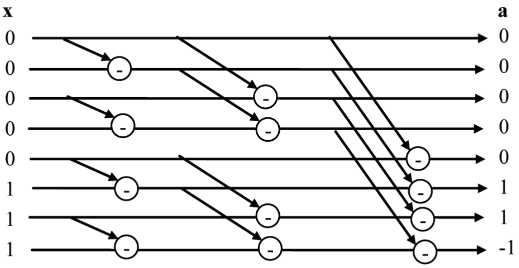


Figure 4. Calculation of the coefficients of the probabilistic form of the structure function.

We can see, the availability of the system computed by new methods and traditional way are equal. But the calculation of system availability based on matrix procedure is formal background for the development of parallel algorithms (Kucharev et al. 1990). For example, the flow diagrams for the calculation of the coefficient of the probabilistic form to compute the system availability are design based on the transformation (9) and is presented in Fig. 4.

4 IMPORTANCE ANALYSIS

The availability is one of the most important characteristics of any system. It can also be used to compute other reliability characteristics, e.g. mean time to failure, mean time to repair, etc. (Barlow & Proschan 1975, Schneeweiss 2009). But they do not permit to identify the influence of individual system components on the proper work of the system. For this purpose, there exist other measures that are known as Importance Measures (IM). The IMs are used in part of reliability analysis that is known as importance analysis. The comprehensive study of these measures has been performed in work [4]. IMs have been widely used for identifying system weaknesses and supporting system improvement activities from design perspective. With the known values of IMs of all components, proper actions can be taken on the weakest component to improve sys-tem availability at minimal costs or effort.

There exist a lot of IMs, but the most often used are the Structural Importance (SI), Birnbaum's Importance (BI), Criticality Importance (CI) (Table 2).

Different mathematical methods and algorithms can be used to calculate these in-dices. Ones of them are Direct Partial Boolean Derivatives (DPBDs) that have been introduced for importance analysis in paper (Moret & Thomason 1984). In paper (Zaitseva & Levashenko 2013), the mathematical background of DPBDs application has been considered. But efficient algorithm for computation of DPBDs has not been proposed. In this paper,

Table 2. Basic importance measures.

Importance Measure	Meaning
SI	The SI concentrates only on the topological structure of the system. It is defined as the relative number of situations in which a given component is critical for the system activity
BI	The BI of a given component is defined as the probability that the component is critical for the system work.
CI	The CI of a given component is calculated as the probability that the system failure has been caused by the component failure, given that the system is failed.

a new parallel algorithm for the calculation of a DPBD is developed.

As alternative result for the new algorithm, algorithms in (Zaitseva et al 2015) can be considered. The authors of the paper (Zaitseva et al 2015) proposed algorithms for calculation of a DPBD based on the structure function representation by a Binary Decision Diagram (BDD) that includes parallel procedure too. But the algorithms in (Zaitseva et al 2015) need a special transformation of initial representation of the structure function into a BDD, and this increases the computation complexity.

4.1 *Direct partial Boolean derivatives*

A DPBD is a part of Logical Differential Calculus (Moret & Thomason 1984, Bochmann & Posthoff 1981). In analysis of Boolean functions, a DPBD allows identifying situations in which the change of a Boolean variable results the change of the value of Boolean function. In case of reliability analysis, the system is defined by the structure function (1) that is a Boolean function. Therefore, a DPBD can be used for the structure function analysis too. In terms of reliability analysis, a DPBD allows investigation the influence of a structure function variable (= component state) change on a function value change (= system state). Therefore, a DPBD of the structure function permits indicating components states (state vectors) for which the change of one component state causes a change of the system state (availability). These vectors agree with the system boundary states (Moret & Thomason 1984, Zaitseva & Levashenko 2013).

DPBD $\partial\phi(j \to \bar{j})/\partial x_i(a \to \bar{a})$ of the structure function $\phi(\boldsymbol{x})$ with respect to variable xi is defined as follows (Bochmann & Posthoff 1981):

$$\frac{\partial\phi(j\to\bar{j})}{\partial x_i(a\to\bar{a})}=\left\{\phi(a_i,\boldsymbol{x})\leftrightarrow j\right\}\wedge\left\{\phi(\bar{a}_i,\boldsymbol{x})\leftrightarrow\bar{j}\right\},. \quad (17)$$

where $\phi(a_i, \boldsymbol{x}) = \phi(x_1, x_2,\ldots, x_{i-1}, a, x_{i+1},\ldots, x_n)$, $a, j \in \{0, 1\}$ and $\leftrightarrow$ is the symbol of equivalence operator (logical bi-conditional).

Clearly, there exist four DPBDs for every variable x_i (Bochmann & Posthoff 1981, Zaitseva & Levashenko 2013):

$$\frac{\partial\phi(1\to 0)}{\partial x_i(1\to 0)},\frac{\partial\phi(0\to 1)}{\partial x_i(0\to 1)},\ \frac{\partial\phi(1\to 0)}{\partial x_i(0\to 1)},\frac{\partial\phi(0\to 1)}{\partial x_i(1\to 0)}.$$

In reliability analysis, the first two DPBDs can be used to identify situations in which a failure (repair) of component *i* results system failure (repair). Similarly, the second two DPBDs identify situations when the system failure (repair) is caused by the *i*-th component repair (failure). The second two derivatives exist (are not equal to zero) for a noncoherent systems (Zaitseva & Levashenko 2013).

For example, consider a system of three components ($n = 3$) in Fig. 1. The influence of the first component failure on the system can be analyzed by DPBD $\partial\phi(1\to0)/\partial x_1(1\to0)$. This derivative has three nonzero values for state vectors $\boldsymbol{x} = (x_1, x_2, x_3)$: $(\underline{1\to0}, 1, 1)$, $(\underline{1\to0}, 0, 1)$ and $(\underline{1\to0}, 1, 0)$. Therefore, the failure of the first component causes a system breakdown for working state of the second and the third component or working state of one of them. The system is not functioning if the second and the third components are failed and, therefore, a failure of the first component does not influence system availability.

4.2 *Importance measures and Direct Partial Boolean Derivatives*

In reliability analysis, the structure function and the system components are used instead of the Boolean function and the Boolean variables, respectively. Using this coincidence, the authors of the papers (Zaitseva & Levashenko 2013) have developed techniques for analysis of influence of individual system components on system failure/functioning using DPBDs. Let us summarize the definitions of IMs (Table 2) for the system failure based on DPBDs.

The SI of component is defined as the relative number of situations, in which the component is critical for system failure. Therefore, the SI of component can be defined by DPBD $\partial\phi(1\to0)/\partial x_i(1\to0)$ as the relative number of state vectors for which the considered DPBD has nonzero values (Zaitseva & Levashenko 2013, Zaitseva 2012):

$$SI_i = \frac{\rho_i^{(1\to 0)}}{2^{n-1}}. \quad (18)$$

where is a number of nonzero values of DPBD $\partial\phi(1\to 0)/\partial x_i(1\to 0)$ and 2^n-1 is a size of the DPBD.

Similarly, the modified SI, which takes into account the necessary condition for component being critical, can be defined as follows (Zaitseva & Levashenko 2013, Zaitseva 2012):

$$MSI_i = \frac{\rho_i^{(1\to 0)}}{\rho_i}. \quad (19)$$

where ρ_i is a number of state vectors for which $\phi(1_i, \boldsymbol{x}) = 1$.

The BI of component *i* defines the probability that the *i*-th system component is critical for system failure. Using DPBDs, this IM can be defined as the probability that the DPBD is nonzero (Zaitseva & Levashenko 2013)

$$BI_i = \Pr\{\partial\phi(1\to 0)/x_i(1\to 0) \leftrightarrow 1\}. \quad (20)$$

A lot of IMs are based on the BI, e.g. the CI, Barlow-Proschan, Bayesian, redundancy, etc. For example, the CI is calculated as follows (Kuo & Zhu 2012):

$$CI_i = BI_i \cdot \frac{q_i}{U}. \quad (21)$$

where q_i is component state probability (1) and U is the system unavailability.

To illustrate the calculation of all IMs using DPBDs consider the system in Fig. 1. Values of IMs for this system are computed in Table 3. According to these IMs, the first component has the most influence on the system failure from point of view of the system structure, because the values of the SI, MSI and BI are greatest for this component. The CI is maximal for the second and third components and, therefore, it indicates the first component as non-important taking into account the probability of failure of this component (it is minimal for this component, i.e. q1 = 0.10). The FVIs implies that the second and third components contribute to system failure with the most probability.

So, DPBDs are one of possible mathematical approaches that can be used in importance analysis, and they allow us to calculate all often used IMs (Table 2). Mathematical background of its application for the definition of IM has been considered in papers (Zaitseva & Levashenko 2013, Zaitseva 2012). In this paper new algorithm for the calculation of DPBD based on a parallel procedure is developed.

Table 3. IMs for the system in Fig. 1.

Component	x_1	x_2	x_3
Probability of component state, p_i	0.90	0.70	0.65
SI$_i$	0.75	0.25	0.25
MSI$_i$	1.00	0.50	0.50
BI$_i$	0.90	0.32	0.27
CI$_i$	0.46	0.49	0.49

4.3 *Parallel algorithm for the calculation of direct partial Boolean derivatives*

One of possible way for the formal development of parallel algorithms is transform mathematical background into matrix algebra. Therefore, consider DPBD (17) in matrix interpretation. As the first step in such transformation, the initial data (structure function) has to be presented as a vector or matrix.

The structure function is defined as a truth vector (Fig. 1) in matrix algorithm for calculation of DPBD. The truth vector of DPBD (derivative vector) is calculated based on the truth vector of the structure function as:

$$\begin{aligned}&\partial\mathbf{x}(j\to\bar{j})/\partial x_i(a\to\bar{a})\\ &\quad = \left(\mathbf{P}^{(i,a)}\cdot(j\leftrightarrow\mathbf{x})\right)\wedge\left(P^{(i,\bar{a})}\cdot(\bar{j}\leftrightarrow\mathbf{x})\right)\end{aligned} \quad (22)$$

where $\mathbf{P}^{(i,l)}$ is the differentiation matrix with size $2^{n-1} \times 2^n$ that is defined as:

$$\mathbf{P}^{(i,l)} = \mathbf{M}^{(i-1)} \otimes \left[l\,\bar{l}\right] \otimes \mathbf{M}^{(n-i)}. \quad (23)$$

and $\mathbf{M}^{(w)}$ is diagonal matrix with size $2^w \times 2^w$, $\left[l\,\bar{l}\right]$ is the vector for which $l = s$ for the matrix $\mathbf{P}^{(i,a)}$ and $l = \bar{a}$ for matrix $\mathbf{P}^{(i,\bar{a})}$, and ⊗ is the Kronecker product (Kucharev et al. 1990).

Note that the calculation $(j\leftrightarrow\mathbf{x})$ and $(\bar{j}\leftrightarrow\mathbf{x})$ in (22) agrees with the definition of state vectors for which the function value is j and $\bar{j}$, respectively. The matrices $\mathbf{P}^{(i,a)}$ and $\mathbf{P}^{(i,\bar{a})}$ allows indicating variables with values a and $\bar{a}$, respectively. The operation AND (∧) integrates these conductions.

DPBD $\partial\phi(j\to\bar{j})/\partial x_i(a\to\bar{a})$ does not depend on the *i*-th variable (Bochmann & Posthoff 1981). Therefore, the derivative vector (22) $\partial\mathbf{x}(j\to\bar{j})/\partial x_i(a\to\bar{a})$ has size of 2^{n-1}.

Consider an example for calculation of derivative vector $\partial\mathbf{x}(1\to 0)/\partial x_1(1\to 0)$ for the structure function with the truth vector $\mathbf{x} = [0\ 0\ 0\ 0\ 0\ 1\ 1\ 1]^T$ (it is the truth vector of the structure function of the system depicted in Fig. 1). According to (22), the rule for the calculation of this derivative is:

$$\begin{aligned}&\partial\mathbf{x}(1\to 0)/\partial x_1(1\to 0)\\ &= \left(\mathbf{P}^{(1,1)}\cdot(1\leftrightarrow\mathbf{x})\right)\wedge\left(\mathbf{P}^{(1,0)}\cdot(0\leftrightarrow x)\right) = [0111]^T.\end{aligned} \quad (24)$$

where matrices $\mathbf{P}^{(1,1)}$ and $\mathbf{P}^{(1,0)}$ are defined based on the rule (23) as:

$$\mathbf{P}^{(1,1)} = \mathbf{M}^{(0)} \otimes [1 \quad 0] \otimes \mathbf{M}^{(2)}$$

and

$$\mathbf{P}^{(1,1)} = \mathbf{M}^{(0)} \otimes [0 \quad 1] \otimes \mathbf{M}^{(2)}.$$

The derivative vector $\partial\mathbf{X}(1\to0)/\partial x_2(1\to0)$ DPBD indicate three state vectors $\boldsymbol{x} = (x_1, x_2, x_3)$: ($\underline{1\to0}$, 1, 1), ($\underline{1\to0}$, 0, 1) and ($\underline{1\to0}$, 1, 0). Therefore, the failure of the first component causes a system breakdown for working state of the second and the third components or working state of one of them. This result is equal to result that has been calculated by definition (17) for DPBD $\partial\phi(1\to0)/\partial x_1(1\to0)$.

A matrix procedure can be transform in parallel procedure according to (Kucharev et al. 1990). Therefore the equation (22) can be interpreted by parallel procedure. For example, the flow diagrams for the calculation of the derivative vectors $\partial\mathbf{x}(1\to0)/\partial x_1(1\to0)$, $\partial\mathbf{x}(1\to0)/\partial x_2(1\to0)$ and $\partial\mathbf{x}(1\to0)/\partial x_3(1\to0)$ for the structure function of the system in Fig. 2 according (22) are presented in Fig. 5. These diagrams illustrate the possibility to use parallel procedures for the calculation of DPBD.

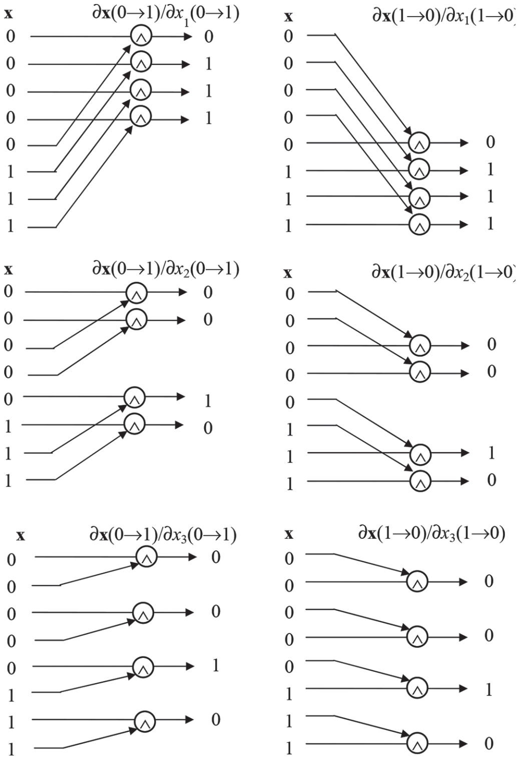


Figure 5. Calculation of DPBDs based on parallel procedures.

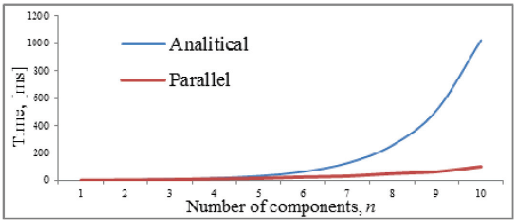


Figure 6. Computation time for calculation of DPBDs based on analytical and parallel procedures.

5 CONCLUSION

In this paper the new algorithm based on the parallel procedures is proposed for the calculation of system availability and most often-used IMs (Table 2). The algorithm for the computation of the system availability are based on the use of the probabilistic form of the structure function in point of view of Boolean algebra. The parallel procedure is used for the calculation the coefficients of this form (14).

The algorithm for the calculation of IMs are based on the use of the DPLDs (17). The parallel procedure allows to compute the values of the derivative (22). The computational complexity of the proposed algorithm is less in comparison with algorithm based on the typical analytical calculation (Fig. 6).

The proposed algorithm for the calculation of IMs based on the parallel procedures can be used in many practical applications. The principal step in these applications is representation of the investigated object by the structure function. As a rule the structure function is defined based on analysis of the structure of investigated object.

ACKNOWLEDGEMENT

This work was partially supported by the grant VEGA 1/0498/14 and the grant of 7th RTD Framework Program No 610425 (RASimAs).

REFERENCES

Barlow, R.E., Proschan, F. (1975). Importance of system components and fault tree events, *Stochastic Processes and their Applications*, 3(2): 153–173.

Beeson, S., Andrews, J.D. (2003) Importance measure for non-coherent-system analysis, *IEEE Trans. Reliability* 52(3) 2003: 301–310.

Bochmann, D., Posthoff, C. (1981). *Binary Dynamic Systems*. Berlin, Academic Verlag.

Brown, S., Vranesic, Z. (2000) *Fundamentals of Digital Logic with VHDL design*, McGraw-Hill.

Chang, Y., Amari, S., Kuo, S. (2004) Computing system failure frequencies and reliability importance measures using OBDD, IEEE Trans Computers, 53(1): 54–68.

Fricks, R., Trivedi, K. (2003) Importance analysis with Markov chains, *Proc 49th IEEE Annual Reliability Maintainability Symp. Tampa, USA, 2003*.

Green, R.C., Lingfeng Wang, Alam, M., Singh, C. (2011). Intelligent and parallel state space pruning for power system reliability analysis using MPI on a multicore platform, *Proc IEEE PES on Innovative Smart Grid Technologies (ISGT)*, 2011.

Kucharev, G.A. Shmerko, V.P., Zaitseva, E.N. (1990) *Multiple-Valued Data Processing Algorithms and Systolic Processors*, Minsk: Nauka and Technica.

Kumar, S.K., Breuer, M. (1981). Probabilistic Aspects of Boolean Switching Function via a New Transform, *Journal of the Association for Computing Machinery*, 28(3): 502–520.

Kuo, W., Zhu, X. (2012). *Importance Measures in Reliability, Risk and Optimization*. John Wiley & Sons, Ltd.

Lingfeng, W, Singh, C. (2009). Multi-deme parallel genetic algorithm in reliability analysis of composite power systems, *Proc. of IEEE Conf. on PowerTech*, 2009.

Moret, B.M.E., Thomason, M.G. (1984). Boolean Difference Techniques for Time-Sequence and Common-Cause Analysis of Fault-Trees, *IEEE Trans. Reliability*, R-33: 399–405.

Schneeweiss, W.G. (2009). A short Boolean derivation of mean failure frequency for any (also non-coherent) system, *Reliability and Engineering System Safety*, 94(8): 1363–1367.

Zaitseva, E. (2012). Importance analysis of a multi-state system based on multiple-valued logic methods. In: Lisnianski A. and Frenkel I. (eds) *Recent Advances in System Reliability: Signatures, Multi-state Systems and Statistical Inference*, London: Springer, pp. 113–134.

Zaitseva, E., Levashenko, V. (2013). Importance Analysis by Logical Differential Calculus. *Automation and Remote Control*, 74(2): 171–182.

Zaitseva, E., Levashenko, V., Kostolny, J., (2015) Application of logical differential calculus and binary decision diagramin importance analysis. *Eksploatacja i Niezawodnosc – Maintenance and Reliability*; 17 (3): 379–388.

Advanced mathematical methods for engineering

Advanced methods to solve partial differential equations

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

A note on the Cauchy-Riemann equation on a class of convex domains of finite and infinite type in $\mathbb{C}^2$

L.K. Ha
Faculty of Mathematics and Computer Science University of Science,
Vietnam National University, Ho Chi Minh City, Vietnam

ABSTRACT: In this paper, we provide an extension of the Hölder regularity result by Range in (Range 1978) to a certain class of strict finite and infinite type convex domains in $\mathbb{C}^2$. A new notion of type is introduced for arbitrary convex domains in $\mathbb{C}^2$ with smooth boundaries. This type generalizes the notion of strict finite type in the original theory (Range 1978) as well as consists many cases of infinite type in which Range's method is fail to be applied.

1 INTRODUCTION

Let $(z_1,\ldots,z_n)$ be the complex Euclidean coordinates of $\mathbb{C}^n$, with $n\geq 1$, and let $\Omega\subset\mathbb{C}^n$ be a bounded domain. The Cauchy-Riemann complex on $C^1(\Omega)$ -functions is defined to be

$$\bar{\partial}u=\sum_{j=1}^{n}\frac{\partial u}{\partial\bar{z}_j}d\bar{z}_j,$$

where $\frac{\partial}{\partial\bar{z}_j}=\frac{1}{2}\left(\frac{\partial}{\partial x_j}+\sqrt{-1}\frac{\partial}{\partial y_j}\right)$ with $z_j=x_j+\sqrt{-1}y_j$, $j=1,\ldots,n.$

One of most fundamental and important problems in multidimensional complex analysis is to solve the Cauchy-Riemann equation

$$\bar{\partial}u=\varphi$$

for a given (0, 1)-form $\varphi=\sum_{j=1}^{2}\varphi_j d\bar{z}_j$. In the complex plane, this problem is trivial (Hörmander 1990). In higher dimension spaces, the solution of $\bar{\partial}$ -equation is explicitly constructed and the regularity theory is also well-understood on the unit ball (Rudin 1980). Moreover, the study to this problem is completely established on strongly pseudoconvex domains which are most "beautiful" domains in several complex variables, see (Hörmander 1965), (Henkin 1969), (Henkin and Romanov 1971), (Romanov 1976). Recently, some existence and regularity results have been proved on certain analytic convex domains in $\mathbb{C}^n$, see (Bruma and Castillo 1984), (Ahn and Cho 2003), (Fornaess et al. 2011), (Khanh 2013), (Ha et al. 2014). On general domains, the solvability and regularity to the $\bar{\partial}$ -equation are still open.

In the lecture of Michael Range (Range 1978) given at the International Conferences Cortona, Italy, 1976–1977, he proved the following facts: on the smooth boundary convex domain of strict finite type m ($m=1,2,\ldots.$)

$$\Omega^m=\{(z_1,z_2)\in\mathbb{C}^2:|z_1|^{2m}+|z_2|^2-1<0\},\tag{1}$$

the Cauchy-Riemann equation $\bar{\partial}u=\varphi$ is solvable. Moreover, the solution u is Hölder continuous of order $\alpha\leq\frac{1}{2m}$ whenever φ is a (0, 1) $C^1(\Omega^m)$ -form. Here, $\bar{\partial}$ is defined in the sense of distributions.

Also in this important lecture, he showed that on the infinite type smooth boundary convex domain

$$\Omega^\infty=\{(z_1,z_2)\in\mathbb{C}^2:\exp(1+2/s).\exp\left(\frac{-1}{|z_1|^s}\right)+|z_2|^2-1<0\},\tag{2}$$

for $0<s<1$, the Cauchy-Riemann equation is although solvable, there is no solution which is Hölder continuous of any positive order. Hence, it is reasonable that we can conjecture if the $\bar{\partial}$ -equation is solvable in other Hölder class in some weak sense on Ω^∞ .

Recently, sup-norm estimates for the solution to the $\bar{\partial}$-equation on Ω^∞ have been established by Fornaess-Lee-Zhang in (Fornaess et al. 2011) and Khanh in (Khanh 2013). The main purpose in this paper is to give a positive and general answer to the above conjecture. The main method in this paper is based on a new proof in (Khanh 2013), (Ha et al. 2014).

2 MAIN RESULT

Let Ω be a bounded domain in $\mathbb{C}^2$ with smooth boundary $b\Omega$. Let ρ be a defining function for Ω, that means, ρ is a real value C^∞-function defined on a neighborhood of $b\Omega$ such that

$$\Omega = \{(z_1, z_2) : \rho(z_1, z_2) < 0\}$$

and $d\rho \neq 0$ on $b\Omega$. Then Ω is said to be (analytic) convex if

$$\sum_{j,k=1}^{2} \frac{\partial^2 \rho}{\partial x_j \partial y_k}(x, y) a_j a_k \geq 0 \quad \text{on } b\Omega,$$

for every $(a_1, a_2) \neq 0$ with $\sum_{j=1}^{2} a_j \frac{\partial \rho}{\partial x_j(y_j)}(x, y) = 0$ on $b\Omega$.

The Leray map on Ω is defined by

$$\Phi(\zeta, z) = \frac{\partial \rho}{\partial \zeta_1}(\zeta)(z_1 - \zeta_1) + \frac{\partial \rho}{\partial \zeta_2}(\zeta)(z_2 - \zeta_2)$$

for $\zeta \in b\Omega$. Since the convexity of Ω, $\text{Re}\left(\sum_{j=1}^{2} \frac{\partial \rho}{\partial \zeta_j}(z_j - \zeta_j)\right) \neq 0$ for $\zeta \in b\Omega$ and $z \in \Omega$ and so $\Phi(\zeta, z) \neq 0$ for all $(\zeta, z) \in b\Omega \times \Omega$. It is well-known that for each $\zeta \in b\Omega$, the complex hypersurface $\{\Phi(\zeta, z) = 0\}$ and the complex tangent space to $b\Omega$ at ζ are actually the same. Moreover, the Leray map has the following properties:

1. Φ is of C^1-class in (ζ, z).
2. $\Phi(\zeta, .)$ is holomorphic on Ω.
3. $|\Phi(\zeta, z)| \geq A > 0$ for $z \in \Omega$, $|z - \zeta| \geq c$ for some constant $c > 0$.

Definition 2.1. Let *Type* be a set of all smooth, increasing functions $F : [0, \infty) \to [0, \infty)$ such that

1. $F(0) = 0$;
2. $\int_0^\delta |\ln F(r^2)| \, dr < \infty$ for some $\delta > 0$;
3. $\frac{F(r)}{r}$ is increasing.

The convex domain Ω is called of admitting an maximal type $F \in Type$ at $P \in b\Omega$ if on the neighborhood $0 \leq |P - z| < c'$, for some $0 < c' \leq c$, we have

$$|\Phi(\zeta, z)| \gtrsim |\rho(z)| + |Im[\Phi(\zeta, z)]| + F(|z_1 - \zeta_1|^2), \quad (3)$$

for every $\zeta \in b\Omega \cap B(P, c)$, and $z \in \overline{\Omega}, |z - \zeta| < c$.

Here and in what follows, the notations $\lesssim$ and $\gtrsim$ denote inequalities up to a positive constant, and $\approx$ means the combination of $\lesssim$ and $\gtrsim$.

Remark 2.2.

1. The definition 2.1 is independent of the choice on holomorphic coordinates in a neighborhood of P and of the particular defining function ρ.
2. The domain Ω is called convex of maximal type F if it has these above properties at every point $P \in \Omega$, with the common function F. Actually, it follows that by compactness of $b\Omega$, we can choose the common function F for all boundary points $P \in b\Omega$.

Example 2.1

- *Let*

$$\Omega^m = \{(z_1, z_2) \in \mathbb{C}^2 : |z_1|^{2m} + |z_2|^2 - 1 < 0\}.$$

Then, Ω^m is convex of maximal type $F(t) = t^m$, see (Range 1978).

- *Let*

$$\Omega^\infty = \{(z_1, z_2) \in \mathbb{C}^2 : \exp(1 + 2/s).\exp\left(\frac{-1}{|z_1|^s}\right) + |z_2|^2 - 1 < 0\},.$$

Then, for $0 < s < 1$, Ω^∞ is convex of maximal type $F(t) = \exp(\frac{-1}{32.t^s})$, see (Verdera 1984).

Let f be an increasing function such that $\lim_{t \to \infty} f(t) = +\infty$, from (Khanh 2013), (Ha and Khanh 2015), we recall the "weak" f-Hölder space on Ω as

$$\Lambda^f(\Omega) = \{u : \|u\|_f := \|u\|_\infty + sup_{z, z+h \in \Omega} f(|h|^{-1}).|u(z+h) - u(z)| < \infty\}.$$

When $f(t) = t^\alpha$, for $0 < \alpha < 1$, we obtain the standard Hölder space $H^\alpha(\Omega)$.

The main results in this paper are follows.

Theorem 2.3. *Let Ω be a bounded convex domain in $\mathbb{C}^2$ with smooth boundary $b\Omega$. Let $F \in Type$ and assume that Ω is convex of maximal type F.*

Then, for every (0, 1) form φ whose coefficients belong to $L^\infty(\Omega)$ and $\partial\varphi = 0$ on Ω in the weak sense, there exists a function $u \in L^\infty(\Omega)$ such that

$$\bar{\partial} u = \varphi$$

in the weak sense and $\|u\|_\infty \lesssim \|\varphi\|_\infty$.

Moreover, we also have

Theorem 2.4. *Let Ω be a domain as in Theorem 2.3 and we define*

$$f(d^{-1}) := \left(\int_0^d \frac{\sqrt{F^*(t)}}{t} dt\right)^{-1},$$

where F^* is the inverse of F.

Then, for every (0, 1) *form φ whose coefficients belong to $L^\infty(\Omega)$ and $\bar\partial\varphi = 0$ on Ω in the weak sense, there exists a function $u \in \Lambda^f(\Omega)$ such that*

$$\bar\partial u = \varphi$$

in the weak sense and $\| u \|_f \lesssim \| \varphi \|_\infty$.

The proof of Theorem 2.3 is actually contained in the proof of Theorem 2.4 with more easier computations. Hence, we omit the details of the proof of Theorem 2.3.

3 PROOF OF THE MAIN RESULT

The proof of Theorem 2.4 is separated to two parts: The first one is to recall briefly the Henkin's construction for solutions to the $\bar\partial$-equation. For more general definitions and properties, we refer to the excellent book by Chen and Shaw (Chen and Shaw 2001). The second one is to estimate all integral terms in this construction.

In the following definitions, only the convexity of Ω is required for defining Φ.

Definition 3.1. (Homotopy Kernel for $\bar\partial$-solution on convex domains).

For $\lambda \in [0,1]$, let define:

- $w_j^0(\zeta, z) = \frac{\bar z_j - \bar\zeta_j}{|z-\zeta|^2}$ and $w_j^1(\zeta, z) = \frac{\partial\rho}{\partial\zeta_j}(\zeta) . \frac{1}{\Phi(\zeta, z)}$, for $j = 1,2$.
- $w_j(\zeta, \lambda, z) = (1-\lambda) w_j^0(\zeta, z) + \lambda w_j^1(\zeta, z)$, for $j = 1$, 2.
- $\omega_{2,0}(\zeta, \lambda, z) = (w_1 \bar\partial_{\zeta,\lambda} w_2 - w_2 \bar\partial_{\zeta,\lambda} w_1) \wedge d\zeta_1 \wedge d\zeta_2$, where $\bar\partial_{\zeta,\lambda} := \bar\partial_\zeta + d_\lambda$.
- $\omega_{2,1}(\zeta, \lambda, z) = -(w_1 \bar\partial_z w_2 - w_2 \bar\partial_z w_1) \wedge d\zeta_1 \wedge d\zeta_2$.

Let choose a differentiable triangulation $\{S_k : k = 1, \ldots, l\}$ of the boundary $b\Omega$, in which the simplices S_k being so small that the above constructions can be carried out for $\zeta \in S_k$, and the functions Φ and w_j^1 depending on the index k. Then, the forms $\omega_{2,0}^k$ and $\omega_{2,1}^k$ are defined as the restrictions of $\omega_{2,0}$ and $\omega_{2,1}$ on S_k.

Theorem 3.2. (Existence). *Let define the linear operator $T : C_{0,1}^1(\Omega) \to C^1(\Omega)$ as follows*

$$T\varphi = \frac{1}{4\pi^2}\left[\sum_{k=1}^{l} \int_{S_k \times [0,1]} \varphi \wedge \omega_{2,0}^k - \int_{\Omega \times [0,1]} \varphi \wedge \omega_{2,0}\right]. \tag{4}$$

Then, if $\bar\partial\varphi = 0$ on Ω, we have

$$\bar\partial(T\varphi) = \varphi \quad \text{on } \Omega.$$

Moreover,

$$\left\| \int_{\Omega \times [0,1]} \varphi \wedge \omega_{2,0} \right\|_f \lesssim \| \varphi \|_\infty$$

for any f with $0 < f(d^{-1}) < d^{-1}$.

For the proof of the above theorem, we refer the reader to (Range 1978), (Bruma and Castillo 1984), (Range 1986) or (Chen and Shaw 2001).

Now, in order to proof the *f*-Hölder estimate for the first integral in (4), we recall the following General Hardy-Littlewood Lemma proved by Khanh (Khanh 2013).

Lemma 3.3. *Let Ω be a bounded smooth domain in $\mathbb{R}^n$ and let ρ be a defining function of Ω. Let $G : \mathbb{R}^+ \to \mathbb{R}^+$ be an increasing function such that $\frac{G(t)}{t}$ is decreasing and $\int_0^d \frac{G(t)}{t} dt < \infty$ for $d > 0$ small enough. If $u \in C^1(\Omega)$ such that*

$$|\nabla u(x)| \lesssim \frac{G(|\rho(x)|)}{|\rho(x)|} \quad \textit{for every } x \in \Omega,$$

then

$$f(|x-y|^{-1})|u(x) - u(y)| < \infty$$

uniformly in $x, y \in \Omega$, $x \neq y$, and where $f(d^{-1}) := \left(\int_0^d \frac{G(t)}{t} dt\right)^{-1}$.

Hence, to complete the proof of Theorem 2.4, it is enough to prove the following result.

Proposition 3.4. *For the above definitions and notations, we have*

$$\int_{S_k \times [0,1]} |\nabla_z \omega_{2,0}^k(\zeta, \lambda, z)| \lesssim \frac{\sqrt{F^*(|\rho(z)|)}}{|\rho(z)|},$$

and $G(t) := \sqrt{F^*(t)}$ satisfies the hypothesis of Lemma 3.3, where F^* is the inverse of F.

Proof. For simplicity, we can drop the index *k*. From the definition 3.1, integrating in $\lambda \in [0,1]$, we have

$$\int_{S \times]0,1]} |\nabla_z \omega_{2,0}(\zeta, z)| \lesssim \int_{S \subset b\Omega} \left(\frac{1}{|\Phi(\zeta, z)| . |\zeta - z|^2} + \frac{1}{|\Phi(\zeta, z)|^2 . |\zeta - z|} \right) d\sigma(\zeta), \tag{5}$$

where $d\sigma$ is the surface measure of $b\Omega$.

Since $|\Phi(\zeta, z)| \geq A > 0$ for z fixed in Ω, $|z - \zeta| \geq c$ for some constant $c > 0$, it is enough to estimate the integral over $S \cap B(z, c)$. Based on Henkin's techniques, we re-introduce the following real coordinate system $t = (t', t_3) = (t_1, t_2, t_3) \in \mathbb{R}^2 \times [0, \infty)$

$$\begin{cases} t_1(z^*) = t_2(z^*) = 0, \quad \text{where } z^* \in b\Omega \\ \text{satisfies } |z - z^*| = dist(z, b\Omega), \\ t_3 = |\operatorname{Im} \Phi(\zeta, z)|, \end{cases}$$

such that $\mathrm{S}\cap B(z,c)\subset\{t:|t|\le R\}$ and $d\sigma(\zeta)|_{S\cap B(z,c)}\, dt_1dt_2dt_3$. The existence of such a coordinate system follow from the implicit function theorem and the convexity for $|p(z)|$ and c sufficiently small.

Since $|\zeta - z|\ge|t'|+|\rho(z)|$, we obtain

$$\begin{aligned}&\int_{(S\cap B(z,c))\times[0,1]}|\nabla_z\omega_{2,0}(\zeta,,\lambda,z)|\\ &\lesssim \underbrace{\int_{|t|\le R}\frac{dt_1dt_2dt_3}{(t_3+|\rho(z)|+F(|t'|^2))(|t'|+|\rho(z)|)^2}}_{:=I(|\rho(z)|)} \qquad (6)\\ &+\underbrace{\int_{|t|\le R}\frac{dt_1dt_2dt_3}{(t_3+|\rho(z)|+|F(t'|^2))^2|t'|}}_{:=II(|\rho(z)|)}.\end{aligned}$$

Some simple computations imply that

$$I_1(|\rho(z)|)\lesssim|\mathrm{n}(|\rho(z)|)|^2\lesssim\frac{G(|\rho(z)|)}{|\rho(z)|} \qquad (7)$$

for any G satisfying Lemma 3.3.

On the other hand, we also have

$$\begin{aligned}I_2(|\rho(z)|)&\lesssim\int_0^R\frac{1}{|\rho(z)|+F(r^2)}dr\\ &=\int_0^{\sqrt{F^*(|\rho(z)|)}}\frac{1}{|\rho(z)|+F(r^2)}dr\\ &+\int_{\sqrt{F^*(|\rho(z)|)}}^R\frac{1}{|\rho(z)|+F(r^2)}dr,\end{aligned}$$

Since $\frac{F(r)}{r}$ is increasing, we have

$$\frac{F(r^2)}{|\rho(z)|}\ge\frac{r^2}{F^*(|\rho(z)|)}\quad\text{for all } r\ge\sqrt{F^*(|\rho(z)|)}.$$

Hence,

$$\int_{\sqrt{F^*(|\rho(z)|)}}^R\frac{1}{|\rho(z)|+F(r^2)}dr\le\frac{\pi}{4}\frac{\sqrt{F^*(|\rho(z)|)}}{|\rho(z)|}.$$

It is easy to see that

$$\int_0^{\sqrt{F^*(|\rho(z)|)}}\frac{1}{|\rho(z)|+F(r^2)}dr\le\frac{\sqrt{F^*(|\rho(z)|)}}{|\rho(z)|}.$$

Thus,

$$I_2(|\rho(z)|)\lesssim\frac{\sqrt{F^*(|\rho(z)|)}}{|\rho(z)|}.$$

The last step in this part is to check the function $G(t):=\sqrt{F^*(t)}$ satisfies the conditions in Lemma 3.3.

Then, by (7) we have

$$I_1(z)+I_2(z)\lesssim\frac{\sqrt{F^*(|\rho(z)|)}}{|\rho(z)|},$$

and so $u\in\Lambda^f(\Omega)$ in which $f(d^{-1}):=\left(\int_0^d\frac{\sqrt{F^*(t)}}{t}dt\right)^{-1}$, for small $d>0$.

Now, since $\sqrt{F^*(t)}$ is increasing and $\frac{\sqrt{F^*(t)}}{t}$ is decreasing, for some small $R>0$, $|\ln(F(t^2))|$ is decreasing for all $0\le t\le R$. Thus, by the hypothesis (2) of F, we have

$$|\ln F(\eta^2)|\,\eta\le\int_0^\eta|\ln F(t^2)|\,dt\le\int_0^R|\ln F(t^2)|\,dt<\infty$$

for all $0\le\eta\le R$. As a consequence, $\sqrt{F^*(t)}\,|\ln t|$ is finite for all $0\le t\le\sqrt{F^*(R)}$ and $\lim_{t\to0}t|\ln F(t^2)|$ is zero. These facts and the second hypothesis of F imply

$$\begin{aligned}\int_0^d\frac{\sqrt{F^*(t)}}{t}dt&=\int_0^{\sqrt{F^*(d)}}y(\ln F(y^2))'dy\\ &=\sqrt{F^*(d)}\ln d-\int_0^{\sqrt{F^*(d)}}(\ln F(y^2))dy\\ &<\infty\end{aligned}$$

for $d>0$ small enough.

Hence, we have the conclusion that $u\in\Lambda^f(\Omega)$.

REFERENCES

Ahn, H. & H.R. Cho (2000). Optimal Hölder and L^p estimates for $\bar{\partial}_b$ on boundaries of convex domains of finite type. *J. Math. Anal. Appl.* 286(1), 281–294.

Bruma, J. & J. del Castillo (1984). Hölder and L^p-estimates for the $\bar{\partial}$ -equation in some convex domains with real-analytic boundary. *Math. Ann.* 296(4), 527–539.

Chen, S.C. & M.C. Shaw (2001). *Partial Differential Equations in Several Complex Variables.* AMS/IP, Studies in Advanced Mathematics, AMS.

Fornaess, J. E. & L. Lee, Y. Zhang (2011). On supnorm estimates for $\bar{\partial}$ on infinite type convex domains in $\mathbb{C}^2$. *J. Geom. Anal.* 21, 495–512.

Ha, L.K. & Khanh, T.V. & A. Raich (2014). L^p-estimates for the $\bar{\partial}$ -equation on a class of infinite type domains. *Int. J. Math.* 25, 1450106 [15pages].

Ha, L.K. & T.V. Khanh (2015). Boundary regularity of the solution to the Complex Monge-Ampère equation on pseudoconvex domains of infinite type. *Math. Res. Lett.* 22(2), 467–484.

Henkin, G.M. & A.V. Romanov (1971). Exact Hölder estimates for the solutions of the $\bar{\partial}$ -equation. *Math USSR Izvestija*, 5, 1180–1192.

Henkin, G.M. (1969). Integral representations of functions holomorphic in strictly-pseudoconvex domains and some applications. *Math. USSR Sbornik.* 7(4), 597–616.

Henkin, G.M. (1970). Integral representations of functions in strictly-pseudoconvex domains and applications to the $\overline{\partial}$ -problem. *Math. USSR Sbornik.* 11, 273–281.

Hörmander, L. (1965). L^2 estimates and existence theorems for the $\overline{\partial}$ operator. *Acta. Math.* 113, 89–125.

Hörmander, L. (1990). *An introduction to complex analysis in Several Complex Variables*. Third edition, Van Nostrand, Princeton, N. J.

Khanh, T.V. (2013). Supnorm and *f*-Hölder estimates for $\overline{\partial}$ on convex domains of general type in $\mathbb{C}^2$. *J. Math. Anal. Appl.* 430, 522–531.

Range, R.M. (1978). On the Hölder estimates for $\overline{\partial}u = f$ on weakly pseudoconvex domains, *Proc. Inter. Conf. Cortona, Italy 1976–1977. Scoula. Norm. Sup. Pisa,* 247–267.

Range, R.M. (1986). *Holomorphic Functions and Integral Representations in Several Complex Variables*. Springer-Vedag, Berlin/New York.

Romanov, A.V. (1976). A formula and estimates for solutions of the tangential Cauchy-Riemann equation. *Math. Sb.* 99, 58–83.

Rudin, W. (1980). *Function theory in the unit ball of* $\mathbb{C}^n$. Springer-Varlag, New York.

Verdera, J. (1984). L^∞-continuity of Henkin operators solving $\overline{\partial}$ in certain weakly pseudoconvex domains of $\mathbb{C}^2$. *Proc. Roy. Soc. Edinburgh,* 99, 25–33.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

A review on global and non-global existence of solutions of source types of degenerate parabolic equations with a singular absorption: Complete quenching phenomenon

D.N. Anh & K.H. Van
Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

J.I. Díaz
Instituto de Matemática Interdisciplinar, Universidad Complutense de Madrid, Madrid, Spain

ABSTRACT: We study the global and non-global existence of solutions of degenerate singular parabolic equation with sources. In the case of global existence, we prove that any solution must vanish identically after a finite time if either the initial data or the source term or the measure of domain is small enough.

1 INTRODUCTION

This paper is to study the nonnegative solutions of the source types of one dimensional degenerate parabolic equations with a singular absorption

$$\begin{cases} \partial_t u - (|u_x|^{p-2} u_x)_x + u^{-\beta}\mathcal{X}_{\{u>0\}} = & \begin{array}{l} f(u,x,t) \\ \text{in } I \times (0,\infty), \end{array} \\ u(x_1,t) = u(x_2,t) = 0 & t \in (0,\infty), \\ u(x,0) = u_0(x) & \text{in } I, \end{cases} \quad (1)$$

where $I = (x_1, x_2)$ is an open bounded interval in $\mathbb{R}$, $\beta \in (0, 1)$, $p > 2$, and $\mathcal{X}_{\{u>0\}}$ denotes the characteristic function of the set of points (x, t) where $u(x, t) > 0$, i.e

$$\mathcal{X}_{\{u>0\}} = \begin{cases} 1, & \text{if } u > 0, \\ 0, & \text{if } u \leq 0. \end{cases}$$

Note that the absorption term $u^{-\beta}\mathcal{X}_{\{u>0\}}$ becomes singular when u is near to 0, and we impose $u^{-\beta}\mathcal{X}_{\{u>0\}} = 0$ whenever $u = 0$. Through this paper, we always as that $f : \mathbb{R} \times \overline{I} \times [0,\infty) \to \mathbb{R}$ is a nonnegative function satisfying the following hypothesis

$$(H)\begin{cases} f \in \mathcal{C}^1(\mathbb{R} \times \overline{I} \times [0,\infty)), \textit{ and } f(0,x,t) = 0 \\ \qquad\qquad\qquad\qquad \forall (x,t) \in I \times (0,\infty). \\ \text{There is a nonnegative real funcation } h \in \mathcal{C}^1(\mathbb{R}) \\ \text{such that } f(u,x,t) \leq h(u),\ \forall (x,t) \in I \times (0,\infty). \end{cases}$$

In the case N -dimension and $p = 2$, equation (1) becomes

$$\begin{cases} \partial_t u - \Delta u + u^{-\beta}\mathcal{X}_{\{u>0\}} = f(u,x,t) & \text{in } \Omega \times (0,\infty), \\ u = 0 & \partial\Omega \times (0,\infty), \\ u(x,0) = u_0(x) & \text{in } \Omega, \end{cases} \quad (2)$$

Problem (2) can be considered as a limit of mathematical models describing enzymatic kinetics (see (Banks 1975)), or the Langmuir-Hinshelwood model of the heterogeneous chemical catalyst (see, e.g. (W. Strieder 1973) p. 68, (Díaz 1985), (Phillips 1987) and references therein). This case was studied in (Phillips 1987), (Kawohl 1996), (Levine 1993), (Dávila & Montenegro 2004), (Winkler 2007), and so forth. These authors focused on studying the existence of solution, and the behaviors of solutions. For example, D. Phillips (Phillips 1987) proved the existence of solution for the Cauchy problem associating (2) in the case $f = 0$. He also showed that any solution must quench after a finite time.

In (Dávila & Montenegro 2004), J. Davila and M. Montenegro proved the existence of solution of equation (2) if the source term $f(u)$ is sub-linear, i.e: $|f(u)| \leq C(u + 1)$, for $u \geq 0$. Moreover, they also showed that the measure of the set $\{(x, t) \in \Omega \times (0, \infty): u(x, t) = 0\}$ is positive. In other words, the solution may exhibit the quenching behavior. Still in the sub-linear case, M. Montenegro (Montenegro 2011) considered equation (2) with the source term $\lambda.f(u)$ instead of $f(u, x, t)$. He showed that there exists a positive real number λ_0 so that if $\lambda \in (0, \lambda_0)$, then any solution must vanish identically after a finite time.

Recently, problem (1) in N-dimension was considered by Giacomoni et al., (Giacomoni, Sauvy, & Shmarev 2014), with the source term $f(u, x)$ satisfying a natural growth condition, i.e:

$$(H_1)\; 0 \le f(u,x) \le \lambda . u^{q-1} + v,$$

with λ, $v \ge 0$, and $q \ge 1$. These authors proved first a local existence result. Unfortunately, their proof of local existence of solution is not correct. Then, our first purpose is to prove the local existence of solution of equation (1), even for a more general class of functions $f(u, x, t)$ satisfying (H) instead of (H_1) in (Giacomoni, Sauvy, & Shmarev 2014).

For example, the function $f(u,x,t) = \frac{|x|^2}{t+1}(e^u - 1)$ satisfies (H). But, it does not satisfy any natural growth conditions in (H_1).

We note that the assumption $f(0, x, t) = 0$ in (H) is a necessary condition for the existence of solution. If this one is violated, then equation (1) may have no solution. For instance, we will show at the end that equation (1) has no weak solution if $f(u) = \lambda . u^{q-1} + v$, for $v > 0$. Thus, this assumption seems to be the sufficient condition for the existence of solution for a particular class of functions $f(u, x, t)$ satisfying a certain growth condition.

The second purpose of this article is to study the existence and nonexistence of global solution of equation (1) for the case where f satisfies a natural growth condition (H_1). Let us first remind some classical results for the global and non-global existence of solution of equation (1) without the singular absorption term:

$$\begin{cases} \partial_t u - (|u_x|^{p-2} u_x)_x = f(u,x,t) & \text{in } I \times (0,\infty), \\ u(x_1,t) = u(x_2,t) = 0 & t \in (0,\infty), \\ u(x,0) = u_0(x) & \text{in } I, \end{cases} \tag{3}$$

For a simple introduction, we only discuss the case: $f(u, x, t) = \lambda . u^{q-1}$, $q > 1$, $\lambda > 0$. For a more general class of f, we refer to (Levine 1990), (Zhao 1993), (Galaktionov 1994), (Galaktionov & Vazquez 2002), and references therein.

In (Tsutsumi 1973), M. Tsutsumi proved that if $q < p$, then problem (3) has global nonnegative solutions whenever initial data u_0 belongs to some Sobolev space. The case $q \ge p$ is quite delicate that there are both nonnegative global solutions, and solutions which blow up in a finite time. Indeed, J. N. Zhao (Zhao 1993) showed that when $q \ge p$, equation (3) has a global solution if the measure of I is small enough, and it has no global solution if the measure of I is large enough. The fact that the first eigenvalue of $-\Delta_p$ (denoted as λ_I) decreases with increasing domain can be also used as an intuitive explanation for Zhao's result. In the critical case $q = p$, Y. Li and C. Xie (Li & Xie 2003) showed that if $\lambda_I > \lambda$, equation (3) has then a unique global bounded solution. While, the unique solution of equation (3) blows up in a finite time if provided $\lambda_I < \lambda$. We also note that the unique solution is globally bounded if provided that $\lambda_I = \lambda$ and initial data $u_0(x) \le \kappa . \phi_I(x)$, for some $\kappa > 0$, and ϕ_I is the eigenfunction corresponding to λ_I.

Roughly speaking, such a weak solution of equation (1) is a sub-solution of equation (3). Thus, the strong comparison theorem implies that the global existence result holds for equation (1) if provided either $q < p$, or $q \ge p$ and u_0 (resp. λ, the measure of I) is small enough. By this observation, we will show that any weak solution of equation (1) exists globally if provided that either

i. $q < p$, or
ii. $q \ge p$ and u_0 (resp. the measure of I, λ) is small enough, or
iii. $q = p$, and $|\lambda_I - \lambda|$ is sufficiently small.

Note that the result of (*iii*) is new because the solutions exist globally even $\lambda > \lambda_I$, while the unique solution of equation (3) blows up whenever $\lambda > \lambda_I$ (compare to Theorem 2.2 and Theorem 3.1, (Giacomoni, Sauvy, & Shmarev 2014)).

The conclusion (*iii*) can be explained as follows. As mentioned above, we will prove an estimate for $|u_x|$ involving a certain power of u.

$$|u_x(x,t)|^p \le C.u^{1-\beta}(x,t), \quad \text{for a.e } (x,t) \in I \times (0,T). \tag{4}$$

Intuitively, inequality (4) says that the absorption $u^{-\beta}\mathcal{X}_{\{u>0\}}$ strengthen the diffusion term to against the effect of the source term. By this reason, the global existence result can be extended to the case: $\lambda > \lambda_I$, and $0 < \lambda - \lambda_I$ is small. At the end, we will provide some numerical experiences in order to illustrate the difference between solutions of both equations (1) and (3).

The final goal of this paper is to consider the quenching phenomenon of solutions of equation (1), that nonnegative solution is extinct after a finite time. As already known, in the case $p = 2, f \equiv 0$, any weak nonnegative solution of equation (1) vanishes identically after a finite time, even beginning with a large initial data, see e.g (Phillips 1987), (Dao, Diaz, & Sauvy 2016), (Winkler 2007), (Dávila & Montenegro 2004), and references therein. This property arises due to the presence of the singular term $u^{-\beta}\mathcal{X}_{\{u>0\}}$.

For the case $f(u) = \lambda . u^{q-1}$, Giacomoni et al. showed that the quenching phenomenon occurs if $q \le p$, and $\lambda_I > \lambda$, see Theorem 2:2, (Giacomoni, Sauvy, & Shmarev 2014). Their argument is based on the observation that the diffusion term dominates the source term $f(u)$ in these cases (see also (Montenegro 2011) for the case $p = 2$). However, this argument is no longer applicable to the remains, such as $q = p$ and $\lambda_I \le \lambda$; or $q > p$. Thus, we are interested in the following question that whether or not the quenching phenomenon occurs for the remain cases. Our answer is positive under the additional conditions on u_0, λ_I, or λ. Then, a brief of our quenching results is as follows:

Any weak solution of equation (1) must vanish identically after a finite time if provided either

a. $q \geq p$, and λ is small enough (Note that u_0 can be large in this case); or
b. $q \geq p$, and $\| u_0 \|_{L^\infty(I)}$ (resp. the measure of I) is small enough; or
c. $q = p$, and $|\lambda - \lambda_I|$ is small enough.

The conclusion (a) means that the source term $\lambda.u^{q-1}$ is so small that this perturbation does not effect so much to the quenching property of solutions of equation (1). A simulation result at the end will illustrate the above result.

2 PRELIMINARY AND MAIN RESULTS

At the beginning, let us introduce the notion of a weak solution of equation (1).

Definition 1 *Given* $0 \leq u_0 \in W_0^{1,p}(I)$. *A function* $u \geq 0$ *is called a weak solution of equation (1) if* $f(u, x, t)$, $u^{-\beta} \mathcal{X}_{\{u>0\}} \in L^1(\mathrm{I} \times (0, T))$, *and* $u \in L^p(0, T; W_0^{1,p}(I)) \cap L^\infty(\overline{I} \times (0, T)) \cap \mathcal{C}([0, \mathrm{T}); L^1(I)$ *satisfies equation (1) in the sense of distributions* $\mathcal{D}'(I \times (0, \infty))$, *i.e,*

$$\int_0^\infty \int_I (-u\phi_t + |u_x|^{p-2} u_x\phi_x + u^{-\beta}\mathcal{X}_{\{u>0\}}\phi) + f(u,x,t)\phi)dxdt = 0, \forall \phi \in \mathcal{C}_c^\infty(I \times (0,\infty)) \tag{5}$$

Note that $u_0 \in C^{0,\alpha}(I)$, with $\alpha = 1 - \frac{1}{p}$, since the Sobolev imbedding. Then, we have the local existence theorem.

Theorem 2 *Let* $0 \leq u_0 \in W_0^{1,p}(I)$, *and f satisfy* (*H*). *Then, there exists a time* $T_0 > 0$ *so that equation* (1) *has a maximal weak solution u in* $I \times (0, T_0)$. *Moreover, u satisfies the following estimate*

$$|u_x(x,t)| \leq C.u^{1-\frac{1}{\gamma}}(x,t)\left(t^{-\frac{1}{p}}\Gamma^{\frac{1+\beta}{p}}(T_0) + \Gamma^{\frac{1+\beta}{p}}(T_0).\Lambda_1(D_u f) + \Gamma^{\frac{1+\beta\gamma}{p\gamma}}(T_0).\Lambda_2(D_x f) + 1 \right), \tag{6}$$

for a.e $(x, t) \in I \times (0, T_0)$, *where* Γ *is the flat solution satisfying the ordinary differential equation:*

$$\Gamma_t = h(\Gamma), \quad \textit{and } \Gamma(0) = \| u_0 \| \infty.$$

And

$$\begin{cases} \Lambda_1(D_u f) = \max\limits_{0 \leq u \leq \Gamma(T_0),\,(x,t) \in \overline{I} \times [0,T_0]} |D_u f(u,x,t)|^{\frac{1}{p}}, \\ \Lambda_2(D_x f) = \max\limits_{0 \leq u \leq \Gamma(T_0),\,(x,t) \in \overline{I} \times [0,T_0]} |D_x f(u,x,t)|^{\frac{1}{p}}. \end{cases}$$

As a consequence of (6), for any $\tau > 0$ *there is a positive constant* $C = C(\beta, p, \tau)$ *such that*

$$|u(x,t) - u(y,s)| \leq C\left(|x-y| + |t-s|^{\frac{1}{3}} \right), \quad \forall x, y \in \overline{I}, \quad \forall t, s \geq \tau. \tag{7}$$

Next, let us denote by ϕ_J and λ_J the first nonnegative normalized eigenfunction and the first eigenvalue of the problem

$$\begin{cases} -\partial_x(|\partial_x \phi_J|^{p-2}\partial_x \phi_J) = \lambda_J \phi_J^{p-1} \text{ in } J, \\ \qquad\qquad J = (l_1, l_2) \subset\subset \mathbb{R}, \\ \phi_J(l_1) = \phi_J(l_2) = 0. \end{cases}$$

It is well known that the formula of the first eigenvalue (see (R. L. Biezuner & Martins 2009)) is

$$\lambda_J = (p-1)\left(\frac{\pi_p}{l_2 - l_1}\right)^p, \text{ with } \pi_p = 2\frac{\pi / p}{\sin(\pi / p)}. \tag{8}$$

As mentioned above, equation (1) has no solution if $f(u) = \lambda.u^{q-1} + v$, for some $v > 0$. By this reason, we only consider $f(u) = \lambda.u^{q-1}$ for the theorems below. Then, we first have the global existence result when $\| u_0 \|_{L^\infty(I)}$ is small.

Theorem 3 *Given* $\lambda > 0$, *and* $q \geq p$. *Let* $f(u) = \lambda.u^{q-1}$. *Assume that* $u_0 \in W_0^{1,p}(I)$ *such that* $\| u_0 \|_{L^\infty(I)}$ *is small enough. Then, the weak solutions of equation* (1) *are globally bounded. Moreover, they vanish identically after a finite time.*

Next, we have the global existence result if λ (resp. the measure of I) is small enough.

Theorem 4 *Given* $u_0 \in W_0^{1,p}(I)$, *and* $q \geq p$. *Let* $f(u) = \lambda.u^{q-1}$. *Assume that* λ *(resp. the measure of I) is small enough. Then, the weak solutions of equation* (1) *are globally bounded. Moreover, they vanish identically after a finite time.*

Particularly, we have the following result for the critical case $q = p$.

Theorem 5 *Given* $u_0 \in W_0^{1,p}(I)$, *and* $q = p$. *Let* $f(u) = \lambda.u^{q-1}$. *Assume that* $|\lambda_I - \lambda|$ *is small enough. Then, the weak solutions of equation* (1) *is globally bounded. Moreover, they vanish identically after a finite time.*

Remark 6 *By the comparison principle, the conclusions of Theorem 3, Theorem 4 and Theorem 5 still hold if f satisfies (H) and* $f(u, x, t) \leq \lambda.u^{q-1}$.

Concerning the non-global existence of solutions of equation (1), we first remind a result of (Giacomoni, Sauvy, & Shmarev 2014).

Proposition 7 *Let* $f(u) = \lambda.u^{q-1}$. *Let* $q > p$, *and* $u_0 \in W_0^{1,p}(I)$. *Assume that* $E(0) < 0$ *with*

$$E(t) = \int_I \left(\frac{1}{p} |u_x(t)|^p + \frac{1}{1-\beta} u^{1-\beta}(t) - \frac{\alpha}{q} u^q(t) \right) dx.$$

Then every solution of equation (1) *blows up in a finite time.*

In the critical case $q = p$, we show that the maximal solution u cannot be globally bounded if provided $E(0) \leq 0$.

Theorem 8 *Let and* $u_0 \in W_0^{1,p}(I)$, *and* $q = p$. *Let* $f(u) = \lambda.u^{q-1}$. *Assume that* $E(0) \leq 0$. *Then, the solution u cannot be globally bounded.*

3 SIMULATION RESULTS

In this part, we will illustrate our theoretical results with some numerical experiences. In the sequel, we consider equation (1) and equation (3) for the case: $q = p = 2.3$, $\beta = 0.8$, $I = (0, L)$, and $u_0(x) = x(L - x)$, and $f(u) = \lambda.u^{q-1}$.

We fix $L = 3.1273$. It follows then from (8) that $\lambda_I = 0.9999$.

With $\lambda = 1 > \lambda_I$ (just a little bit difference), the unique solution of equation (3) blows up after $t = 4286$, see Figure 1. While $\lambda = 1.269$, the maximal solution of equation (1) vanishes after $t = 7.6$, see Figure 2.

With $\lambda = 1.270$, the maximal solution of equation (1) blows up at $t = 23$, see Figure 3. Intuitively, the absorption $u^{-\beta}\ \mathcal{X}_{\{u>0\}}$ supports the nonlinear diffusion an amount $\lambda_{0.}u^{p-1}$, with $\lambda_0 = 1.269 - 0.9999 = 0.2691$. By this reason, for any $\lambda \in (0, 1.269)$, the solutions of equation (1) exist globally and they vanish after a finite time.

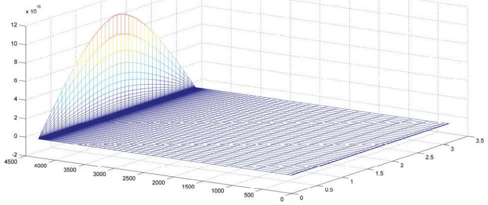


Figure 1. Evolution of the unique solution of equation (3).

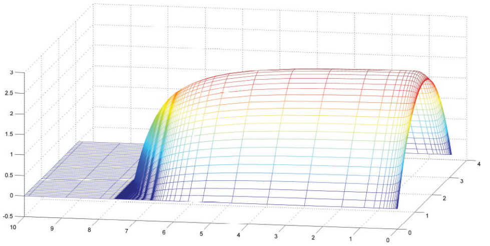


Figure 2. Evolution of the maximal solution of equation (1).

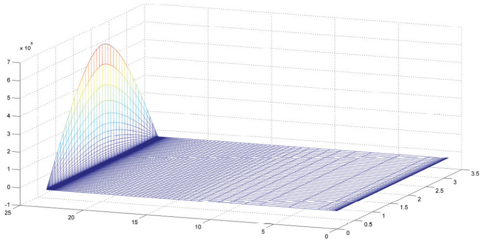


Figure 3. Evolution of the maximal solution of equation (1).

REFERENCES

Aris, R. (1975). *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts*. Oxford University Press.

Bandle, C. & C.M. Brauner (1986). Singular perturbation method in a parabolic problem with free boundary. *Boole Press Conf. Ser. 8*, 7–14.

Banks, H.T. (1975). Modeling and control in the biomedical sciences. lecture notes in biomathematics. *Springer-Verlag, Berlin-New York 6*.

Biezuner, R.L., G.E. & E.M. Martins (2009). Computing the first eigenvalue of the *p*-laplacian via the inverse power method. *Funct. Anal. 257*, 243–270.

Boccardo, L. & F. Murat (1992). Almost everywhere convergence of the gradients of solutions to elliptic and parabolic equations. *Nonlinear Anal. Theory, Methods and Applications 19(6)*, 581–597.

Boccardo, L. & T. Gallouet (1989). Nonlinear elliptic and parabolic equations involving measure data. *Funct. Anal. 87*, 149–169.

Coddington, E. & N. Levinson (1955). *Theory of Ordinary Differential Equations*. New York: McGraw-Hill.

Dao, A.N. & J.I. Diaz. A gradient estimate to a degenerate parabolic equation with a singular absorption term: global and local quenching phenomena. *To appear Jour. Math. Anal. Appl.*.

Dao, A.N., J.I. Diaz, & P. Sauvy (2016). Quenching phenomenon of singular parabolic problems with l^1 initial data. *In preparation*.

Dávila, J. & M. Montenegro (2004). Existence and asymptotic behavior for a singular parabolic equation. *Transactions of the AMS 357*, 1801–1828.

Díaz, J.I. (1985). Nonlinear partial differential equations and free boundaries, research notes in mathematics. *Pitman 106*.

Fila, M. & B. Kawohl (1990). Is quenching in infinite time possible. *Q. Appl. Math. 48(3)*, 531–534.

Galaktionov, V.A. & J.L. Vazquez (2002). The problem of blow-up in nonlinear parabolic equations. *Discrete and continuous dynamical systems 8*, 399–433.

Galaktionov, V.A. (1994). Blow-up for quasilinear heat equations with critical fujita's exponents. *Proceedings of the Royal Society of Edinburgh 124A*, 517–525.

Giacomoni, J., P. Sauvy, & S. Shmarev (2014). Complete quenching for a quasilinear parabolic equation. *J. Math. Anal. Appl. 410*, 607–624.

Herrero, M.A., J.L.V. (1982). On the propagation properties of a nonlinear degenerate parabolic equation. *Comm. in PDE 7(12)*, 1381–1402.

Kawohl, B. & R. Kersner (1992). On degenerate diffusion with very strong absorption. *Mathematical Methods in the Applied Sciences 7(15)*, 469–477.

Kawohl, B. (1996). Remarks on quenching. *Doc. Math., J. DMV 1*, 199–208.

Ladyzenskaja, O.A., V.A.S. & N.N. Uralceva (1988). *Linear and Quasi-Linear Equations of Parabolic Type*. AMS 23.

Levine, H.A. (1990). The role of critical exponents in blowup theorems. *SIAM Review 32(2)*, 262–288.

Levine, H.A. (1993). Quenching and beyond: a survey of recent results. nonlinear mathematical problems in industry ii. *Internat. Ser. Math. Sci. Appl. 2*, 501–512.

Li, Y. & C. Xie (2003). Blow-up for p-laplacian parabolic equations. *Electronic Jour. Diff. Equa. 20*, 1–12.

Montenegro, M. (2011). Complete quenching for singular parabolic problems. *J. Math. Anal. Appl. 384*, 591–596.

Ph. Benilan, J.I.D. (2004). Pointwise gradient estimates of solutions of one dimensional nonlinear parabolic problems. *Evolution Equations 3*, 557–602.

Phillips, D. (1987). Existence of solutions of quenching problems. *Appl. Anal. 24*, 253–264.

Strieder, W., R.A. (1973). *Variational Methods Applied to Problems of Diffusion and Reaction*. Berlin: Springer-Verlag.

Tsutsumi, M. (1972–1973). Existence and nonexistence of global solutions for nonlinear parabolic equations. *Publ. RIMS, Kyoto Univ. 8*, 211–229.

Winkler, M. (2007). Nonuniqueness in the quenching problem. *Math. Ann. 339*, 559–597.

Zh. Q. Wu, J. N. Zhao, J. X. Y. & H. L. Li (2001). *Nonlinear Diffusion Equations*. World Scientific, Singapore.

Zhao, J.N. (1993). Existence and nonexistence of solutions for $u_t = div(|\nabla u|^{p-2}\nabla u) + f(\nabla u, u, x, t)$. *J. Math. Anal. Appl. 172*, 130–146.

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

A spectral decomposition in vector spherical harmonics for Stokes equations

M.-P. Tran
Faculty of Mathematics—Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

T.-N. Nguyen
Department of Mathematics, HCMC University of Education, Ho Chi Minh City, Vietnam

ABSTRACT: The main goal of this paper is to present a spectral decomposition of velocity and pressure fields of the Stokes equations outside a unit ball. These expansions bases on the basis of vector spherical harmonics. Moreover, we show that this basis diagonalises the Neumann to Dirichlet operator.

1 INTRODUCTION

We consider the Stokes problem in the domain $\Omega_0 \cup B(0,1)$ where $\Omega_0 := \mathbf{R}^3 \setminus \overline{B(0,1)}$. Given a velocity field $\mathbf{g}$ defined on $S^2 := \partial B(0,1)$, we seek the velocity and pressure fields $(\mathbf{u}, p)$ satisfying

$$\begin{cases} -\Delta \mathbf{u} + \nabla p = 0 & \text{in } \Omega_0 \cup B(0,1), \\ \nabla \cdot \mathbf{u} = 0 & \text{in } \Omega_0 \cup B(0,1), \\ \mathbf{u} = \mathbf{g} & \text{on } S^2. \end{cases} \tag{1.1}$$

The well-posedness and regularity results of this equation can be found in the book of Galdi (Galdi 1994). If $(\mathbf{u}, p)$ is sufficiently smooth, the surface density of forces applied by the boundary of $B(0, 1)$ is defined by

$$\mathrm{f} = \left(\nabla \mathrm{u} + \nabla \mathrm{u}^T - pId\right) \cdot \mathrm{n}.$$

The Dirichlet to Neumann operator $\mathcal{DN}$ can be defined as

$$\mathcal{DN}g = \mathrm{f} \in W^{-1/2,2}(S^2, \boldsymbol{R}^3),\ g \in W^{1/2,2}(S^2, \boldsymbol{R}^3),$$

where $W^{1/2,2}(S^2, \boldsymbol{R}^3), W^{-1/2,2}(S^2, \boldsymbol{R}^3)$ define the fractional Sobolev and its dual space. This operator is a continuous linear isomorphism. Its inverse is called the Neumann to Dirichlet operator $\mathcal{ND}$ which is defined by the convolution between the Green function with the surface force density

$$\mathcal{ND}\ \mathrm{f} := G * \mathrm{f},$$

The Green function G is given by

$$G(\boldsymbol{r}) = \frac{1}{8\pi r}\left(Id + e_r \otimes e_r\right).$$

We refer the readers to (Nguyen 2013) for more detail properties of these operators. The main goal of this paper is to give a spectral decomposition of Neumann to Dirichlet operator in basis of vector spherical harmonics.

The sequel of this paper is organized as follows. In the next section, we describe the basis of vector spherical harmonics. We follow the notation in (Nédélec 2001) where these objects are introduced in the context of electromagnetism. In Section 3, we present the decomposition of the solution of the Stokes problem. Eventually, using the decomposition of the velocity and pressure field, we obtain an expansion of the corresponding Dirichlet to Neumann operator $\mathcal{DN}$ in vector spherical harmonics.

2 VECTOR SPHERICAL HARMONICS

Let us recall the definition and some properties of vector spherical harmonics. We consider the unit sphere S^2 in $\mathbf{R}^3$. The case of a sphere of arbitrary radius follows by a change of scale. In this geometry, it is natural to define a point of $\mathbf{R}^3$ by its spherical coordinates (r, θ, φ), where r is the radius and θ, φ the two Euler angles. These coordinates are related to the euclidean coordinates (x_1, x_2, x_3) by

$$\begin{cases} x_1 = r\sin\theta\cos\varphi, \\ x_2 = r\sin\theta\sin\varphi, \\ x_3 = r\cos\theta. \end{cases}$$

In these coordinates, the surface gradient of the function **u**, denoted $\nabla_{S^2}\mathbf{u}$, is defined as

$$\nabla_{S^2}\mathbf{u} = \frac{1}{\sin\theta}\frac{\partial\mathbf{u}}{\partial\varphi}\vec{e}_{\varphi} + \frac{\partial\mathbf{u}}{\partial\theta}\vec{e}_{\theta}, \tag{2.1}$$

where $\vec{e}_r, \vec{e}_\theta$ and $\vec{e}_\varphi$ are the unitary vectors. Let $H^1(S^2)$ denotes the Hilbert space

$$\mathrm{H}^1(S^2) = \left\{\mathbf{u} \in L^2(S^2, \mathbf{R}) : \nabla_{S^2}\mathbf{u} \in L^2(S^2, \mathbf{R}^3)\right\},$$

with its hermitian product

$$(\mathbf{u},\mathbf{v})_{H^1(S^2)} = \tfrac{1}{4}\int_{S^2} \mathbf{u}\bar{\mathbf{v}}d\sigma + \int_{S^2} \nabla_{S^2}\mathbf{u}\cdot\nabla_{S^2}\bar{\mathbf{v}}d\sigma.$$

We will denote by Δ_{S^2} the Laplace-Beltrami operator on the unit sphere S^2, defined as

$$\Delta_{S^2}\mathbf{u} = \frac{1}{\sin^2\theta}\frac{\partial^2\mathbf{u}}{\partial\varphi^2} + \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\mathbf{u}}{\partial\theta}\right).$$

The Laplace-Betrami operator is self-adjoint in the space $L^2(S^2)$ and it is coercive on the space $H^1(S^2) \cap L_0^2(S^2)$. It admits a family of eigenfunctions which constitutes an orthogonal Hilbert basis of the space $L^2(S^2)$. This basis is also orthogonal for the scalar product in $H^1(S^2)$. These eigenfunctions are called spherical harmonics. They are described in Theorem 2.1.

Let $\mathcal{H}_l$ be the space of homogeneous polynomials of degree l in three variables that are moreover harmonic in $\mathbf{R}^3$. Let $\mathcal{Y}_l$ be the space of the restrictions to the unit sphere S^2 of polynomials in $\mathcal{H}_l$.

Theorem 2.1. ((Nédélec 2001)). *Let $Y_l^m, |m| \le l$, denote an orthonormal basis of $\mathcal{Y}_l$ for the hermitian product of $L^2(S^2)$. The functions Y_l^m, for $l \ge 0$ and $|m| \le l$, form an orthogonal basis in $L^2(S^2)$, which is also orthogonal in $H^1(S^2)$. Moreover, $\mathcal{Y}_l$ coincides with the subspace spanned by the eigenfunctions of the Laplace-Beltrami operator associated with the eigenvalue $-l(l+1)$, i.e.,*

$$\Delta_{S^2}Y_l^m + l(l+1)Y_l^m = 0.$$

The eigenvalue $-l(l+1)$ has multiplicity $2l+1$.

By Theorem 2.1 and the Green's formula, we have

$$\left\|\nabla_{S^2}Y_l^m\right\|_{L^2}^2 = l(l+1).$$

We consider the Legendre polynomial $\mathbb{P}_l$:

$$\mathbb{P}_l(x) = \frac{(-1)^l}{2^l l!}\frac{d^l}{dx^l}(1-x^2)^l, x \in [-1,1].$$

The associated Legendre functions $\mathbb{P}_l^m$ are given by

$$\mathbb{P}_l^m = (-1)^m(1-x^2)^{m/2}\frac{d^m}{dx^m}\mathbb{P}_l(x).$$

The spherical harmonics of order l are the $2l+1$ functions as follows: for $l \ge 0, -l \le m \le l$

$$Y_l^m(\theta,\varphi) = \sqrt{2}C_l^m P_l^m(\cos\theta)\cos(m\varphi), \text{ if } m > 0,$$
$$Y_l^m(\theta,\varphi) = \sqrt{2}C_l^m P_l^{|m|}(\cos\theta)\sin(|m|\varphi), \text{ if } m < 0,$$
$$Y_l^m(\theta,\varphi) = C_l^m P_l^0(\cos\theta), \text{ if } m = 0,$$

where

$$C_l^m = \sqrt{\frac{(2l+1)}{4\pi}\frac{(l-|m|)!}{(l+|m|)!}}.$$

For $x \in S^2$, we respectively define $T_{l,m}, I_{l,m}, N_{l,m}$ as the traces on S^2 of the harmonic polynomials

$$T_{l,m}(x) = \nabla_{S^2}Y_l^m(x) \wedge x =: \nabla_{S^2}^{\perp}Y_l^m(x) \in TS^2,$$
$$I_{l,m}(x) = \nabla_{S^2}Y_{l+1}^m(x) + (l+1)Y_{l+1}^m(x)x,$$
$$N_{l,m}(x) = -\nabla_{S^2}Y_{l-1}^m(x) + lY_{l-1}^m(x)x.$$

Notice that by construction the components of $T_{l,m}, I_{l,m}, N_{l,m}$ belong to $\mathcal{Y}_l$, that is

$$\Delta_{S^2}Y + l(l+1)Y = 0, \quad \text{for } Y = T_{l,m}, I_{l,m}, N_{l,m}.$$

Using the tangential gradient defined by (2) and the Euler relation for the normal derivatives, we obtain

Theorem 2.2 ((Nédélec 2001)). *For each $l \ge 0$, the family $\left\{(T_{l,m})_{|m|\le l}; (I_{l,m})_{|m|\le l+1}; (N_{l,m})_{|m|\le l-1}\right\}$ forms an orthogonal basis of $H^1(S^2)$ and of $L^2(S^2, \mathbf{R}^3)$. Further, they satisfy*

$$\int_{S^2} |T_{l,m}(x)|^2\, d\sigma = l(l+1),$$
$$\int_{S^2} |I_{l,m}(x)|^2\, d\sigma = (l+1)(2l+3),$$
$$\int_{S^2} |N_{l,m}(x)|^2\, d\sigma = l(2l-1).$$

Let $\mathbf{u} \in L^2(S^2, \mathbf{R}^3)$, then **u** decomposes as

$$\mathbf{u}(x) = \sum_{l\ge1}\sum_{m=-l}^{l} i_{l,m}T_{l,m}(x) + \sum_{l\ge0}\sum_{m=-l-1}^{l+1} j_{l,m}I_{l,m}(x) + \sum_{l\ge1}\sum_{m=-l+1}^{l-1} k_{l,m}N_{l,m}(x).$$

For simplicity, we use the symbol Σ instead of using $\Sigma_{l\geq 1}\ \Sigma_{m=-l}^{l}$, $\Sigma_{l\geq 0}^{l}\ \Sigma_{m=-l-1}^{l+1}$ and $\Sigma_{l\geq 1}\ \Sigma_{m=-l+1}^{l-1}$.

3 DECOMPOSITION OF VELOCITY AND PRESSURE FIELD

The regularity results of Stoke equation are classical and can be found in the book of (Galdi 1994). The main point of this section is to establish the decomposition of the velocity and pressure field in vector spherical harmonics.

Theorem 3.1. *Let* $\mathbf{g} \in W^{1/2,2}(S^2, \mathbf{R}^3)$ *such that* $\int_{S^2} \mathbf{g}\cdot\mathbf{n} = 0$ *and let* $(\mathbf{u}, p)$ *be the variational solution of* (1.1) ($\mathbf{u} \in \mathcal{D}(\Omega_0 \cup B(0,1))$, $p \in L^2(\Omega_0 \cup B(0,1))$ *and* $\int_{B(0,1)} p = 0$ *). If the decomposition of* $\mathbf{g}$ *in the basis of vector spherical harmonics reads*

$$\mathbf{g}(x) = \sum g_{l,m}^{T} T_{l,m}(x) + \sum g_{l,m}^{I} I_{l,m}(x) + \sum g_{l,m}^{N} N_{l,m}(x), \tag{3.1}$$

then we obtain the decomposition of the velocity field $\mathbf{u}$ *and of the pressure field p in vector spherical harmonics for* $r > 1$ *, as follows,*

$$\begin{aligned} \mathbf{u}(x) &= \sum g_{l,m}^{T} r^{-(l+1)} T_{l,m} + \sum g_{l,m}^{I} r^{-(l+1)} I_{l,m} + \\ &\sum \left[\frac{(2l-3)(l-1)}{2l} g_{l-2,m}^{I} (r^2-1) + g_{l,m}^{N} \right] r^{-(l+1)} N_{l,m}, \\ p(x) &= \sum_{m=-l}^{l\geq 1} \frac{l(2l-1) g_{l-1,m}^{I}}{(l+1) r^{l+1}} \left[I_{l,m}(x/r) + N_{l+1,m}(x/r) \right] \cdot e_r. \end{aligned} \tag{3.2}$$

Proof. We recall that by the regularity results in (Galdi 1994), the velocity and pressure field can be decomposed in the basis of vector spherical harmonics. Since $\Delta p = 0$, we put

$$p(x) = \sum_{l\geq 0} \sum_{m=-l}^{l} \alpha_{l,m} r^{-(l+1)} Y_{l,m}(x/r).$$

We decompose $\mathbf{u}$ in the form

$$\begin{aligned} \mathbf{u}(x) &= \sum i_{l,m} \left(r^{-(l+1)} T_{l,m}(x) \right) \\ &+ \sum j_{l,m} \left(r^{-(l+1)} I_{l,m}(x) \right) \\ &+ \sum k_{l,m} \left(r^{-(l+1)} N_{l,m}(x) \right). \end{aligned}$$

This form is chosen because $r^{-(l+1)} T_{l,m}(x)$, $r^{-(l+1)} I_{l,m}(x)$ and $r^{-(l+1)} N_{l,m}(x)$ are harmonics. Using vector spherical harmonics and the formula

$$\operatorname{div}(a e_r) = \partial_r a + (2/r) a,$$

we obtain

$$\begin{aligned} \operatorname{div}\mathbf{u}(x) &= \sum (l+1) r^{-(l+2)} \left(r j'_{l,m} - (2l+1) j_{l,m} \right) Y_{l+1,m} \\ &+ \sum l r^{-(l+1)} k'_{l,m} Y_{l-1,m}. \end{aligned}$$

Since $\operatorname{div}\mathbf{u} = 0$, we deduce

$$k'_{1,0} = 0, \tag{3.3}$$

$$(l+1)\left(r^2 j'_{l,m} - r(2l+1) j_{l,m} \right) + (l+2) k'_{l+2,m} = 0. \tag{3.4}$$

We now decompose the first relation of equations (1.1). We have

$$\begin{aligned} \nabla p(x) &= \sum \frac{\alpha_{l,m}}{r^{l+2}} \left(\nabla_{S^2} Y_{l,m}\left(\frac{x}{r}\right) - (l+1) Y_{l,m}\left(\frac{x}{r}\right) e_r \right) \\ &= \sum (-\alpha_{l-1,m}) r^{-(l+1)} N_{l,m}(x/r), \end{aligned}$$

and

$$\begin{aligned} \Delta\mathbf{u}(x) &= \sum r^{-(l+2)} \left(r i''_{l,m} - 2l i'_{l,m} \right) T_{l,m} \\ &+ \sum r^{-(l+2)} \left(r j''_{l,m} - 2l j'_{l,m} \right) I_{l,m} \\ &+ \sum r^{-(l+2)} \left(r k''_{l,m} - 2l k'_{l,m} \right) N_{l,m}. \end{aligned}$$

Identifying these expansions in vector spherical harmonics, we obtain

$$r i''_{l,m} - 2l i'_{l,m} = 0, l \geq 1, |m| \leq l, \tag{7}$$

$$r j''_{l,m} - 2l j'_{l,m} = 0, l \geq 0, |m| \leq l+1, \tag{8}$$

$$r k''_{l,m} - 2l k'_{l,m} = -\alpha_{l-1,m} r, l \geq 1, |m| \leq l-1. \tag{9}$$

We deduce from (3.5), (3.6), the boundary condition $\mathbf{u} = \mathbf{g}$ on $\partial\Omega_0$ and the condition of decay at infinity that

$$i_{l,m}(r) = g_{l,m}^{T}, \quad l \geq 1, |m| \leq l,$$

$$j_{l,m}(r) = g^I_{l,m}, \quad l \geq 0, |\, m \,| \leq l+1.$$

The relations (3.3), (3.4) lead to

$$k_{1,0} = g^N_{1,0}, \quad \alpha_{0,0} = 0,$$
$$k_{l,m} = \frac{(2l-3)(l-1)}{2l} g^I_{l-2,m}(r^2-1) + g^N_{l,m}.$$

From (3.7) we get

$$\alpha_{l,m} = \frac{(2l+1)(2l-1)l}{l+1} g^I_{l-1,m}, l \geq 1,\ 0 \leq |\, m \,| \leq l.$$

Eventually, with the convention $g^I_{-1,0} = 0$ and all of the above equalities, we obtain the decomposition of p and $\mathbf{u}$ in vector spherical harmonics as in (3.2). □

4 DECOMPOSITION OF NEUMANN TO DIRICHLET OPERATOR

In this section, we obtain an expansion of the Dirichlet to Neumann operator $\mathcal{DN}$ in vector spherical harmonics. We refer the readers to (Halpern 2001) for the similar result.

Theorem 4.1. *Let* $\mathbf{g} \in W^{1/2,2}(S^2, \mathbf{R}^3)$ *and let* $(\mathbf{u}, p)$ *be a solution of* (1.1). *Then the vector spherical harmonic basis diagonalizes the Neumann to Dirichlet operator* $\mathcal{ND}$ *defined on* $\partial B(0,1)$.

In particular, if the decomposition of $\mathbf{g}$ *in the basis of vector spherical harmonics is given by* (3.1), *then we have*

$$\mathcal{ND}\mathbf{g} = \sum \frac{g^T_{l,m}}{2l+1} T_{l,m} + \sum \frac{(l+2)g^I_{l,m}}{4l^2+8l+3} I_{l,m} + \sum \frac{(l-1)g^N_{l,m}}{4l^2-1} N_{l,m}. \tag{4.1}$$

Proof. Let us first decompose the following operator

$$\mathcal{DN}_{jump}\mathbf{g} := \left[-e_r \cdot \left(\nabla \mathbf{u} + \nabla \mathbf{u}^t\right) + pe_r\right]_{|S^2}.$$

We have

$$\mathcal{DN}_{jump}\mathbf{g} = \mathcal{DN}_{ext}\mathbf{g} + \mathcal{DN}_{int}\mathbf{g}, \tag{4.2}$$

where $\mathcal{DN}_{ext}$ and $\mathcal{DN}_{int}$ correspond to the exterior and interior solutions.

Let us first decompose $\mathcal{DN}_{ext}$. We have

$$\mathcal{DN}_{ext} = \left[-e_r \cdot \left(\nabla \mathbf{u} + \nabla \mathbf{u}^t\right) + pe_r\right]_{|S^2_{ext}}.$$

For $x \in S^2$, we compute

$$\begin{aligned}
&-(\nabla \mathbf{u} \cdot e_r)(x) = \sum (l+1) g^T_{l,m} (T_{l,m} \cdot e_r) e_r \\
&-\sum g^T_{l,m} \nabla_{S^2} T_{l,m} \cdot e_r - \sum g^I_{l,m} \nabla_{S^2} I_{l,m} \cdot e_r \\
&+\sum (l+1) g^I_{l,m} (I_{l,m} \cdot e_r) e_r - \sum g^N_{l,m} \nabla_{S^2} N_{l,m} \cdot e_r \\
&+\sum \left[\frac{(2l-3)(1-l)}{l} g^I_{l-2,m} + (l+1) g^N_{l,m}\right] (N_{l,m} \cdot e_r) e_r
\end{aligned} \tag{4.3}$$

To reduce the three first terms, we use vector spherical harmonics and obtain $T_{l,m} \cdot e_r = 0$ and

$$I_{l,m} \cdot e_r = (l+1) Y_{l+1,m},$$
$$N_{l,m} \cdot e_r = l Y_{l-1,m}.$$

For the three remaining terms, we remark that for a regular vector field V of TS^2 and a spherical harmonic Y of S^2 in $\mathbf{R}^3$, we have

$$\{\nabla_{S^2} V\} \cdot e_r = -V,$$
$$\{\nabla_{S^2} (Y e_r)\} \cdot e_r = \nabla_{S^2} Y.$$

Using vector spherical harmonics, it leads

$$\{\nabla_{S^2} T_{l,m}\} \cdot e_r = -T_{l,m},$$
$$\{\nabla_{S^2} I_{l,m}\} \cdot e_r = l \nabla_{S^2} Y_{l+1,m},$$
$$\{\nabla_{S^2} N_{l,m}\} \cdot e_r = (l+1) \nabla_{S^2} Y_{l-1,m}.$$

The equality (4.3) then becomes

$$\begin{aligned}
&-(\nabla \mathbf{u} \cdot e_r)(x) = \sum (l+1)^2 g^I_{l,m} Y_{l+1,m} e_r \\
&-\sum (2l-3)(l-1) g^I_{l-2,m} Y_{l-1,m} e_r \\
&+\sum (l+1) l g^N_{l,m} Y_{l-1,m} e_r + \sum g^T_{l,m} T_{l,m} \\
&+\sum (-l) g^I_{l,m} \nabla_{S^2} Y_{l+1,m} + \sum (-l-1) g^N_{l,m} \nabla_{S^2} Y_{l-1,m}.
\end{aligned}$$

After simplifying and using again the formula of vector spherical harmonics, we obtain,

$$-(\nabla \mathbf{u} \cdot e_r)(x) = \sum g^T_{l,m} T_{l,m} - \sum l g^I_{l,m} I_{l,m} + \sum (l+1) g^N_{l,m} N_{l,m}.$$

With the same kind of computation, we obtain (see L. Halpern in (Halpern 2001))

$$\Lambda\mathbf{g}(x) = \sum (l+1) g^T_{l,m} T_{l,m} + \sum \frac{3(l+1)^2}{l+2} g^I_{l,m} I_{l,m}$$
$$+\sum (l+1) g^N_{l,m} N_{l,m},$$

where $\Lambda\mathbf{g}(x) := \left(-e_r \cdot \nabla\mathbf{u}^t + pe_r\right)_{|S^2_{ext}}$. Eventually, we get

$$\mathcal{DN}_{ext}\mathbf{g}(x) = \sum (l+2) g^T_{l,m} T_{l,m}$$
$$+\sum \frac{2l^2+4l+3}{l+2} g^I_{l,m} I_{l,m} + \sum 2(l+1) g^N_{l,m} N_{l,m}, \tag{4.4}$$

To decompose $\mathcal{DN}_{int}$, we solve the interior problem (1) in the unit ball $B(0,1)$ with $\mathbf{g} = T_{l,m}, \mathbf{g} = I_{l,m}$ and then $\mathbf{g} = N_{l,m}$.

For $\mathbf{g} = T_{l,m}$, since

$$x \mapsto r^l T_{l,m}(x/r)$$

is harmonic and divergence free, we have a solution of the form $\mathbf{u} = r^l T_{l,m}(x/r)$ and $p = 0$. Using the above formulas, it is easy to check that

$$pe_r - (\nabla\mathbf{u} + \nabla\mathbf{u}^T)\cdot e_r = (l-1)T_{l,m}.$$

For $\mathbf{g} = I_{l,m}$, we still have a solution of the form $\mathbf{u} = r^l I_{l,m}(x/r)$ and $p = 0$. Similarly, we calculate

$$pe_r - (\nabla\mathbf{u} + \nabla\mathbf{u}^T)\cdot e_r = 2lI_{l,m}.$$

For $\mathbf{g} = N_{l,m}$, the mapping

$$x \mapsto r^l N_{l,m}(x/r)$$

is not divergence free. Proceeding as in the case of the exterior domain, we look for a solution of the form,

$$\mathbf{u} = r^l N_{l,m} + \alpha r^{l-2}(1-r^2) I_{l-2,m},$$
$$p = \beta r^{l-1} Y_{l-1,m}.$$

The condition $\nabla\cdot\mathbf{u} = 0$ yields

$$\alpha = \frac{l(2l+1)}{2(l-1)}.$$

Using the first equation of (1.1), we get

$$\beta = -2(2l-1)\alpha = -\frac{l(4l^2-1)}{l-1}.$$

Then we compute

$$pe_r - (\nabla\mathbf{u} + \nabla\mathbf{u}^T)\cdot e_r = \frac{2l^2+1}{l-1} N_{l,m}.$$

We remark that the coefficient of $N_{1,0}$ is zero. Eventually, $\mathcal{DN}_{int}$ writes as:

$$\mathcal{DN}_{int}\mathbf{g}(x) = \sum (l-1) g^T_{l,m} T_{l,m} + \sum 2l g^I_{l,m} I_{l,m}$$
$$+\sum \frac{2l^2+1}{l-1} g^N_{l,m} N_{l,m}. \tag{4.5}$$

Finally, the decomposition of the Dirichlet to Neumann operator is obtained by (4.2), (4.4) and (4.5),

$$\mathcal{DN}_{jump}\mathbf{g}(x) = \sum (2l+1) g^T_{l,m} T_{l,m}$$
$$+\sum \frac{4l^2+8l+3}{l+2} g^I_{l,m} I_{l,m} + \sum \frac{4l^2-1}{l-1} g^N_{l,m} N_{l,m}. \tag{4.6}$$

This is the desired decomposition. □

REFERENCES

Galdi, G.P. (1994). *An introduction to the mathematical theory of the Navier-Stokes equations. Vol. I, Volume 38 of Springer Tracts in Natural Philosophy*. New York: Springer-Verlag. Linearized steady problems.

Halpern, L. (2001). A spectral method for the Stokes problem in three-dimensional unbounded domains. *Math. Comp. 70*(236), 1417–1436 (electronic).

Nédéelec, J.C. (2001). *Acoustic and electromagnetic equations,* Volume 144 of *Applied Mathematical Sciences*. New York: Springer-Verlag. Integral representations for harmonic problems.

Nguyen, T.N. (2013). *Convergence to equilibrium for discrete gradient-like flows and An accurate method for the motion of suspended particles in a Stokes fluid.* Dissertation. Ecole Polytechnique.

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

On convergence result for a finite element scheme of Landau-Lifschitz-Gilbert equation

M.-P. Tran
Faculty of Mathematics—Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

ABSTRACT: In this paper, we prove a convergence result of discrete solution of Landau-Lifschitz-Gilbert equation using a precise finite element scheme which was proposed by F. Alouges *et al.* in (Alouges, Kritsikis, Steiner, & Toussaint 2014). The convergence result is established in both space and time-space discretization.

1 INTRODUCTION

Recently, some convergence results for solution of gradient-like system have been studied by may authors in both continuous and discrete problems, such as (Haraux & Jendoubi 1998, Haraux 2012, Haraux & Jendoubi 2015) for continuous problem and (Grasselli & Pierre 2012, Alaa & Pierre 2013, Merlet & Pierre 2010, Merlet & Nguyen 2013) for discrete problem. These results have many applications in partial differential equations. Continue the works in (Merlet & Nguyen 2013), we apply some convergence results in this paper to a discretization of the Landau-Lifschitz-Gilbertequations using a precise finite element scheme which was proposed by F. Alouges *et al.* in (Alouges, Kritsikis, Steiner, & Toussaint 2014).

The Landau-Lifschitz-Gilbert equations was first proposed by Landau and Lifschitz in (Landau & Lifschitz 1935). These equations describe the evolution of the magnetization $m:\Omega\times(0,+\infty)\to S^2$ inside a ferromagnetic body occupying an open region $\Omega\subset\mathbf{R}^3$. This system of equations reads

$$\alpha\partial_t m - m\times\partial_t m = -m\times\mathbf{H}_{\text{eff}}, \text{ in } \Omega, \tag{1.1}$$

where $\alpha > 0$ is a damping parameter and '×' denotes the three dimensional cross product. The so-called effective magnetic field $\mathbf{H}_{\text{eff}}$ is given by the functional derivative of micromagnetic energy $\mathcal{D}$, more precisely

$$\mathbf{H}_{\text{eff}}(m) = -\frac{\partial\mathcal{D}}{\partial m} = d^2\Delta m + \mathbf{H}_{\text{d}}(m) + \mathbf{H}_{\text{ext}} + Q(e\cdot m)e,$$

where the energy $\mathcal{D}$ is given by

$$\mathcal{D}(m) = \frac{1}{2}\Big(d^2\int_\Omega |\nabla m|^2\,dx - \int_\Omega \mathbf{H}_{\text{d}}(m)\cdot mdx - 2\int_\Omega \mathbf{H}_{\text{ext}}\cdot mdx - Q\int_\Omega (e\cdot m)^2 dx\Big).$$

We use the same notations as in (Alouges, Kritsikis, Steiner, & Toussaint 2014), i.e., the vector field $\mathbf{H}_{\text{ext}}$ models an applied magnetic field, $\mathbf{H}_{\text{aniso}} = Q(e\cdot m)e$ denotes the anisotropy field, the stray field $\mathbf{H}_{\text{d}}$ is the magnetic field, d is the exchange constant and Q is the anisotropy constant. It is supplemented with initial and boundary conditions

$$\begin{cases} \dfrac{\partial m}{\partial n} = 0 \quad \text{on } \partial\Omega, \\ m(x,0) = m_0(x) \in S^2. \end{cases}$$

Notice that, at least formally, this evolution system preserves the constraint $|\mathrm{m(x,t)}| = 1, \forall x\in\Omega$.

We will consider a discretization of the following variational formulation of (1.1),

$$\alpha\int_\Omega \partial_t m\cdot\psi - \int_\Omega m\times\partial_t m\cdot\psi = d^2\int_\Omega \nabla m\cdot\nabla\psi, - \int_\Omega m\times(\mathbf{H}_{\text{d}}(m) + \mathbf{H}_{\text{ext}} + \mathbf{H}_{\text{aniso}}(m))\cdot\psi, \tag{1.2}$$

for every $\psi\in H^1(\Omega,\boldsymbol{R}^3)$ which furthermore satisfies $\psi(x)\cdot m(x) = 0$ a.e. in Ω. It is known that for every initial data $m_0\in H^1(\Omega,S^2)$, this variational formulation admits a solution for all time (see (Alouges 2008)).

The main idea comes from the fact that the Dirichlet energy function $\mathcal{D}$ of Landau-Lifschitz-Gilbert equation is a Lyapunov function for (1.2). Indeed, considering a smooth solution $m(x,t)$, we compute,

$$\frac{d}{dt}\mathcal{D}(m(\cdot,t)) = \int_{\Omega} \nabla m \cdot \nabla \partial_t m(x,t)dx.$$

Since, for every, $x \in \Omega, t \mapsto \| m(x,t) \|^2$ is constant, we have $\partial_t m(x,t) \cdot m(x,t) = 0$. So, we can choose $\psi = \partial_t m(\cdot, t)$ in (1.2) and deduce,

$$\frac{d}{dt}\mathcal{D}(m(\cdot,t)) = -\alpha \int_{\Omega} \| \partial_t m \|^2 (x,t)dx \le 0,$$

as claimed.

The sequel of this parer is organized as follows. In the next section, we recall some convergence results for the gradient-like systems. These results will be applied for our works. The convergence of numerical solution of Landau-Lifschitz-Gilbert equation are established in both cases space and time-space discretization. The first case is presented in Section 3 and the second one is in Section 4.

2 CONVERGENCE RESULTS FOR GRADIENT-LIKE SYSTEMS

In this section, we recall some abstract convergence results in recent study in (Nguyen 2013). Let M be a Riemannian manifold embedded in $\mathbf{R}^d$ and the inner product on every tangent space $T_u M$ is the restriction of the euclidian inner product on $\mathbf{R}^d$. We consider a tangent vector field $G \in C(M, TM)$ and a function $F \in C^1(M, \mathbf{R})$. We say that G and ∇F satisfy the angle and comparability condition if there exists a real number $\gamma > 0$ such that for all $u \in M$,

$$\langle G(u), -\nabla F(u) \rangle \ge \gamma \left(\| G(u) \|^2 + \| \nabla F(u) \|^2 \right). \quad (2.1)$$

We assume that F is a strict Lyapunov function for the gradient-like system

$$u'(t) = G(u(t)), u(t) \in M. \quad (2.2)$$

Theorem 2.1. *(Łojasiewicz 1971) If* $F : \mathbf{R}^d \to \mathbf{R}$ *is real analytic in some neighborhood of a point* φ *then F satisfies the Lojasiewicz inequality at* φ*, that means: there exist* $\beta, \sigma > 0$ *and* $\nu \in [0, 1/2)$ *such that*

$$| F(u) - F(\varphi) |^{1-\nu} \le \beta \| \nabla F(u) \|, \forall u \in B(0, \sigma) \cap M. \quad (2.3)$$

Theorem 2.2. *(Nguyen 2013) Assume that G and* ∇F *satisfy the angle and comparability condition 2.1 and let u be a global solution of 2.2 and there exists* φ *such that F satisfies the Lojasiewicz inequality 2.3 at* φ*. Then u(t) converges to* φ *as t goes to infinity.*

We consider the θ-scheme of the gradient-like system as follows

$$u_{n+1} - u_n = \Delta t \theta G(u_{n+1}) + \Delta t (1 - \theta) G(u_n). \quad (6)$$

Theorem 2.3. *(Nguyen 2013) Let* $\theta \in [0,1]$ *and* u_n *be the sequence defined by the* θ*-scheme* (2.4)*. If F is one-sided Lipschitz and G is Lipschitz then the sequence* u_n *converges to* φ*.*

For more general theorem, the authors studied a convergence result for a projected θ-scheme. Let $u_0 \in M$ and $n = 0, 1, 2, \ldots$, the projected θ-scheme has two steps

$$\begin{bmatrix} \text{Step 1: find } v_n \in T_{u_n} M \text{ such that} \\ \qquad v_n = \theta G_{u_n}(u_n + \Delta t v_n) + (1-\theta) G_{u_n}. \\ \text{Step 2: set } u_{n+1} := \Pi_M (u_n + \Delta t v_n) \end{bmatrix} \quad (2.5)$$

We assume that the family G_u satisfies these following conditions for all $u, u' \in M$ and $v, v' \in T_u M$:

$$\begin{aligned} & G_u(u) = G(u), \quad \| G(u) \| \le C, \\ & \| G_u(u+v) - G_u(u+v') \| \le K \| v - v' \|, \end{aligned} \quad (2.6)$$

Moreover, we also assume that the projection acts only at second order, that is there exists $\delta, R > 0$ such that

$$\| \Pi_M(u+v) - (u+v) \| \le R \| v \|^2, \text{for } \| v \| < \delta. \quad (2.7)$$

Theorem 2.4. *(Nguyen 2013) Let* u_n *be the sequence defined by the project* θ*-scheme* (2.5) *and assume that these above conditions* (2.6) *are satisfied. Then the sequence* u_n *converges to* φ*.*

3 SPACE DISCRETIZATION

We discretize the problem in space using P1-Finite Elements. Let us introduce some notation. Let $(\tau_h)_h$ be a regular family of conformal triangulations of the domain Ω parameterized by the space step h. Let $(x_i^h)_i$ be the vertices of τ_h and $(\phi_i^h)_{1 \le i \le N_{(h)}}$ the set of associated basis functions of the so-called $P^1(\tau_h)$ discretization. That is to say the functions $(\phi_i^h)_i$ are globally continuous and linear on each triangle (or tetrahedron in 3D) and satisfy $\phi_i^h(x_j^h) = \delta_{ij}$. We define

$$V^h := \left\{ m = \sum_{i=1}^{N_h} m_i \phi_i^h : \forall i, m_i \in \mathrm{R}^3 \right\},$$

$$M^h := \left\{ m \in V^h : \forall i, m_i \in S^2 \right\}.$$

Notice that M^h is a manifold isomorphic to $(S^2)^{N_h}$. For any $m = \sum_{i=1}^{N} m_i \phi_i^h \in M^h$, we introduce the tangent space

$$T_{m^h}M^h = \left\{ v = \sum_{i=1}^{N} v_i^h \phi_i^h : \forall i, m_i^h \cdot v_i^h = 0 \right\}.$$

The space discretization of the variational formulation (1.2) reads,

$$\begin{cases} m^h(0) = m_0^h \in M^h, \quad \text{and} \quad \forall \psi^h \in T_{m^h(t)}M^h, \\ \alpha \int_\Omega \partial_t m^h . \psi^h - \sum_{i=1}^{N^h} (m_i^h \times \partial_t m_i^h) \cdot \psi_i^h \int_\Omega \phi_i^h \\ \qquad = -d^2 \int_\Omega \nabla m^h . \nabla \psi^h + \\ \int_\Omega m^h \times (\mathbf{H}_\mathrm{d}(m^h) + \mathbf{H}_\mathrm{ext} + \mathbf{H}_\mathrm{aniso}(m^h)) \cdot \psi^h. \end{cases} \tag{3.1}$$

Remark 3.1. *We have replaced the term* $\int_\Omega (m^n \times p^n) \cdot \psi^h$ *in the original scheme of [?] by*

$$\sum_{i=1}^{N^h} (m_i^n \times p_i^n) \cdot \psi_i^h \int_\Omega \phi_i^h.$$

This modification is equivalent to using the quadrature formula:

$$\int_\Omega f\, dx \simeq \sum_{i=1}^{N_h} f(x_i^h) \int_\Omega \phi_i^h,$$

for the computation of this integral. The convergence to equilibrium results below are still true with an exact quadrature formula, but the proof is slightly more complicated, see Remark 4.2.

We now interpret this variational formulation as a gradient-like differential system of the form (2.2). For this we introduce the Lyapunov functional $F : M^h \subset H^1(\Omega, \mathbf{R}^3) \to \mathbf{R}$ defined by

$$F(m^h) = \frac{1}{2}\Big(d^2 \int_\Omega |\nabla m^h|^2\, dx - \int_\Omega \mathbf{H}_\mathrm{d}(m^h) \cdot m^h dx \\ -2\int_\Omega \mathbf{H}_\mathrm{ext} \cdot m^h dx - Q \int_\Omega (e \cdot m^h)^2 dx \Big).$$

As usual, the gradient of this functional is $q^h = \nabla F(m^h) = A^h m^h$, where A^h is the rigidity matrix associated to the P^1-FE discretization:

$$\begin{aligned} \left\langle q^h, \psi^h \right\rangle_{L^2} &= d^2 \int_\Omega \nabla m^h \cdot \nabla \psi^h \\ &\quad - \int_\Omega m^h \times (\mathbf{H}_d(m^h) + \mathbf{H}_{ext} + \mathbf{H}_{aniso}(m^h)) \cdot \psi^h \\ &= d^2 \sum_{i,j} m_i^h \psi_j^h \int_\Omega \nabla \phi_i^h \cdot \nabla \phi_j^h \\ &\quad - \int_\Omega m^h \times (\mathbf{H}_d(m^h) + \mathbf{H}_{ext} + \mathbf{H}_{aniso}(m^h)) \cdot \psi^h \\ &=: \left\langle A^h m^h, \psi^h \right\rangle_{L^2}. \end{aligned} \tag{3.2}$$

We also introduce the section $G : M^h \to TM^h$ defined by $G(m^h) := p^h$ where $p^h \in T_{m^h}M^h$ solves: $\forall \psi^h \in T_{m^h}M^h$,

$$\begin{aligned} &\alpha \int_\Omega p^h \cdot \psi^h - \sum_{i=1}^{N^h} (m_i^h \times p_i^h) \cdot \psi_i^h \int_\Omega \phi_i^h \\ &= -d^2 \int_\Omega \nabla m^h . \nabla \psi^h \\ &\quad + \int_\Omega m^h \times (\mathbf{H}_\mathrm{d}(m^h) + \mathbf{H}_\mathrm{ext} + \mathbf{H}_\mathrm{aniso}(m^h)) \cdot \psi^h. \end{aligned} \tag{3.3}$$

The function G is well defined. Indeed, it is sufficient to check that the bilinear form b_{m^h} defined on $T_{m^h}M^h \times T_{m^h}M^h$ by

$$b_{m^h}(p^h, \psi^h) = \alpha \int_\Omega p^h \cdot \psi^h - \sum_{i=1}^{N^h} (m_i^h \times p_i^h) \cdot \psi_i^h \int_\Omega \phi_i^h \tag{3.4}$$

has a positive symmetric part. Using $p_i^h \times p_i^h = 0$, we see that $b_{m^h}(p^h, p^h) = \alpha \| p^h \|^2_{L^2(\Omega)^2}$ and b_{m^h} is coercive on $T_{m^h}M^h \times T_{m^h}M^h$. So, by definition, $m^h \in C^1(R_+, M^h)$ solves the variational formulation (3.1) if and only if

$$\frac{d}{dt} m^h = G(m^h) \quad \forall t > 0, \quad m^h(0) = m_0^h.$$

We now check that the hypotheses of Theorem 2.2 hold.

Theorem 3.2. *The functions G and* ∇F *defined above satisfy the angle and comparability condition (2.1). Moreover, the Lyapunov function F satisfies a Łojasiewicz inequality (2.3) in the neighborhood of any point* m^h *of the manifold* $M = M^h$.

Proof. For the first point, let us fix $m^h \in M^h$ and write $p^h = G(m^h)$ and $q^h = \nabla F(m^h)$. Choosing $\psi^h = q^h$ in (3.3) and using (3.2), we obtain

$$\alpha \left\langle p^h, q^h \right\rangle_{L^2} = \sum_{i=1}^{N^h} (m_i^h \times p_i^h) \cdot q_i^h \int_\Omega \phi_i^h - d^2 \| q^h \|^2_{L^2} \\ + \int_\Omega m^h \times (\mathbf{H}_\mathrm{d}(m^h) + \mathbf{H}_\mathrm{ext} + \mathbf{H}_\mathrm{aniso}(m^h)) \cdot q^h.$$

We use the classical estimate from elliptic regularity theory, namely

$$\| \mathbf{H}_\mathrm{d}(m) \|_{L^2(\Omega)} \le C \| m \|^2_{L^2(\Omega)}.$$

Moreover, we can obtain the same estimate for applied field $\mathbf{H}_\mathrm{ext}$ and anisotropy field $\mathbf{H}_\mathrm{aniso}$, where the constant C depends on Q and $|\Omega|$. Then the Cauchy-Schwarz inequality, the identities

$\| m_i^h \| = 1$ and the equivalence of norms in finite dimension yield

$$\| q^h \|_{L^2} \leq C \| p^h \|_{L^2} .$$

On the other hand, choosing $\psi^h = p^h$ in (3.3), we get

$$\alpha \| p^h \|_{L^2}^2 = -d^2 \left\langle q^h, p^h \right\rangle_{L^2} \\ + \int_\Omega m^h \times (\mathbf{H}_{\mathrm{d}}(m^h) + \mathbf{H}_{\mathrm{ext}} + \mathbf{H}_{\mathrm{aniso}}(m^h)) \cdot p^h.$$

So, we have

$$\left\langle -q^h, p^h \right\rangle_{L^2} \geq \gamma \left(\| p^h \|_{L^2}^2 + \| q^h \|_{L^2}^2 \right),$$

with γ depends on Q, $|\Omega|$, α and d: i.e. the pair $(-\nabla F, G)$ satisfies the tangential angle condition and comparability condition (2.1).

For the second point, $F(m^h)$ is a polynomial function of $(m_i^h)_{1 \leq i \leq N_h} \in (S^2)^{N_h}$, hence it is analytic. The manifold $M^h = (S^2)^{N_h}$ being analytic, we can use an analytic chart φ (for example a product of stereographic projections) defined in a neighborhood of m^h. We apply Theorem 2.1 to the □ analytic function $F \circ \varphi^{-1}$ and deduce that it satisfies a Łojasiewicz inequality in the neighborhood of $\varphi(m^h)$. □

We deduce from the Theorem 3.2:

Theorem 3.3. *Assume $m^h(t)$ is a solution of (3.1). Since $M = M^h$ is compact $\omega(m^h)$ is not empty. Consequently there exists $\varphi \in M^h$ such that $u = m^h$ satisfies the conclusion of Theorems 2.2.*

4 TIME-SPACE DISCRETIZATION

We now consider the θ-scheme proposed by F. Alouges in (Alouges, Kritsikis, Steiner, & Toussaint 2014). Given an initial $m^0 \in M^h$, choose $\theta \in [0,1]$ and a time step Δt. For $n = 0, 1, 2, \ldots$, the algorithm has two steps:

$$\left[\begin{array}{l} \text{Find } p^n \in T_{m^n} M^h \text{ such that } \forall \psi^h \in T_{m^h} M^h, \\ \alpha \int_\Omega p^n \cdot \psi^h - \sum_{i=1}^{N^h} (m_i^n \times p_i^n) \cdot \psi_i^h \int_\Omega \phi_i^h \\ \qquad = -d^2 \int_\Omega \nabla (m^n + \theta \Delta t \, p^n) \cdot \nabla \psi^h \\ + \int_\Omega m^n \times (\mathbf{H}_{\mathrm{d}}(m^n) + \mathbf{H}_{\mathrm{ext}} + \mathbf{H}_{\mathrm{aniso}}(m^n)) \cdot \psi^h. \\ \text{Set } m^{n+1} := \sum_{i=1}^{N_h} m_i^n + \Delta t \, p_i^n \left| m_i^n + \Delta t \, p_i^n \right| \phi_i^h, \text{ and iterate.} \end{array} \right. \tag{4.1}$$

Let us rewrite this scheme as a projected θ-scheme of the form (2.5). For this we introduce the family of mappings $\left\{ G_{m^h} : m^h + T_{m^h} \to T_{m^h} M^h \right\}$ defined by $G_{m^h}(u^h) = p^h$ where $p^h \in T_{m^h} M^h$ solves the variational formulation $\forall \psi^h \in T_{m^h} M^h$,

$$\alpha \int_\Omega p^h \cdot \psi^h - \sum_{i=1}^{N^h} (u_i^h \times p_i^h) \cdot \psi_i^h \int_\Omega \phi_i^h \\ = -d^2 \int_\Omega \nabla u^h \cdot \nabla \psi^h \\ + \int_\Omega m^n \times (\mathbf{H}_{\mathrm{d}}(m^n) + \mathbf{H}_{\mathrm{ext}} + \mathbf{H}_{\mathrm{aniso}}(m^n)) \cdot \psi^h.$$

Notice that G_{m^h} only depends on m^h through the space of test functions $T_{m^h} M^h$. As above, we see that p^h is well defined and uniquely defined by this variational formulation through the coercivity of the bilinear for b_{m^h} (see 3.4).

Lemma 4.1. *Let m^n, p^n be defined in the scheme (4.1). Then,*

$$p^n = \theta G_{m^n}(m^n + \Delta t p^n) + (1 - \theta) G_{m^n}(m^n). \tag{4.2}$$

Proof. Let us set $q^h = G_{m^n}(m^n + \Delta t p^n)$, $r^h = G_{m^n}(m^n)$. By definition of G_{m^n} and linearity, we see that the function $p^h = \theta q^h + (1 - \theta) r^h$ satisfies

$$\alpha \int_\Omega p^h \cdot \psi^h - \sum_{i=1}^{N^h} (m_i^h \times p_i^h) \cdot \psi_i^h \int_\Omega \phi_i^h \\ - \theta \Delta t \sum_{i=1}^{N^h} (p_i^n \times r_i^h) \cdot \psi_i^h \int_\Omega \phi_i^h \\ = -d^2 \int_\Omega \nabla (m^n + \theta \Delta t \, p^n) \cdot \nabla \psi^h, \\ + \int_\Omega m^n \times (\mathbf{H}_{\mathrm{d}}(m^n) + \mathbf{H}_{\mathrm{ext}} + \mathbf{H}_{\mathrm{aniso}}(m^n)) \cdot \psi^h, \\ \forall \psi^h \in T_{m^h} M^h.$$

We see that in the third term of the left hand side, the triple product $(p_i^n \times r_i^h) \cdot \psi_i^h$ vanishes. Indeed, the three vectors p_i^n, r_i^h, ψ_i^h belong to the two dimensional tangent space $\left\{ v_i^h \in \mathrm{R}^3 : v_i^h \cdot m_i^n = 0 \right\}$. So, it turns out that p^h and p^n solve the same (well-posed) variational formulation. We conclude that $p^h = p^n$ as claimed. □

Remark 4.2. *If we had used the original variational formulation, with obvious changes in the definition of G_{m^h}, then the term $\theta \Delta t \int_\Omega (p^n \times r^h) \cdot \psi^h$ would not vanish in general and the identity (4.2) would be wrong. In this case, we can not link the scheme of (Alouges, Kritsikis, Steiner, & Toussaint 2014) to our projected θ-scheme. However, this term is of small magnitude and using the present ideas, it is not difficult to establish that Theorems 2.3 and 2.4 apply to this scheme and conclude to the convergence to equilibrium of the sequence (m^n).*

Theorem 4.3. *The functions F, G and* $\{G_{m^h}\}$ *satisfy hypotheses* (2.6)*. Moreover, the projection* $\Pi_{M^h}(z^h) := \sum_{i=1}^{N_h} z_i^h \left| z_i^h \right| \phi_i^h$, *satisfies* (2.7).

Proof. The first identity in (2.6) is obvious. Next, for $m^h \in M^h$ and $p^h = G(m^h)$, using $\psi^h = p^h$ in (3.3), we obtain

$$\alpha \| p^h \|_{L^2}^2 \leq (1+\alpha^2) \| \nabla m^h \|_{L^2} \| \nabla p^h \|_{L^2},$$

and we conclude from the equivalence of the norms in finite dimensional spaces, that G is bounded on the compact manifold M^h. The Lipschitz estimate in (2.6) is also a consequence of this fact and of the uniform coercivity of the bilinear forms b_{m^h}. The Lipschitz estimate on ∇F is also obvious since F is smooth on the compact manifold M^h.

Eventually, we easily see that (2.7) holds. Indeed, if $v^h \in T_{m^h} M^h$, then $| m_i^h + v_i^h |^2 = | m_i^h |^2 + | v_i^h |^2 \geq 1$, so $\Pi_{M^h}(m^h + v^h)$ is just the L^2-projection of $(m_i^h + v_i^h)$ on the product of balls $(\overline{B}(0,1))^{N_h} \subset (\mathbf{R}^3)^{N^h}$. □

The previous Theorem 4.1 and 4.3 show that the sequence $(u_n = m^n)$ satisfies all the hypotheses for Theorem 2.3. Hence, we have:

Theorem 4.4. *There exists* $\Delta t' > 0$ *such that if* $\Delta t \in (0, \Delta t')$ *and* $(m^n) \subset M^h$ *is a sequence that complies to the scheme* (4.1)*, then there exists* $\varphi \in M^h$ *such that* (m^n) *converges to* φ.

REFERENCES

Alaa, N.E. & M. Pierre (2013). Convergence to equilibrium for discretized gradient-like systems with analytic features. ***IMA** J. Numer. Anal. 33*(4), 1291–1321.

Alouges, F. (2008). A new finite element scheme for Landau-Lifchitz equations. *Discrete Contin. Dyn. Syst. Ser. S 1*(2), 187–196.

Alouges, F., E. Kritsikis, J. Steiner, & J.C. Toussaint (2014). A convergent and precise finite element scheme for Landau-Lifschitz-Gilbert equation. *Numer. Math. 128*(3), 407–430.

Grasselli, M. & M. Pierre (2012). Convergence to equilibrium of solutions of the backward Euler scheme for asymptotically autonomous second-order gradient-like systems. *Commun. Pure Appl. Anal. 11*(6), 2393–2416.

Haraux, A. (2012). Some applications of the łojasiewicz gradient inequality. *Commun. Pure Appl. Anal. 11*(6), 2417–2427.

Haraux, A. & M.A. Jendoubi (1998). Convergence of solutions of second-order gradient-like systems with analytic nonlinearities. *J. Differential Equations 144*(2), 313–320.

Haraux, A. & M.A. Jendoubi (2015). *The convergence problem for dissipative autonomous systems*. Springer Briefs in Mathematics. Springer, Cham; BCAM Basque Center for Applied Mathematics, Bilbao. Classical methods and recent advances, BCAM SpringerBriefs.

Landau, L. & I. Lifschitz (1935). On the theory of the dispersion of magnetic permeability in feromagnetic bodies. *Phys. Zeitsch. der Sow. 8*, 153–169.

Łojasiewicz, S. (1971). Sur les ensembles semi-analytiques. In *Actes du Congr`es International des Math'ematiciens (Nice, 1970), Tome 2,* pp. 237–241. Gauthier-Villars, Paris.

Merlet, B. & T.N. Nguyen (2013). Convergence to equilibrium for discretizations of gradient-like flows on Riemannian manifolds. *Differential Integral Equations 26*(5–6), 571–602.

Merlet, B. & M. Pierre (2010). Convergence to equilibrium for the backward Euler scheme and applications. *Commun. Pure Appl. Anal. 9*(3), 685–702.

Nguyen, T.N. (2013). *Convergence to equilibrium for discrete gradient-like flows and An accurate method for the motion of suspended particles in a Stokes fluid. Dissertation. Ecole Polytechnique*.

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Some results on the viscous Cahn-Hilliard equation in $\mathbb{R}^N$

L.T.T. Bui
Faculty of Mathematics and Computer Science, University of Science, Vietnam National University, Ho Chi Minh City, Vietnam

N.A. Dao
Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

ABSTRACT: We study existence and uniqueness of solution for the viscous Cahn-Hilliard equation under weak growth assumptions on the nonlinearity φ and in whole domain $\mathbb{R}^N$. We also address a priori estimates which are sufficient to investigate singular passage to the limit over different small parameters.

1 INTRODUCTION

Forward-backward parabolic equations arise in a variety of applications, such as edge detection in image processing (Perona & Malik (1990)), aggregation models in population dynamics (Padrón (1998)), and stratified turbulent shear flow (Barenblatt & Bertsch & Dal Passo & Prostokishin & Ughi (1993)), theory of phase transitions (Brokate & Sprekels (1996), Bellettini & Fusco & Guglielmi (2006)). A well-known equation of this type is the *Perona-Malik equation*,

$$w_t = \operatorname{div}\left(\frac{\nabla w}{1+|\nabla w|^2}\right), \tag{1}$$

which is parabolic if $|\nabla w| < 1$ and backward parabolic if $|\nabla w| > 1$. Similarly, the equation

$$u_t = \Delta\left(\frac{u}{1+u^2}\right) \tag{2}$$

is parabolic if $|u| < 1$ and backward parabolic if $|u| > 1$. Observe that in one space dimension the above equations are formally related setting $u = w_x$. A different well-known equation of application in theory of phase transitions is

$$u_t = \Delta\varphi(u) \tag{3}$$

where the famous choice of nonlinearity is $\varphi(u) = u^3 - u$.

Clearly, forward-backward parabolic equations lead to ill-posed problems. Often a higher order term is added to the right-hand side to regularize the equation. Two main classes of additional terms are encountered in the mathematical literature, which, *e.g.* in case of equation (2), (3), reduce to:

i. $\epsilon\Delta[\psi(u)]_t$ with $\psi' > 0$, leading to third order *pseudo-parabolic equations* ($\epsilon > 0$ being a small parameter; for example, see (Novick-Cohen & Pego (1991), Plotnikov (1994), Smarrazzo & Tesei (2012), Smarrazzo & Tesei (2013), Bui et al. 2014a));
ii. $-\epsilon\Delta^2 u$, leading to fourth-order *Cahn-Hilliard type equations* (for example, see (Bellettini & Fusco & Guglielmi (2006), Plotnikov (1997), Slemrod (1991)) and references therein).

Remarkably, when $\psi(u) = u$ either of the above regularizations can be regarded as a particular case of the viscous *Cahn-Hilliard equation*,

$$\nu u_t = \Delta\left[\varphi(u) - \alpha\Delta u + \beta u_t\right] \ (\alpha,\beta,\nu > 0), \tag{4}$$

choosing either $\alpha = \epsilon$ or $\beta = \epsilon$; here $\varphi(u) = u^3 - u$ or $\varphi(u) = \frac{u}{1+u^2}$ for equation (2), whereas in general it denotes a *non-monotonic* function.

Equation (4) has been derived by several authors using different physical considerations (in particular, see (Gurtin (1996), Jäckle & Frisch (1986), Novick-Cohen (1988))). It is worth mentioning the wide literature concerning both the relationship between the viscous Cahn-Hilliard equation and *phase field models*, and generalized versions of the equation suggested in (Gurtin (1996)).

Concerning equation (4) with $\nu = 1 - \beta$, the existence results were obtained under suitable nonlinearity φ in bounded smooth domain of $\mathbb{R}^N$ (see (Carvalho & Dłotko (2007)), (Elliott & Stuart (1996)), (Bui et al. 2014b)). Moreover, in the latter reference authors give us the rigorous proof of convergence to solutions of either the Cahn-Hilliard equation, or of the Allen-Cahn equation, or of the Sobolev equation, depending on the choice of the parameter α, β. Recently, in (Dłotko et al. 2012) authors give the analysis of equation (4) in $\mathbb{R}^N$ under some restrictive assumptions on the growth of nonlinearity φ.

In the light of the above considerations, we study the following viscous Cahn-Hilliard parabolic problem

$$\begin{cases}(1-\beta)u_t = \Delta\left[\varphi(u)-\alpha\Delta u+\beta u_t\right] & \text{in}\,\mathbb{R}^N\times(0,T)\\ \lim_{|x|\to\infty} u = \lim_{|x|\to\infty}\Delta u = 0 & \text{for}\, t\in(0,T)\\ u=u_0 & \text{in}\,\mathbb{R}^N\times\{0\},\end{cases} \tag{5}$$

where the nonlinearity φ satisfies the following assumptions:

(H_1) there exists $K > 0$ such that

$$|\varphi'(u)|\leq K(1+|u|^{q-1}) \tag{6}$$

for some $q \in (1, \infty)$ if $N = 1$, 2, or $q\in\left(1,\frac{N+2}{N-2}\right]$ if $N \geq 3$.

We obtain the existence results with more extensive of class of nonlinearities φ which include the critical growth in (Dłotko et al. 2012). By the same way but more technical, we also give the analysis of singular limits of problem (5) as in (Bui et al. 2014b). Here are the description of our main method. Firstly, we state and prove the existence of weak solution of the viscous Cahn-Hilliard problem in a ball B_n which has center at origin and radius $n \in \mathbb{N}$:

$$\begin{cases}(1-\beta)u_t = \Delta\left[\varphi(u)-\alpha\Delta u+\beta u_t\right] & \text{in}\, B_n\times(0,T)\\ u=\Delta u = 0 & \text{on}\,\partial B_n\times(0,T)\\ u=u_0 & \text{in}\, B_n\times\{0\}.\end{cases} \tag{7}$$

It is worth to mention that this result (see Theorem 3) is also an improvment of that in (Bui et al. 2014b). Second, we establish the family of uniformly bounded estimates on those solution independent of n (see Lemma 4). Then we can pass to the limit as $n \to \infty$ w to get a desired result (see Theorem 5). Finally, by taking advantage of the set of uniformly bounded estimates on solution of problem (5) with respect to appropriate parameters, we investigate its singular limits to get solutions of Sobolev equation or Cahn-Hilliard equation. Here is the different way to get solutions of well-known equations which are extensively investigated in the literature.

This paper was organized as follow : Section 1 is for the introduction of our problem. Our main results are presented in Section 2.

2 MAIN RESULTS

In this paper, let $\Omega = B_n$ and $Q_n = B_n \times (0, T)$, $Q = \mathbb{R}^N \times (0, T)$.

Definition 1. *Let $\alpha \in (0, \infty)$, $\beta \in (0,1)$, and let $u_0\in H^2(\Omega)\cap H_0^1(\Omega)$. By a strict solution of problem* (7) *we mean any function $u\in C([0,T];$ $H^2(\Omega)\cap H_0^1(\Omega))\cap C^1([0,T];L^2(\Omega))$ such that $\varphi(u)$ $\in C([0,T];L^2(\Omega))$, and*

$$\begin{cases}u_t=\Delta v & \mathit{in}\ Q_n\\ u=u_0 & \mathit{in}\ \Omega\times\{0\}\end{cases} \tag{8}$$

in strong sense. Here $v\in C([0,T];H^2(\Omega)\cap H_0^1(\Omega))$ and for every $t \in [0, T]$ the function $v(\cdot,t)$ is the unique solution of the elliptic problem

$$\begin{cases}-\beta\Delta v+(1-\beta)v=\varphi(u)-\alpha\Delta u & \mathit{in}\ \Omega\\ v(\cdot,t)=0 & \mathit{on}\ \partial\Omega.\end{cases} \tag{9}$$

The function v is called chemical potential.

Definition 2. *Let $\alpha\in(0,\infty)$, $\beta\in(0,1)$, and let* $\mathrm{u}_0 \in H^2(\mathbb{R}^N)$. *By a strict solution of problem* (5) *we mean any function $u\in C([0,T];H^2(\mathbb{R}^N))\cap C^1([0,T];L^2(\mathbb{R}^N))$ such that $\varphi(u)\in C([0,T];L^2(\mathbb{R}^N))$, and*

$$\begin{cases}u_t=\Delta v & \mathit{in} Q\\ u=u_0 & \mathit{in}\mathbb{R}^N\times\{0\}\end{cases} \tag{10}$$

in strong sense. Here $v\in C([0,T];H^2(\mathbb{R}^N)\cap \mathrm{H}_0^1(\mathbb{R}^N))$ and for every $t\in[0,T]$ the function $v(\cdot,t)$ is the unique solution of the elliptic problem

$$\begin{cases}-\beta\Delta v+(1-\beta)v=\varphi(u)-\alpha\Delta u & \mathit{in}\ \Omega\\ \lim_{|x|\to\infty} v(x,t)=0.\end{cases} \tag{11}$$

The *function v is called* chemical potential.

A well-posedness result for problem (7) under assumption (H_1) is the content of the following theorem.

Theorem 3. *Let $\alpha \in (0, \infty)$, $\beta \in (0, 1)$, and let φ satisfy assumption (H_1). Then for every $u_0\in H^2(\Omega)\cap H_0^1(\Omega)$ there exists a unique strict solution of problem* (7).

Lemma 4. *Let $\alpha\in(0,\infty)$, $\beta\in(0,1)$, ,$u_0\in H^2(\Omega)\cap \mathrm{H}_0^1(\Omega)$ and let φ satisfy assumption (H_1). Let $\{u_n\}$ be the sequence of solutions to problems* (7) *given by Theorem 3, with $u_{0n}=u_0$. Then for every $t \in (0,T]$ there holds*

$$\begin{aligned}&\int_\Omega \Phi_n(u_n)(x,t)dx+\frac{\alpha}{2}\int_\Omega|\nabla u_n|^2(x,t)dx\\ &+\beta\int_0^t\int_\Omega u_{nt}^2dxds+(1-\beta)\int_0^t\int_\Omega|\nabla u_n|^2dxds\\ &=\int_\Omega\Phi_n(u_{0n})(x)dx+\frac{\alpha}{2}\int_\Omega|\nabla u_{0n}|^2dx,\end{aligned} \tag{12}$$

where $\Phi(u)=\int_0^u\varphi(s)ds$.

By using the above uniform estimates, we can state and prove our main theorem as follows:

Theorem 5. *Let* $\alpha \in (0, \infty)$, $\beta \in (0, 1)$, *and let* φ *satisfy assumption* (H_1). *Then for every* $u_0 \in H^2(\mathbb{R}^N)$ *there exists a unique strict solution of problem* (5). *Moreover, for every* $\bar{\alpha} > 0$ *there exists* $M > 0$ (*only depending on the norm* $\| u_0 \|_{H^2(\mathbb{R}^N)}$) *such that for any* $\alpha \in (0, \bar{\alpha})$ *and* $\beta \in (0,1)$

$$\|\Phi(u)\|_{L^\infty((0,T);L^1(\mathbb{R}^N))} \leq M, \tag{13}$$

where $\Phi(u) = \int_0^u \varphi(s)ds$;

$$\sqrt{\alpha}\|u\|_{L^\infty((0,T);H^1(\mathbb{R}^N))} \leq M; \tag{14}$$

$$\sqrt{\beta}\|u_t\|_{L^2(Q)} \leq M; \tag{15}$$

$$\sqrt{\alpha}\|\varphi(u)\|_{L^2(Q)} \leq M; \tag{16}$$

$$\alpha^{\frac{3}{2}}\|\Delta u\|_{L^2(Q)} \leq M; \tag{17}$$

$$\sqrt{1-\beta}\|v\|_{L^2((0,T);H^1(\mathbb{R}^N))} \leq M; \tag{18}$$

$$\sqrt{\alpha\beta}\|v\|_{L^\infty((0,T);H^1(\mathbb{R}^N))} \leq M; \tag{19}$$

$$\sqrt{\beta(1-\beta)}\,\|v\|_{L^2((0,T);H^2(\mathbb{R}^N))} \leq M. \tag{20}$$

Further estimates of the solution given by Theorem 5 are the content of the following theorem.

Theorem 6. *Let* $\alpha \in (0, \infty)$, $\beta \in (0, 1)$ *and* $u_0 \in H^2(\mathbb{R}^N)$. *Let* φ *satisfy* (H_1) *and the following one*:

(H_2) *there exists* $u_0 > 0$ *such that* $\varphi'(u) > 0$ *if* $|u| \$ u_0$.

Let u be the solution of problem (5) *given by Theorem* 5. *Then for every* $\alpha \in (0, \infty)$ *and* $\beta \in (0, 1)$

$$\|u\|_{L^\infty((0,T);L^2(\mathbb{R}^N))} \leq \|u_0\|_{H^1(\mathbb{R}^N)}\sqrt{\frac{1+e^{\frac{2LT}{\beta}}}{1-\beta}}; \tag{21}$$

$$\|u\|_{L^\infty((0,T);H^1(\mathbb{R}^N))} \leq \|u_0\|_{H^1(\mathbb{R}^N)}\sqrt{\frac{2\left(1+e^{\frac{2LT}{\beta}}\right)}{\beta}}; \tag{22}$$

$$\|\Delta u\|_{L^2(Q)} \leq \|u_0\|_{H^1(\mathbb{R}^N)}\sqrt{\frac{1+e^{\frac{2LT}{\beta}}}{\alpha}}. \tag{23}$$

Moreover, for every $\bar{\alpha} > 0$ and $\beta \in (0,1)$ *there exists* $\bar{M} > 0$ (*only depending on the norm* $\| u_0 \|_{H^1(\mathbb{R}^N)}$ *and on* β, *and diverging as* $\beta \to 0^+$, $\beta \to 1^-$) *such that for any* $\alpha \in (0, \bar{\alpha})$ and $n \in \mathbb{N}$

$$\|\varphi(u)\|_{L^2(Q)} \leq \bar{M}. \tag{24}$$

REFERENCES

Barenblatt, G.I., M. Bertsch, R. Dal Passo, V.M. Prostokishin & M. Ughi (1993). A mathematical problem of turbulent heat and mass transfer in stably stratified turbulent shear flow. *J. Fluid Mech*. **253**, 341–358.

Bellettini, G., G. Fusco & N. Guglielmi (2006). A concept of solution for forward-backward equations of the form $u_t = \frac{1}{2}\left(\zeta\left(u_x\right)\right)_x$ and numerical experiments for the singular perturbation $u_t = \varepsilon^2 u_{xxxx} + \frac{1}{2}\left(\zeta\left(u_x\right)\right)_x$. *Discrete Cont. Dyn. Syst*., **16**, 259–274.

Brokate M. & J. Sprekels (1996). Hysteresis and Phase Transitions. *Applied Mathematical Sciences,* **121** (Springer, 1996).

Bui, L.T.T., F. Smarrazzo & A. Tesei (2014a). Sobolev regularization of a class of forward-backward parabolic equations. *Journal of differential Equations,* 257 (5), 1403–1456.

Bui, L.T.T., F. Smarrazzo & A. Tesei (2014b). Passage to the limit over small parameters of a viscous Cahn-Hilliard equation. *J. Math. Analysis App*., 420 (2), 1265–1300.

Carvalho A.N. & T. Dłotko (2007). Dynamics of viscous Cahn-Hilliard equation. *Cadernos de Matemática*, **8**, 347–373.

Elliott, C.M. & A.M. Stuart (1996). Viscous Cahn-Hilliard Equation, II. Analysis. *Journal of differential Equations.* **128**, 387–414.

Gurtin M. (1996). Generalized Ginzburg-Landau and Cahn-Hilliard equations based on a microforce balance. *Physica D* **92**, 178–192.

Jäckle, J. & H.L. Frisch (1986). Properties of a generalized diffusion equation with memory. *J. Chem. Phys.* **85**, 1621–1627.

Novick-Cohen, A. & R.L. Pego (1991). Stable patterns in a viscous diffusion equation. *Trans. Amer. Math. Soc.* **324**, 331–351.

Novick-Cohen, A. (1988). On the viscous Cahn-Hilliard equation. in "*Material Instabilities in Continuum Mechanics and Related Mathematical Problems*" *(J. M. Ball, Ed.)*, pp. 329–342, Clarendon Press.

Padrón, V. (1998). Sobolev regularization of a nonlinear ill-posed parabolic problem as a model for aggregating populations. *Comm. Partial Differential Equations*, **23**, 457–486.

Perona, P. & J. Malik (1990). Scale space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**, 629–639.

Plotnikov, P.I. (1994). Passing to the limit with respect to viscosity in an equation with variable parabolicity direction. *Diff. Equ.* **30**, 614–622.

Plotnikov, P.I. (1997). Passage to the limit over a small parameter in the Cahn-Hilliard equations. *Siberian Math. J.* **38**, 550–566.

Slemrod, M. (1991). Dynamics of measure-valued solutions to a backward-forward heat equation. *J. Dynam. Differential Equations* **3**, 1–28.

Smarrazzo F. & A. Tesei (2012). Degenerate regularization of forward-backward parabolic equations: The regularized problem. *Arch. Rational Mech. Anal.* **204**, 85–139.

Smarrazzo, F. & A. Tesei (2013). Degenerate regularization of forward-backward parabolic equations: The vanishing viscosity limit. *Math. Ann.* **355**, 551–584.

Tomasz Dłotko, Maria B. Kania & Chunyou Sun (2012). Analysis of the viscous Cahn-Hilliard equation in R^N. *Journal of differential Equations,* **252**, 2771–2791.

This page intentionally left blank

Inverse problems

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

On a multi-dimensional initial inverse heat problem with a time-dependent coefficient

C.D. Khanh & N.H. Tuan
Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

ABSTRACT: In this paper, we solve an initial inverse problem for an inhomogeneous heat equation. The problem is ill-posed, as the solution exhibits unstable dependence on the given data functions. Up to now, most of studies are focused on the homogeneous problem, and with constant coefficients. Recently, we solved the heat problem with time-dependent coefficients in 1-D for a homogeneous heat equation. This work is a continuous expansion of previous results (See (Quan 2011)) (N.H. Tuan & Triet 2013). Herein we introduce two efficient regularization methods, the quasi-boundary-type and high frequency truncation methods. Some error estimates between the regularization solutions and the exact solution are obtained even in the Sobolev space H^1 and H^2.

Keywords: ill-posed problems; boundary value method; truncation method; heat equation; regularization

1 INTRODUCTION

Forward heat conduction problems related to the heat equation is to predict the temperature field at a subsequent time of a medium from the knowledge of an initial temperature and boundary conditions. On the other hand, an inverse heat conduction problem is to recover the temperature distribution at a certain time from the knowledge of the temperature and boundary conditions at a later time. The inverse problems for heat equation are of great importance in engineering applications, and aim to detect a previous status of physical field from its present information. They can be applied to several practical areas, such as image processing, mathematical finance, physics and mechanics of continuous media, etc. In this paper, we consider the problem of finding the temperature $u(x,t)$, such that

$$\begin{cases} \dfrac{\partial u}{\partial t} = b(t)\Delta u + f(x,t), & (x,t) \in \Omega \times (0,T), \\ u\,|_{\partial\Omega} = 0, & t \in (0,T), \\ u(x,T) = g(x), & x \in \Omega, \end{cases} \tag{1}$$

where Ω is an open bounded and connected domain in R^n with sufficiently smooth boundary, Δ is Laplace operator on R^n, $\partial\Omega$ is the boundary of Ω, and $b(t)$, $g(x)$, $f(x,t)$ are given. It is well-known that the backward problem is ill-posed, i.e., its solution does not always exist, and in the case of existence, it does not depend continuously on the given datum. In fact, from a small noise of contaminated physical measurement, the corresponding solutions may have a large error. This makes the numerical computation difficult, hence a regularization is needed.

Many papers are devoted to special cases of the problem (1) in one dimension. For instance, when $b(t) = 1$ and $f(x,t) = 0$, the problem (1) has been investigated by many authors, such as John (John 1960) who introduced a fundamental concept to prescribe a bound on the solution at $t = T$ with relaxation of an initial data g; Lattes and Lions (Lattès & J.L. Lion 1967), Showalter (Showalter 1974), and Ewing (Ewing 1975) used quasi-reversibility method. Other approaches including the least squares methods with Tikhonov-type regularization were introduced by Ames and Epperson (Ames & Epperson 1997), and Miller (Miller 1970). A parallel method for backward parabolic problems is proposed by Lee and Sheen (J. Lee 2006, J. Lee 2009). This problem was also investigated by many other authors, such as Clark and Oppenheimer (G.W. Clark & S.F. Oppenheimer 1994), Ames et al. (Ames & Epperson 1997), Denche and Bessila (Denche & Bessila 2005), Tautenhahn et al. (T. Schroter 1996), Melnikova et al. (I.V. Melnikova 1993b, I.V. Melnikova 1993a), Fu (X.L. Feng 2008, C.L. Fu 2007), Yildiz et al. (B. Yildiz 2000, B.Yildiz 2003). When $b(t) = 1$ and $f(x,t) \neq 0$, the problem (1) has been studied by Trong et al. (Trong & Tuan 2006, Trong & Tuan 2008).

Up to now, the backward heat problem with the time-dependent coefficient of Δu in the main equation is still continuously investigated. This kind of equation $u_t - b(t)\Delta u = f(x,t)$ has many applications in groundwater pollution. It is a simple form of the advection-convection, which appear in groundwater pollution source identification problems (See (Atmadja & Bagtzoglou 2003)).

This work is a continuous expansion of our previous results (Quan 2011). We solve the heat equation in the multi-dimensional case by two regularization methods, the modified quasi-boundary value method and the truncation method. The first method is the perturbation method, whereby we modified the source term *f* and the final data *g*. The main idea of the quasi-boundary-value method is to replace the boundary value problem with an approximate well-posed one, then to construct approximate solutions of the given boundary value problem. This method has been applied in many problems, such as the evolution operator differential equation (G.W. Clark & S.F. Oppenheimer 1994), the hyper-parabolic partial differential equation (Showalter 1983), the elliptic equations (Feng 2010), etc.

The second method is based on ideas of the paper (C.L. Fu 2007). Moreover, using the truncation method can be easily obtained an error estimate to archive a better convergence rate. This fact has been confirmed in (X.L. Feng 2010, C.L. Fu 2007). The truncated regularization method is an effective method for solving some ill-posed problems, and it has been successfully applied to some inverse heat conduction problems (Berntsson 1999).

The outline of the rest of the paper is as follows. In the next section, we simply analyze the ill-posedness of the problem (1). In Sections 3 and 4, we introduce two regularized methods and error estimates between the exact solution and the regularized solutions, respectively.

2 MATHEMATICAL INITIAL INVERSE PROBLEM OF HEAT CONDUCTION

Let $b:[0,T]\to R$ be a continuous function on $[0,T]$ satisfying $0 < B_1 \le b(t) \le B_2, \forall t \in [0,T]$; it is assumed to be differentiable for every *t* and satisfy $0 < b'(t) \le C_1$ for $t \in (0,T)$.

Throughout this article, we denote the L^2-norm by $\|.\|$, and the inner product on $L^2(\Omega)$ by $<,>$. We also suppose that $f \in L^2((0,T); L^2(\Omega))$ and $g \in L^2(\Omega)$. First, we state a few properties of the eigenvalues of the operator $-\Delta$ on the open, bounded and connected domain Ω with Dirichlet boundary conditions, which can also be referred to Section 6.5 in (Evans 1997).

Known facts (Eigenvalues of the Laplace operator)

1. *Each eigenvalue of* $-\Delta$ *is real. The family of eigenvalues* $\{\lambda_p\}_{p=1}^{\infty}$ *satisfy* $0 < \lambda_1 \le \lambda_2 \le \lambda_3 \le \ldots,$ *and* $\lambda_p \to \infty$ *as* $p \to \infty$.
2. *There exists an orthonormal basis* $\{X_p\}_{p=1}^{\infty}$ *of* $L^2(\Omega)$, *where* $X_p \in H_0^1(\Omega)$ *is an eigenfunction corresponding to* λ:

$$\begin{cases} -\Delta X_p(x) = \lambda_p X_p(x), x \in \Omega \\ X_p(x) = 0, x \in \partial\Omega, \end{cases} \tag{2}$$

for $p = 1, 2, \ldots$

Let $0 \ne q < \infty$. By $S^q(\Omega)$ we denote the space of all functions $g \in L^2(\Omega)$ with the property

$$\sum_{p=1}^{\infty} (1+\lambda)^{2q} \,|\, g_p \,|^2 < \infty, \tag{3}$$

where $g_p = \int_\Omega g(x) X_p(x) dx$. We define $\|g\|_{S^q(\Omega)} = \sqrt{\sum_{p=1}^{\infty} (1+\lambda_p)^{2q} \,|\, g_p \,|^2}$. If $q=0$ then $S^q(\Omega)$ is $L^2(\Omega)$. (See [?] Chapter V) and (X.L. Feng 2008) (page 179).

As we know, the forward heat problem

$$\begin{cases} \dfrac{\partial u}{\partial t} = b(t)\Delta u + f(x,t), & (x,t) \in \Omega \times (0,T) \\ u\,|_{\partial\Omega} = 0, & t \in (0,T) \\ u(x,0) = g(x), & x \in \Omega \end{cases} \tag{4}$$

where $f \in L^2((0,T); L^2(\Omega))$ $g \in L^2(\Omega)$ has a unique solution. However, for the backward problem (1) where $f \in L^2((0,T); L^2(\Omega))$, $g \in L^2(\Omega)$, there is no guarantee that the solution exists. In the following Theorem, we consider the existence condition of solution to the problem (1) under the following condition on *f* and *g*.

Theorem 2.2. *The problem (1) has a unique solution u if and only if*

$$\sum_{p=1}^{\infty} \exp\left(2\lambda_p \int_0^T b(s)ds\right) \times \left[g_p - \int_0^T \exp\left(-\lambda_p \int_s^T b(\xi)d\xi\right) f_p(s)ds \right]^2 < \infty, \tag{5}$$

where $g_p = \int_\Omega g(x) X_p(x) dx, f_p(x) = \int_\Omega f(x,s) X_p(x) dx.$

Proof: Suppose the problem (1) has an exact solution $u \in C([0,T]; H_0^1(\Omega)) \cap C^1((0,T); L^2(\Omega))$. We have

$$< \frac{\partial u}{\partial t}, X_p(.) > = b(t) < \Delta u, X_p(.) > + < f(.,t), X_p(.) > . \tag{6}$$

Integrating by parts, we have

$$< \Delta u, X_p(.) > = \int_\Omega u(x,t) X''_p(x) dx \\ = -\lambda_p < u(.,t), X_p(.) >. \quad (7)$$

Combining (6) and (7), we obtain $u'_p(t) + b(t)\lambda_p u_p(t) = f_p(t)$ which is equivalent to

$$\left[e^{-\lambda_p \int_t^T b(\xi)d\xi} u_p(t) \right]'(t) \\ = e^{-\lambda_p \int_t^T b(\xi)d\xi} f_p(t). \quad (8)$$

where upon

$$\int_t^T \left[e^{-\lambda_p \int_s^T b(\xi)d\xi} u_p(s) \right]'(s)ds \\ = \int_t^T \exp\left(-\lambda_p \int_s^T b(\xi)d\xi\right) f_p(s)ds, \quad (9)$$

or

$$g_p - \exp\left(-\lambda_p \int_s^T b(\xi)d\xi\right) u_p(t) \\ = \int_t^T \exp\left(-\lambda_p \int_s^T b(\xi)d\xi\right) f_p(s)ds. \quad (10)$$

Hence

$$u_p(t) = < u(.,t), X_p(.) > = \exp\left(\lambda_p \int_t^T b(\xi)d\xi\right) \\ \left[g_p - \int_t^T \exp\left(-\lambda_p \int_s^T b(\xi)d\xi\right) f_p(s)ds \right]. \quad (11)$$

Letting $t = 0$ in (11), we have

$$u_p(0) = \exp\left(\lambda_p \int_0^T b(s)ds\right) \\ \left[g_p - \int_0^T \exp\left(-\lambda_p \int_s^T b(\xi)d\xi\right) f_p(s)ds \right]. \quad (12)$$

Then

$$\| u(.,0) \|^2 = \sum_{p=1}^\infty \exp\left(2\lambda_p \int_0^T b(s)ds\right) \\ \left[g_p - \int_0^T \exp\left(-\lambda_p \int_s^T b(\xi)d\xi\right) f_p(s)ds \right]^2 < \infty. \quad (13)$$

If (5) holds, then we define

$$v(x) = \sum_{p=1}^\infty \exp\left(\lambda_p \int_0^T b(s)ds\right) \\ \left[g_p - \int_0^T \exp\left(-\lambda_p \int_s^T b(\xi)d\xi\right) f_p(s)ds \right] X_p(x). \quad (14)$$

It is easy to see that $v \in L^2(\Omega)$. Then, we consider the problem of finding u from the forward heat problem

$$\begin{cases} \dfrac{\partial u}{\partial t} = b(t)\Delta u + f(x,t), \\ u|_{\partial\Omega} = 0, \ t \in (0,T) \\ u(x,0) = v(x), \ x \in \Omega. \end{cases} \quad (15)$$

The problem (15) is the forward problem so it has a unique solution u (See (Evans 1997)). We have

$$u(x,t) = \sum_{p=1}^\infty [\exp\left(-\lambda_p \int_0^t b(\xi)d\xi\right) < v(x), X_p(x) > \\ + \int_0^t \exp\left(-\lambda_p \int_s^t b(\xi)d\xi\right) f_p(s)ds] X_p(x). \quad (16)$$

Thus

$$u(x,T) = \sum_{p=1}^\infty \left[\exp\left(-\lambda_p \int_0^T b(\xi)d\xi\right) < v(x), X_p(x) > \\ + \int_0^T \exp\left(-\lambda_p \int_s^T b(\xi)d\xi\right) f_p(s)ds \right] X_p(x). \quad (17)$$

Combining (14), (17), and by a simple computation, we get

$$u(x,T) = \sum_{p=1}^\infty g_p X_p(x) = g(x). \quad (18)$$

Hence, u is the unique solution of the problem (1).

Remark 1:

1. *When $b(t) = 1, f(x,t) = 0$, the problem (1) has a unique solution if and only if g satisfies the following strong regularity assumption*

$$\sum_{p=1}^\infty e^{2T\lambda} | < g(.), X_p(.) > |^2 < \infty. \quad (19)$$

This assumption is also given by Lemma 1 in (G.W. Clark & S.F. Oppenheimer 1994).

2. *Some examples of f and g which satisfying (5) are given in the Numerical Experiments Section.*

Theorem 2.3 *The problem (1) has at most one solution in* $C([0,T];H_0^1(\Omega))\cap C^1((0,T);L^2(\Omega))$. *If (1) has a solution u, then it is defined by*

$$u(x,t)=\sum_{p=1}^{\infty}\exp\left(\lambda_p\int_t^T b(\xi)d\xi\right)\left[g_p-\int_t^T\exp\left(-\lambda_p\int_s^T b(\xi)d\xi\right)f_p(s)ds\right]X_p(x). \quad (20)$$

In spite of the uniqueness, the problem (1) is still ill-posed and some regularization methods are necessary. In next sections, we propose two approximating problems.

Proof: The proof of this Theorem is divide into two steps.

Step 1. The problem (1) has at most one solution.

Let $u(x, t)$, $v(x, t)$ be two solutions of the problem (1) such that u, $v \in C([0, T]; H_0^1(\Omega))\cap C^1((0,T);L^2(\Omega))$. Put $w(x, t) = u(x, t) - v(x, t)$; then w satisfies the problem

$$\begin{cases} w_t-b(t)\Delta w=0, \\ w|_{\partial\Omega}=0,\ t\in(0,T), \\ w(x,T)=0,\ x\in\Omega. \end{cases} \quad (21)$$

Now, setting $G(t)=\int_\Omega w^2(x,t)dx\ (0\le t\le T)$, and by taking the derivative of $G(t)$, we have

$$\begin{aligned} G'(t)&=2\int_\Omega w(x,t).w_t(x,t)dx \\ &=2b(t)\int_\Omega w(x,t).\Delta w(x,t)dx \\ &=2\int_\Omega w(x,t).w_t(x,t)dx. \end{aligned} \quad (22)$$

Using the Green formula, we obtain

$$G'(t)=-2b(t)\int_\Omega(\nabla w(x,t))^2dx. \quad (23)$$

Hence

$$\begin{aligned} G''(t)=&-4b(t)\int_\Omega\nabla w(x,t)\cdot\nabla w_t(x,t)dx \\ &-2b'(t)\int_\Omega(\nabla w(x,t))^2dx. \end{aligned} \quad (24)$$

Moreover, using the integration by parts and $w_t(x,t)=b(t)\Delta w(x,t)$, we get

$$\begin{aligned} -4b(t)&\int_\Omega\nabla w(x,t)\cdot\nabla w_t(x,t)dx \\ &=4b(t)\int_\Omega\Delta w(x,t)\cdot w_t(x,t) \\ &=4b^2(t)\int_\Omega(\Delta w(x,t))^2dx. \end{aligned} \quad (25)$$

So

$$\begin{aligned} G''(t)&=4b^2(t)\int_\Omega(\Delta w(x,t))^2dx-2b'(t)\int_\Omega(\nabla w(x,t))^2dx \\ &=4\int_\Omega w_t^2(x,t)dx-2b'(t)\int_\Omega(\nabla w(x,t))^2dx. \end{aligned} \quad (26)$$

We have

$$\begin{aligned} G(t)&G''(t)-(G'(t))^2-\frac{C_1}{B_1}G(t)G'(t) \\ &=4\int_\Omega w^2dx\int_\Omega w_t^2dx \\ &\quad-2b'(t)\int_\Omega w^2dx\int_\Omega(\nabla w(x,t))^2dx \\ &\quad-4\left(\int_\Omega w(x,t)w_t(x,t)dx\right)^2 \\ &\quad+2\frac{C_1}{B_1}b(t)\int_\Omega w^2dx\int_\Omega(\nabla w(x,t))^2dx \\ &=4\int_\Omega w^2dx\int_\Omega w_t^2dx-4\left(\int_\Omega w(x,t).w_t(x,t)dx\right)^2 \\ &\quad+2\left(\frac{C_1}{B_1}b(t)-b'(t)\right)\int_\Omega w^2dx\int_\Omega(\nabla w(x,t))^2dx. \end{aligned} \quad (27)$$

Using the Hölder inequality, we have

$$\begin{aligned} 4&\int_\Omega w^2dx\int_\Omega w_t^2dx \\ &-4\left(\int_\Omega w(x,t)w_t(x,t)dx\right)^2\ge 0. \end{aligned} \quad (28)$$

Since $0<B_1\le b(t),\ 0<b'(t)\le C_1$, we get $\frac{C_1}{B_1}b(t)-b'(t)\ge 0$. Then

$$G(t)G''(t)-(G'(t))^2-\frac{C_1}{B_1}G(t)G'(t)\ge 0. \quad (29)$$

We define the function $m(t)=e^{\frac{C_1}{B_1}t}$, and then regard G as a function of m. Let us introduce an auxiliary function

$$F(m)=\ln[G(t(m))]. \quad (30)$$

Since $t=\frac{B_1}{C_1}\ln m$, we have

$$F'(m)=\frac{G'(t(m))t'(m)}{G(t(m))}=\frac{B_1}{C_1m}\frac{G'(t(m))}{G(t(m))}. \quad (31)$$

and

$$F''(m)=\frac{-B_1}{C_1m^2}\frac{G'(t(m))}{G(t(m))}+\frac{B_1^2}{C_1^2m^2}\times\frac{G(t(m))G''(t(m))-[G'(t(m))]^2}{G^2(t(m))}$$

$$= \frac{G(t(m))G''(t(m)) - [G'(t(m))]^2}{\left(G(t(m))\frac{C_1}{B_1}m\right)^2} - \frac{-\frac{C_1}{B_1}G(t(m))G'(t(m))}{\left(G(t(m))\frac{C_1}{B_1}m\right)^2} \tag{32}$$

Using (29) and (32), we obtain $F''(m) \geq 0$. Hence F is a convex function on the interval $1 \leq m \leq m_1$ with $m_1 = e^{\frac{C_1}{B_1}T}$. According to the convex property of function $F(m)$, we have

$$F(m) \leq \frac{m-1}{m_1 - 1}F(m_1) + \frac{m_1 - m}{m_1 - 1}F(1). \tag{33}$$

In addition, from (30), inequality (33) is equivalent to

$$G(t) \leq [G(T)]^{\frac{m-1}{m_1-1}}[G(0)]^{\frac{m_1-m}{m_1-1}}. \tag{34}$$

Since $G(T) = 0$, we conclude that $G(t) = 0$ for $0 \leq t \leq T$. This implies that $u(x,t) = v(x,t)$. The proof of the step 1 is completed.

Step 2. The problem (1) has a solution which is defined in (20).

Using (11), we have (20).

In spite of the uniqueness, the problem (1) is still ill-posed and some regularization methods are necessary. In next sections, we propose two approximating problems.

3 A MODIFIED QUASI-BOUNDARY VALUE METHOD AND ERROR ESTIMATES IN L²

In practice, we get the given data g by measuring at discrete data. Hence, instead of g, we shall get an inexact data $g^{\in} \in L^2(\Omega)$ satisfying

$$\| g^{\in} - g \| \leq \in. \tag{35}$$

In this section, we shall regularize the problem (1) in the following one

$$\frac{\partial u^{\in}}{\partial t} = b(t)\Delta u^{\in} + \sum_{p=1}^{\infty} \frac{\exp\left(-\lambda_p \int_0^T b(\xi)d\xi\right) f_p(t)X_p(x)}{\in \lambda^k + \exp\left(-\lambda_p \int_0^T b(\xi)d\xi\right)}, \tag{36}$$
$$(x,t) \in \Omega \times (0,T)$$

$$u^{\in}|_{\partial\Omega} = 0, \quad t \in (0,T),$$
$$u^{\in}(x,T) = \sum_{p=1}^{\infty} \frac{\exp\left(-\lambda_p \int_0^T b(\xi)d\xi\right)}{\in \lambda^k + \exp\left(-\lambda_p \int_0^T b(\xi)d\xi\right)} g_p X_p(x), \quad x \in \Omega. \tag{37}$$

where $0 < \in < 1$, $k \geq 1$ and $f_p(t), g_p$ are defined by

$$f_p(t) = \int_\Omega f(x,t)X_p(x)dx, \qquad g_p = \int_\Omega g(x)X_p(x)dx. \tag{38}$$

The idea of the problem (36) has a long mathematical history going back to (Showalter 1983, G.W. Clark & S.F. Oppenheimer 1994, Denche & Bessila 2005). Adding an appropriate "corrector" into the given data $u(x,T)$ is a key idea in the theory of the quasi-boundary value method (or modified quasi-boundary value method). Using this method, Clark and Oppenheimer (G.W. Clark & S.F. Oppenheimer 1994), and Denche and Bessila (Denche & Bessila 2005), regularized a similar backward problem by replacing the given condition by

$$u(T) + \in u(0) = g \tag{39}$$

and

$$u(T) - \in u'(0) = g, \tag{40}$$

respectively. Tuan and Trong (Trong & Tuan 2008) presented a different perturbation of g by a new term $u(x,T) = A(\in, T)g$, where $A(\in, g)$ satisfies some suitable conditions. The problem (36) is a generalized version of the regularized problem given in (Trong & Tuan 2008).

In the next Theorem, we shall study the existence, uniqueness and stability of a (weak) solution of the problem (36).

Theorem 3.1 *The problem (46) has a unique solution $u^{\in} \in C([0, T]; L^2(\Omega)) \cap L^2((0, T); H_0^1(\Omega)) \cap C^1((0,T); H_0^1(\Omega))$. The solution depends continuously on g in $L^2(\Omega)$.*

In Step 2, the stability of the solution is given. First, we state the following Lemma

Proof:

The proof is divided into two steps. In Step 1, the existence and the uniqueness of a solution of (36) is showed; the (unique) solution $u^{\in}$ of problem (36) is given by

$$u^{\in}(x,t)=\sum_{p=1}^{\infty}\frac{\exp\left(-\lambda_p\int_0^t b(\xi)d\xi\right)}{\in\lambda^k+\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)}\times\left(g_p-\int_t^T e^{-\lambda_p\int_s^T b(\xi)d\xi}f_p(s)ds\right)X_p(x). \tag{41}$$

In Step 2, the stability of the solution is given. First, we state the following Lemma

Lemma 3.1

For $M,\in,x>0,\ k\ \geq\ 1,$ we have the following inequality

$$\frac{1}{\in x^k+e^{-Mx}}\leq\frac{(kM)^k}{\in\ln^k(\frac{M^k}{k\in})}. \tag{42}$$

Proof: Let $f(x)=\frac{1}{\in x^k+e^{-Mx}}$, we have

$$f'(x)=\frac{\in kx^{k-1}-Me^{-Mx}}{-(\in x^k+e^{-Mx})^2}. \tag{43}$$

The equation $f'(x)=0$ gives a unique solution x_0 such that $\in kx_0^{k-1}-Me^{-Mx_0}=0$. It means that $x_0^{k-1}e^{Mx_0}=\frac{M}{k\in}$. Thus the function f achieves its maximum at a unique point $x=x_0$. Hence

$$f(x)\leq\frac{1}{\in x_0^k+e^{-Mx_0}}. \tag{44}$$

Since $e^{-Mx_0}=\frac{k\in}{M}x_0^{k-1}$, we have

$$f(x)\leq\frac{1}{\in x_0^k+e^{-Mx_0}}\leq\frac{1}{\in x_0^k+\frac{k\in}{M}x_0^{k-1}}. \tag{45}$$

By using the inequality $e^{Mx_0}\geq Mx_0$, we get

$$\begin{aligned}\frac{M}{k\in}&=x_0^{k-1}e^{Mx_0}\leq\frac{1}{M^{k-1}}e^{(k-1)Mx_0}e^{Mx_0}\\&=\frac{1}{M^{k-1}}e^{kMx_0}.\end{aligned} \tag{46}$$

This gives $e^{kMx_0}\geq\frac{M^k}{k\varepsilon}$ or $kMx_0\geq\ln(\frac{M^k}{k\in})$. Therefore $x_0\geq\frac{1}{kM}\ln(\frac{M^k}{k\in})$. Hence, we obtain

$$f(x)\leq\frac{1}{\in x_0^k}\leq\frac{(kM)^k}{\in\ln^k(\frac{M^k}{k\varepsilon})}. \tag{47}$$

The Lemma is completely proved, and Now we pass to the proof of Theorem 3.1. Denote W = $C([0,\ T];\ L^2(\Omega))\ \cap\ L^2((0,\ T);\ H_0^1(\Omega))\cap C^1((0,T);H_0^1(\Omega))$.

Step 1. The existence and the uniqueness of a solution of the problem (36). We divide this step into two parts.

Part A If $u^{\in}\in W$ satisfies (51) then $u^{\in}$ is solution of the problem (36). We can verify directly that $u^{\in}\in W$. We have

$$\begin{aligned}&<u_t^{\in}(.,t),X_p(.)>\\&=\frac{-\lambda b(t)\exp\left(-\lambda_p\int_0^t b(\xi)d\xi\right)}{\in\lambda^k+\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)}\\&\quad\times\left(g_p-\int_t^T e^{-\lambda\int_s^T b(\xi)d\xi}f_p(s)ds\right)\\&\quad+\frac{\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)}{\in\lambda^k+\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)}f_p(t)\\&=-\lambda b(t)<u^{\in}(.,t),X_p(.)>\\&\quad+\frac{\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)}{\in\lambda^k+\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)}f_p(t)\\&=b(t)<\Delta u^{\in}(.,t),X_p(.)>\\&\quad+\frac{\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)}{\in\lambda^k+\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)}f_p(t).\end{aligned} \tag{48}$$

This implies that

$$u_t^{\in}=b(t)\Delta u^{\in}+\sum_{p=1}^{\infty}\frac{\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)}{\in\lambda^k+\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)}f_p(t)X_p(x). \tag{49}$$

By letting $t=T$ in (41), we get

$$u^{\in}(x,T)=\sum_{p=1}^{\infty}\frac{\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)}{\in\lambda^k+\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)}g_pX_p(x).$$

Therefore, $u^{\in}$ is the solution of the problem (36).

Part B. The problem (36) has at most one solution in W.

We can prove this part in a similar way as in the step 1 of Theorem 2.2. This ends.

Step 2. The solution of the problem (36) depends continuously on g in $L^2(\Omega)$.

Let w and v be two solutions of (36) corresponding to the given values g and h.

From (41), we have

$$w(x,t)=\sum_{p=1}^{\infty}\frac{\exp\left(-\lambda_p\int_0^t b(\xi)d\xi\right)}{\in\lambda^k+\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)} \times\left(g_p-\int_t^T e^{-\lambda_p\int_s^T b(\xi)d\xi}f_p(s)ds\right)X_p(x). \tag{50}$$

$$v(x,t)=\sum_{p=1}^{\infty}\frac{\exp\left(-\lambda_p\int_0^t b(\xi)d\xi\right)}{\in\lambda^k+\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)} \times\left(h_p-\int_t^T e^{-\lambda_p\int_s^T b(\xi)d\xi}f_p(s)ds\right)X_p(x). \tag{51}$$

where $g_p = \int_\Omega g(x)X_p(x)dx$, $h_p(x) = \int_\Omega h(x) X_p(x)dx$. It follows that

$$\begin{aligned}\|w(.,t)-v(.,t)\|^2 &= \sum_{p=1}^{\infty}\left|\frac{\exp\left(-\lambda_p\int_0^t b(\xi)d\xi\right)}{\in\lambda^k+\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)}(g_p-h_p)\right|^2 \\ &\le \sum_{p=1}^{\infty}\left|\frac{1}{\in\lambda^k+\exp(-B_2T\lambda)}(g_p-h_p)\right|^2.\end{aligned} \tag{52}$$

Using the inequality

$$\begin{aligned}\frac{1}{\in x^k+e^{-B_2Tx}} &\le (kTB_2)^k\in^{-1}\left(\ln\left(\frac{(B_2T)^k}{k\in}\right)\right)^{-k} \\ &= B_4\in^{-1}\left(\ln\left(\frac{(B_3)}{\in}\right)\right)^{-k}\end{aligned} \tag{53}$$

where $B_3=\frac{(B_2T)^k}{k}$, $B_4=(kB_2T)^k$, we conclude that

$$\|w(.,t)-v(.,t)\|_{L^2}\le B_4\in^{-1}\left(\ln\left(\frac{B_3}{\in}\right)\right)^{-k}\|g-h\|_{L^2}. \tag{54}$$

This ends the proof of the step 2 and the proof of Theorem 3.1.

Theorem 3.2 *Let* $g\in S^k(\Omega)$ *for* $k>0$. *Then we have*

$$\|u^{\in}(.,T)-g(.)\|_{L^2}\le B_4\left(\ln\left(\frac{B_3}{\varepsilon}\right)\right)^{-k}\|g\|_{S^k}.$$

Proof: Let $\alpha>0$. Since $g(x)=\sum_{p=1}^{\infty}g_pX_p(x)$, then there exists a positive integer N for which $\sum_{p=N+1}^{\infty}g_p^2<\alpha/2$. We have

$$\|u^{\varepsilon}(x,T)-g(x)\|_{L^2}^2 = \sum_{p=1}^{\infty}\frac{\in^2\lambda_p^{2k}g_p^2}{\left(\in\lambda^k+\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)\right)^2}. \tag{55}$$

Using the following estimate

$$\begin{aligned}&\left(\in\lambda^k+\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)\right)^2 \\ &> \in^2\lambda^{2k}+\exp\left(-2\lambda\int_0^T b(s)ds\right) \\ &> \exp\left(-2\lambda\int_0^T b(s)ds\right)\end{aligned} \tag{56}$$

we get

$$\|u^{\in}(x,T)-g(x)\|_{L^2}^2 \le \in^2\sum_{p=1}^{N}\lambda^{2k}g_p^2\exp\left(2\lambda\int_0^T b(s)ds\right)+\frac{\alpha}{2}. \tag{57}$$

By taking $\in$ such that

$$\in<\sqrt{\frac{\alpha}{2\pi\sum_{p=1}^{N}\lambda^{2k}g_p^2\exp\left(2\lambda\int_0^T b(s)ds\right)}},$$

we obtain $\|u^{\in}(x,T)-g(x)\|_{L^2}^2<\alpha$.

By using (53), we have the following error estimate

$$\begin{aligned}\|u^{\varepsilon}(x,T)-g(x)\|_{L^2}^2 &= \sum_{p=1}^{\infty}\frac{\varepsilon^2\lambda^{2k}g_p^2}{\left(\in\lambda^k+\exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)\right)^2} \\ &\le \sum_{p=1}^{\infty}\frac{\in^2\lambda^{2k}g_p^2}{(\in\lambda^k+e^{-TB_2\lambda})^2} \\ &\le \in^2B_4^2\in^{-2}\left(\ln\left(\frac{(B_3)}{\in}\right)\right)^{-2k}\sum_{p=1}^{\infty}\lambda^{2k}g_p^2 \\ &= B_4^2\left(\ln\left(\frac{B_3}{\in}\right)\right)^{-2k}\|g\|_{S^k}^2.\end{aligned}$$

where upon

$$\|u^{\in}(x,T)-g(x)\|_{L^2}\le B_4\left(\ln\left(\frac{B_3}{\varepsilon}\right)\right)^{-k}\|g\|_{S^k}.$$

Theorem 3.3

Let $g\in L^2(\Omega)$ *be as Theorem 3.2, and* $\int_0^T\|f(.,s)\|_{S^k}^2\,ds<\infty$. *If* $u^{\in}(x,0)$ *converges in*

$L^2(\Omega)$, then the problem (1) has a unique solution u. Furthermore, the regularized solution $u^{\in}(x,t)$ converges to $u(t)$ uniformly in t as $\in$ tends to zero.

Proof: Assume that $\lim_{\in\to 0} u^{\in}(x,0) = u_0(x)$ exists. Let

$$u(x,t) = \sum_{p=1}^{\infty}[\exp\left(-\lambda_p\int_0^t b(\xi)d\xi\right)u_{0p} + \int_0^t \exp\left(-\lambda_p\int_s^t b(\xi)d\xi\right)f_p(s)ds]X_p(x) \tag{58}$$

where $u_{0p} = \int_\Omega u_0(x)X_p(x)dx$, $f_p(s) = <f(x, s), X_p(x)>$. It is easy to check that u satisfies $u_t = b(t)\Delta u + f(x,t)$ and $u(x, t) = 0$ for $x \in \partial\Omega$. We will prove that $u(x,T) = g(x)$. Using the inequality $(a+b)^2 \le 2(a^2+b^2)$, we have

$$\begin{aligned}
&\| u^{\in}(.,t) - u(.,t)\|_{L^2}^2 \\
&\le \sum_{p=1}^{\infty}\exp(-2\lambda\int_0^t b(s)ds)(u_{0p}^{\in} - u_{0p})^2 \\
&+ t^2\sum_{p=1}^{\infty}\left[\int_0^t \exp\left(-2\lambda_p\int_s^t b(\xi)d\xi\right)\right. \\
&\left.\frac{\in^2\lambda^{2k}}{\left(\in\lambda^k + \exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)\right)^2}f_p^2(s)ds\right] \qquad (59)\\
&\le \| u^{\varepsilon}(.,0) - u_0(.)\|_{L^2}^2 \\
&+ T^2\int_0^t\sum_{p=1}^{\infty}\frac{\in^2}{(\in\lambda^k + e^{-TB_2\lambda})^2}\lambda^{2k}f_p^2(s)ds \\
&\le \| u^{\varepsilon}(.,0) - u_0(.)\|_{L^2}^2 \\
&+ B_4^2\left(\ln\left(\frac{B_3}{\varepsilon}\right)\right)^{-2k}\int_0^T \| f(.,s)\|_{S^k}^2\ ds.
\end{aligned}$$

Hence $\lim_{\in\to 0}\| u^{\in}(.,t) - u(.,t)\|_{L^2} = 0$. Thus $\lim_{\in\to 0}\| u^{\in}(.,T) - u(.,T)\|_{L^2} = 0$. Using the Theorem 3.2, we have $\lim_{\in\to 0}\| u^{\in}(.,T) - g(.)\|_{L^2} = 0$. Therefore, $u(x,T) = g(x)$. Hence, $u(x,t)$ is the unique solution of the problem (1). From (59), we also conclude that $u^{\in}(x,t)$ converges to $u(x,t)$ uniformly in t.

Theorem 3.4 *Let $f \in L^2(0,T; L^2(\Omega))$ and $g \in L^2(\Omega)$. Suppose that the problem (1) has a unique solution $u(x,t)$ in $C([0, T]; H_0^1(\Omega)) \cap C^1((0,T); L^2(\Omega))$ which satisfies $u(., t) \in S^k(\Omega)$ for any $t \in [0,T]$. Let $g^{\in} \in L^2(\Omega)$ be a measured data such that $\| g^{\in} - g\| \le \in$ Then there exists a function $v^{\in}$ satisfying*

$$\| u(.,t) - v^{\in}(.,t)\|_{L^2} \le [C+1]\left(\ln\left(\frac{B_3}{\in}\right)\right)^{-k} \tag{60}$$

for every $t \in [0, T]$ and $C = B_4\sup_{t\in[0,T]}\| u(.,t)\|_{S^k(\Omega)}$.

Proof: Let $u^{\in}$ be the solution of the problem (36) corresponding to g, and let $v^{\in}$ be the solution of the problem (46) corresponding to $g^{\in}$, where g and $g^{\in}$ are in right hand side of (36). Using (20) and (41), we get

$$\begin{aligned}
&|<u(.,t) - u^{\in}(.,t), X_p(.)>| \\
&= \frac{\in\lambda^k\exp\left(\lambda_p\int_t^T b(\xi)d\xi\right)}{\in\lambda^k + \exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)} \\
&\quad\left|g_p - \int_t^T \exp\left(-\lambda_p\int_s^T b(\xi)d\xi\right)f_p(s)ds)\right| \\
&\le \frac{\in}{\in\lambda^k + \exp\left(-\lambda_p\int_0^T b(\xi)d\xi\right)} \\
&\quad\left|\lambda^k\exp\left(\lambda_p\int_t^T b(\xi)d\xi\right)g_p\right. \\
&\quad\left.-\int_t^T \lambda^k\exp\left(\lambda_p\int_t^s b(\xi)d\xi\right)f_p(s)ds)\right| \\
&\le \frac{\in}{\in\lambda^k + e^{-B_2T\lambda}}\left|\lambda^k\exp\left(\lambda_p\int_t^T b(\xi)d\xi\right)g_p\right. \\
&\quad\left.-\int_t^T \lambda^k\exp\left(\lambda_p\int_t^s b(\xi)d\xi\right)f_p(s)ds)\right|.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\| u(.,t) - u^{\varepsilon}(.,t)\|_{L^2}^2 \\
&= \sum_{p=1}^{\infty}|<u(.,t) - u^{\varepsilon}(.,t), X_p(.)>|^2 \\
&\le B_4^2\left(\ln\left(\frac{B_3}{\in}\right)\right)^{-2k}\sum_{p=1}^{\infty}\lambda^{2k}|<u(x,t), X_p(x)>|^2 \\
&\le C^2\left(\ln\left(\frac{B_3}{\in}\right)\right)^{-2k}.
\end{aligned} \tag{61}$$

Hence

$$\| u(.,t) - u^{\in}(.,t)\|_{L^2} \le C\left(\ln\left(\frac{B_3}{\varepsilon}\right)\right)^{-k}. \tag{62}$$

Using (62) and Step 2 of Theorem 3.1, we get

$$\begin{aligned}
&\| v^{\in}(.,t) - u(.,t)\|_{L^2} \\
&\le \| v^{\in}(.,t) - u^{\in}(.,t)\|_{L^2} + \| u^{\in}(.,t) - u(.,t)\|_{L^2} \\
&\le \in^{-1}\left(\ln\left(\frac{B_3}{\varepsilon}\right)\right)^{-k}\| g^{\in} - g\| + C\left(\ln\left(\frac{B_3}{\varepsilon}\right)\right)^{-k}
\end{aligned} \tag{63}$$

$$\leq [C+1]\left(\ln\left(\frac{B_3}{\varepsilon}\right)\right)^{-k}$$

for every $t \in (0,T)$ and where C is defined in Theorem 3.4. This ends the proof of Theorem 3.4.

Remark 2.

- If $k=1$ then the condition $u(.,t) \in S^k(\Omega)$ is equivalent to the condition $\Delta u(.,t) \in L^2(\Omega)$. Hence, this condition is natural and acceptable.
- To estimate the error in higher Sobolev spaces such as H^1 and H^2, we can not continue to use the modified quasi-boundary value method. We present a truncation method in the next Section.

4 REGULARIZATION BY THE TRUNCATION METHOD AND ERROR ESTIMATES IN L^2, H^1, H^2

Suppose that the problem (1) has an exact solution $u \in C([0,T]; H_0^1(\Omega)) \cap C^1((0,T); L^2(\Omega))$, according to (20), we have

$$u(x,t) = \sum_{p=1}^{\infty} \exp\left(\lambda_p \int_t^T b(\xi)d\xi\right) \times \left[g_p - \int_t^T \exp\left(-\lambda_p \int_s^T b(\xi)d\xi\right) f_p(s)ds\right] X_p(x) \tag{64}$$

Let

$$R_\in = \{p \geq 1, p \in N, \lambda_p \leq M_\in\},$$
$$Q_\in = \{p \geq 1, p \in N, \lambda_p > M_\in\}. \tag{65}$$

In spite of the uniqueness, the problem is still ill-posed, and a regularization is necessary. For each $\in > 0$, we introduce the truncation mapping $P_\in : L^1(\Omega) \to C^\infty(\Omega) \cap H_0^1(\Omega)$

$$P_\in w(x) = \sum_{p \in R_\in} < w, X_p(x) > X_p(x). \tag{66}$$

In fact, $P_\in$ is a finite-dimensional orthogonal projection on $L^2(\Omega)$. We shall approximate the original problem by the following well-posed problem.

Theorem 4.1

For each $f \in L^2((0,T); L^2(\Omega))$ and $g \in L^2(\Omega)$, let $w \in L^2((0,T); L^2(\Omega))$ be defined by

$$< w(x,t)X_p(x) > = \exp\left(\lambda_p \int_t^T b(\xi)d\xi\right)$$
$$\Big[< P_\in g, X_p(x) > - \int_t^T \exp\left(-\lambda_p \int_s^T b(\xi)d\xi\right) \tag{67}$$
$$< P_\in f(x,s), X_p(x) > ds\Big]$$

for any $p \geq 1$. Then $w = P_\in w$ and it depends continuously on g, i.e. if w_i is the solution with respect to g_i, $i = 1, 2$, then

$$\|w_1(t) - w_2(t)\|_{L^2(\Omega)} \leq e^{B_2(T-t)M_\in} \|g_1 - g_2\|_{L^2(\Omega)}$$

Proof: Note that $w(t)$ is well-defined because $< w(t), X_p(x) > = 0$ if $p \in Q_\in$. This fact also implies that $w = P_\in w$. Now for two solutions w_1, w_2 we have

$$\|w_1(.,t) - w_2(.,t)\|^2_{L^2(\Omega)}$$
$$= \sum_{p \in R_\in} \left| < w_1(t) - w_2(t), X_p(x) > \right|^2 \tag{68}$$

$$= \sum_{p \in R_\in} \left| \exp\left(\lambda_p \int_t^T b(\xi)d\xi\right) < g_1 - g_2, X_p(x) > \right|^2$$
$$\leq e^{2B_2(T-t)M_\in} \|g_1(.) - g_2(.)\|^2_{L^2(\Omega)}. \tag{69}$$

Recalling the value of $M_\in$, we have the desired estimate.

Theorem 4.2

Assume that the problem (1) has at most one (weak) solution $u \in C\big([0,T]; L^2(\Omega)\big) \cap C^1((0,T);$ $L^2(\Omega))$ corresponding to $f \in L^2((0,T); L^2(\Omega))$ and $g \in L^2(\Omega)$. Let $g_\in$ be measured data such that

$$\|g_\in - g\|_{L^2(\Omega)} \leq \in .$$

Define the regularized solution $u_\in \in L^2((0,T);$ $L^2(\Omega))$ from $g_\in$ as in (67). Then for each $t \in [0,T]$, $u_\in(t) \in C^\infty(\Omega) \cap H_0^1(\Omega)$ and $\lim_{\varepsilon \to 0} u_\varepsilon(.,t) = u(.,t)$ in $L^2(\Omega)$ if we choose $M_\in = \frac{\ln(\varepsilon^{-1})}{2TB_2}$.

Proof:

Note that $u_\in(t) = P_\in u_\in(t) \in C^\infty(\Omega) \cap H_0^1(\Omega)$ as in Remark 1. Moreover using the stability in Theorem 4.1 we find that

$$\|u_\in(.,t) - u(.,t)\|_{L^2(\Omega)}$$
$$\leq \|P_\in u_\in(t) - P_\in u(t)\|_{L^2(\Omega)} + \|P_\in u(t) - u(t)\|_{L^2(\Omega)} \tag{71}$$
$$\leq \varepsilon^{\frac{t}{2T}} + \sqrt{\sum_{p \in Q_\in} |< u(.,t), X_p(.) >|^2}$$

Note that $\varepsilon^{\frac{t}{T}}$ converges to zero as $\in \to 0$ and $t > 0$. To obtain the convergence of the second term in the right-hand side of (71), we note that

$$\sum_{p \in Q_\in} |< u(.,t), X_p(.) >|^2 \leq \|u(.,t)\|^2_{L^2(\Omega)} < \infty.$$

and $M_\in \to \infty$ as $\in \to 0$.

In the above theorem, we have not given an error estimate because the condition of the exact solution u is so weak (we even did not require $u(t) \in H_0^1(\Omega)$). And the error estimate at $t = 0$ is useless. However in practical application we may expect that the exact solution is smoother. In such cases, many explicit error estimates are shown in the next section. An essential point should be stated that the regularized solution is the same in any case. This is a substantial usefulless for practical applications because even if we do not know how good the exact solution is, we are always ensured that the regularized solution still works without any further adjustment.

From the usual viewpoint of variational method, it is natural to assume that $u(.,t) \in H_0^1(\Omega)$ for all $t \in [0,T]$. Moreover, if f is smooth and u is a classical solution of the heat equation (1), then $u(.,t) \in H^2(\Omega) \cap H_0^1(\Omega)$ for all $t \in [0,T]$. For these two situations we have the following explicit error estimates.

Theorem 4.3

Let u, $u_\in$ be as in Theorem 4.2, and let $t \in [0,T]$. Let us choose $M_\in = \frac{\ln(\in^{-1})}{2TB_2}$.

i. *Assume that $u(.,t) \in H_0^1(\Omega)$. Then $\lim\limits_{\varepsilon\to 0} u_\varepsilon(.,t) = u(.,t)$ in $H_0^1(\Omega)$ and*

$$\|u_\varepsilon(.,t) - u(.,t)\|_{L^2(\Omega)} \le \varepsilon^{\frac{t}{2T}} \sqrt{\frac{\ln(\frac{1}{\varepsilon})}{2TB_2} + \frac{\sqrt{2TB_2}}{\sqrt{\ln(\varepsilon^{-1})}} \|\nabla u(.,t)\|_{L^2(\Omega)}}. \tag{72}$$

ii. *Assume that $u(.,t) \in H^2(\Omega) \cap H_0^1(\Omega)$. Then $\lim\limits_{\varepsilon\to 0} u_\in(.,t) = u(.,t)$ in $H^2(\Omega)$ and*

$$\|u_\in(.,t) - u(.,t)\|_{L^2(\Omega)} \le \varepsilon^{\frac{t}{2T}} \sqrt{\frac{\ln(\frac{1}{\varepsilon})}{2TB_2} + \frac{2TB_2}{\ln(\varepsilon^{-1})} \|u(.,t)\|_{H^2(\Omega)}} \tag{73}$$

$$\|u_\varepsilon(.,t) - u(.,t)\|_{H_0^1(\Omega)} \le \varepsilon^{\frac{t}{2T}} \sqrt{\frac{\ln(\frac{1}{\varepsilon})}{2TB_2} + \frac{\sqrt{2TB_2}}{\sqrt{\ln(\varepsilon^{-1})}} \|u(.,t)\|^2_{H^2(\Omega)}}. \tag{74}$$

Here we use the norms

$$\|w\|^2_{H_0^1} = \|\nabla w\|^2_{L^2}, \tag{75}$$

$$\|w\|^2_{H^2} = \|w\|^2_{L^2} + \|w\|^2_{H_0^1} + \|\Delta w\|^2_{L^2}. \tag{76}$$

Proof:

i. By using the integration by parts and the Parseval's equality, it is straightforward to check that if $u(t) \in H_0^1(\Omega)$ then

$$\|\nabla u(.,t)\|^2_{L^2(\Omega)} = \sum_{p=1}^{\infty} \lambda \,|< u(.,t), X_p(.) >|^2. \tag{77}$$

Using (77) we have

$$\begin{aligned} &\sum_{p\ge 1}^{\lambda > M_\varepsilon} |< u(.,t), X_p(.) >|^2 \\ &\le \frac{1}{M_\varepsilon} \sum_{p=1}^{\infty} \lambda \,|< u(.,t), X_p(.) >|^2 \\ &= \frac{1}{M_\varepsilon} \|\nabla u(.,t)\|^2_{L^2(\Omega)}. \end{aligned} \tag{78}$$

This implies that

$$\sqrt{\sum_{p\in Q_\varepsilon} |< u(.,t), X_p(.) >|^2} \le \frac{1}{\sqrt{M_\varepsilon}} \|\nabla u(.,t)\|_{L^2(\Omega)}.$$

Substituting the latter inequality into the estimate (71) in the proof of Theorem 4.2, we obtain the error estimate in L^2.

To prove the convergence in H_0^1 we use the identity (77) and the stability of Theorem 2 again

$$\begin{aligned} &\|\nabla u_\varepsilon(.,t) - \nabla u(.,t)\|^2_{L^2(\Omega)} \\ &= \sum_{p=1}^{\infty} \lambda \,|< u_\in(.,t) - u(.,t), X_p(x) >|^2 \\ &= \sum_{p=1}^{\lambda \le M_\in} \lambda \,|< u_\in(.,t) - u(.,t), X_p(.) >|^2 \\ &\quad + \sum_{p=1}^{\lambda > M_\in} \lambda \,|< u(.,t), X_p(.) >|^2 \\ &\le \sum_{p=1}^{\lambda \le M_\in} M_\varepsilon |< P_\varepsilon u_\varepsilon(.,t) - P_\varepsilon u(.,t), X_p(.) >|^2 \\ &\quad + \sum_{p=1}^{\lambda > M_\in} \lambda \,|< u(.,t), X_p(.) >|^2 \\ &\le M_\varepsilon \,\|\, P_\varepsilon u_\varepsilon(.,t) - P_\varepsilon u(.,t) \,\|^2_{L^2(\Omega)} \\ &\quad + \sum_{p=1}^{\lambda_p > M_\in} \lambda \,|< u(.,t), X_p(.) >|^2 \\ &\le \varepsilon^{\frac{t}{T}} \frac{\ln(\frac{1}{\varepsilon})}{2TB_2} + \sum_{p=1}^{\lambda > M_\in} \lambda \,|< u(.,t), X_p(.) >|^2. \end{aligned} \tag{79}$$

The second term in the right-hand side of (79) converges to 0 as $\in \to 0$ because of the convergence in (77). Thus the convergence in H_0^1 has been proved.

ii. We now assume that $u(.,t) \in H^2(\Omega) \cap H_0^1(\Omega)$. We have an identity similar to (77)

$$\sum_{p\ge 1} \lambda^2 \,|< u(.,t), X_p(.) >|^2 = \|\Delta u(.,t)\|^2_{L^2(\Omega)}. \tag{80}$$

The error estimate in $L^2(\Omega)$ follows (71) and the following inequality

$$\begin{aligned}&\sum_{p\ge1}^{\lambda>M_\varepsilon} |<u(.,t), X_p(.)>|^2 \\ &\quad\le \frac{1}{M_\varepsilon^2}\sum_{p\ge1}\lambda^2 |<u(.,t), X_p(.)>|^2 \\ &\quad\le \frac{1}{M_\varepsilon^2}\|u(.,t)\|^2_{H^2(\Omega)}.\end{aligned} \tag{81}$$

Similarly, from (79) and the estimate

$$\begin{aligned}&\sum_{p\ge1}^{\lambda>M_\varepsilon} \lambda|<u(.,t), X_p(.)>|^2 \\ &\quad\le \frac{1}{M_\varepsilon}\sum_{p\ge1}\lambda^2 |<u(.,t), X_p(.)>|^2 \\ &\quad\le \frac{1}{M_\varepsilon}\|u(.,t)\|^2_{H^2(\Omega)},\end{aligned} \tag{82}$$

we find that

$$\begin{aligned}&\|\nabla u_\varepsilon(.,t)-\nabla u(.,t)\|^2_{L^2(\Omega)} \\ &\quad\le \varepsilon^{\frac{t}{T}}\frac{\ln(\frac{1}{\varepsilon})}{2TB_2}+\frac{1}{M_\varepsilon}\|u(.,t)\|^2_{H^2(\Omega)}.\end{aligned} \tag{83}$$

Using the inequality $a+b\le\left(\sqrt{a}+\sqrt{b}\right)^2$ we obtain the error estimate in H_0^1.

Finally we prove the convergence in $H^2(\Omega)$. Similarly to (79) we have

$$\begin{aligned}&\|\Delta(u_\varepsilon-u)(.,t)\|^2_{L^2} \\ &\quad=\sum_{p\ge1}\lambda^2\left|<u_\varepsilon(.,t)-u(.,t), X_p(.)>\right|^2 \\ &\quad\le \sum_{p\ge1}^{\lambda\le M_\varepsilon} M_\varepsilon^2\left|<u_\varepsilon(.,t)-u(.,t), X_p(.)>\right|^2 \\ &\quad+\sum_{p\ge1}^{\lambda>M_\varepsilon}\lambda^2\left|<u(x,t), X_p(x)>\right|^2 \\ &\quad\le M_\varepsilon^2\|P_\varepsilon u_\varepsilon(.,t)-P_\varepsilon u(.,t)\|^2_{L^2(\Omega)} \\ &\quad+\sum_{p\ge1}^{\lambda>M_\varepsilon}\lambda^2\left|<u(.,t), X_p(.)>\right|^2 \\ &\quad\le \varepsilon^{\frac{t}{T}}\frac{\ln(\frac{1}{\varepsilon})}{2TB_2}+\sum_{p\ge1}^{\lambda>M_\varepsilon}\lambda^2\left|<u(.,t), X_p(.)>\right|^2\to0\end{aligned} \tag{84}$$

as $\in\to0$ due to the convergence in (80).

Remark 3.

In the subsection (ii) of Theorem 4.3, an error estimate in $H^2(\Omega)$ is not given because we only know $u(t)\in H^2(\Omega)\cap H_0^1(\Omega)$, and do not have enough information on the exact solution. However, when u is smoother then an explicit error estimate in $H^2(\Omega)$ may be derived. In the last theorem, we shall give the error estimates in some special cases when the exact solution is known. From the proof of Theorem 4.3 shown that in fact $u(t)\in H_0^1(\Omega)$ and $u(t)\in H^2(\Omega)\cap H_0^1(\Omega)$ are equivalent to

$$\sum_{p\ge1}\lambda^{2k}|<u(.,t), X_p(.)>|^2<\infty \tag{85}$$

with $k=1,2$, respectively. We shall see that from the condition (41) above with $k>2$ we may improve the estimate, and particularly give an error estimate in $H^2(\Omega)$. We next consider a stronger condition, although it is quite strict for the linear case, as we discussed in previous section, if the above condition (85) holds then we have a better convergence rate.

Theorem 4.4

Let u, $u_\in$, $M_\in$ be as in Theorem 4.2 and let $t\in[0,T]$.

i. *Assume that*

$$\|u(.,t)\|^2_{S^k(\Omega)}=\sum_{p=1}^{\infty}\lambda^{2k}|<u(.,t), X_p(.)>|^2<\infty, \tag{86}$$

for some constant $k>2$. Then

$$\begin{aligned}&\|u_\varepsilon(.,t)-u(.,t)\|_{L^2(\Omega)} \\ &\quad\le\varepsilon^{\frac{t}{2T}}\sqrt{\frac{\ln(\frac{1}{\varepsilon})}{2TB_2}+\left(\frac{2B_2T}{\ln(\varepsilon^{-1})}\right)^k}\ \|u(.,t)\|_{S^k(\Omega)},\end{aligned} \tag{87}$$

$$\begin{aligned}&\|u_\varepsilon(.,t)-u(.,t)\|_{H_0^1(\Omega)} \\ &\quad\le\varepsilon^{\frac{t}{2T}}\sqrt{\frac{\ln(\frac{1}{\varepsilon})}{2TB_2}+\left(\frac{2B_2T}{\ln(\varepsilon^{-1})}\right)^{\frac{2k-1}{2}}}\ \|u(.,t)\|_{S^k(\Omega)},\end{aligned} \tag{88}$$

$$\begin{aligned}&\|u_\varepsilon(.,t)-u(.,t)\|_{H^2(\Omega)} \\ &\quad\le3\varepsilon^{\frac{t}{2T}}\sqrt{\frac{\ln(\frac{1}{\in})}{2TB_2}+3\left(\frac{2B_2T}{\ln(\varepsilon^{-1})}\right)^{\frac{2k-2}{2}}}\ \|u(.,t)\|_{S^k(\Omega)}.\end{aligned} \tag{89}$$

Here we assume $\in\le e^{-2T}$ for the estimate in $H^2(\Omega)$.

ii. *Assume that*

$$F_r(t)=\sum_{p\ge1}e^{2r\lambda}|<u(.,t), X_p(.)>|^2<\infty$$

for some constant $r>0$. Then

$$\begin{aligned}&\left\|u_\varepsilon(.,t)-u(.,t)\right\|_{L^2(\Omega)}\\&\quad\le \varepsilon^{\frac{t}{2T}}\sqrt{\frac{\ln(\frac{1}{\varepsilon})}{2TB_2}}+\sqrt{F_r(t)}\varepsilon^{\frac{r}{2T}},\end{aligned}\tag{90}$$

$$\begin{aligned}&\left\|u_\varepsilon(.,t)-u(.,t)\right\|_{H_0^1(\Omega)}\\&\quad\le \varepsilon^{\frac{t}{2T}}\sqrt{\frac{\ln(\frac{1}{\in})}{2TB_2}}+\frac{\sqrt{F_r(t)\ln(\varepsilon^{-1})}}{\sqrt{2B_2T}}\varepsilon^{\frac{r}{2T}}\end{aligned}\tag{91}$$

$$\begin{aligned}&\left\|u_\varepsilon(.,t)-u(.,t)\right\|_{H^2(\Omega)}\\&\quad\le 3\varepsilon^{\frac{t}{2T}}\sqrt{\frac{\ln(\frac{1}{\varepsilon})}{2TB_2}}+3\frac{\sqrt{F_r(t)}\ln(\varepsilon^{-1})}{2TB_2}\varepsilon^{\frac{r}{2T}}.\end{aligned}\tag{92}$$

Here we assume $\in\le e^{-2T}$ *for the estimate in* $H_0^1(\Omega)$*, and* $\in\le e^{-4T}$ *for the estimate in* $H^2(\Omega)$.

Proof:

i. We use the same way as in the proof of Theorem 4.3. We shall prove the error estimates in $H^2(\Omega)$ (the other ones are similar and easier). From

$$\begin{aligned}&\sum_{p\ge1}^{\lambda>M_\varepsilon}\lambda^2\,|<u(.,t),X_p(.)>|^2\\&\quad\le\frac{1}{M_\varepsilon^{2k-2}}\sum_{p\ge1}\lambda^{2k}\,|<u(.,t),X_p(.)>|^2\\&\quad\le\frac{\|u(.,t)\|_{S^k(\Omega)}^2}{M_\varepsilon^{2k-2}}\end{aligned}\tag{93}$$

and (84) we find that

$$\left\|\Delta(u_\varepsilon-u)(.,t)\right\|_{L^2}^2\le\varepsilon^{\frac{T+t}{T}}+\frac{\|u(.,t)\|_{S^k(\Omega)}^2}{M_\varepsilon^{2k-2}}.$$

Using

$$\|w\|_{H^2}\le\|w\|_{L^2}+\|w\|_{H_0^1}+\|\Delta w\|_{L^2}\le3\|\Delta w\|_{L^2}\tag{94}$$

and $M_\in\ge1$ we conclude the desired estimate in $H^2(\Omega)$.

ii. From (71) and

$$\begin{aligned}&\sum_{p\ge1}^{\lambda>M_\varepsilon}|<u(.,t),X_p(.)>|^2\\&\quad\le e^{-2rM_\varepsilon}\sum_{p\ge1}e^{2r\lambda}\,|<u(.,t),X_p(.)>|^2\le F_r(t)\varepsilon^{\frac{r}{T}}\end{aligned}\tag{95}$$

we get the error estimate in $L^2(\Omega)$:

$$\begin{aligned}&\left\|u_\varepsilon(.,t)-u(.,t)\right\|_{L^2(\Omega)}\\&\quad\le\varepsilon^{\frac{t}{2T}}\sqrt{\frac{\ln(\frac{1}{\in})}{2TB_2}}+\sum_{p\ge1}^{\lambda>M_\varepsilon}|<u(.,t),X_p(.)>|^2\\&\quad\le\varepsilon^{\frac{t}{2T}}\sqrt{\frac{\ln(\frac{1}{\in})}{2TB_2}}+\sqrt{F_r(t)}\varepsilon^{\frac{r}{\cdot}2T}\end{aligned}\tag{96}$$

Note that the function $\xi\mapsto e^\xi/\xi$ is increasing when $\xi\ge1$. Thus

$$\lambda\le M_\varepsilon e^{2r(\lambda-M_\varepsilon)}\quad\text{when}\quad\lambda>M_\varepsilon\ge1.$$

It implies that

$$\begin{aligned}&\sum_{p\ge1}^{\lambda>M_\varepsilon}|<u(.,t),X_p(.)>|^2\\&\quad\le M_\varepsilon\sum_{p\ge1}e^{2r(\lambda-M_\varepsilon)}\,|<u(.,t),X_p(.)>|^2\\&\quad\le M_\varepsilon F_r(t)\varepsilon^{\frac{r}{\cdot}T}\end{aligned}\tag{97}$$

The error estimate in $H_0^1(\Omega)$ follows from the above estimate and (79).

Similarly, because the function $\xi\mapsto e^\xi/\xi^2$ is increasing when $\xi\ge2$, we find that

$$\lambda^2\le M_\varepsilon^2e^{2r(\lambda-M_\varepsilon)}\quad\text{if}\quad\lambda>M_\varepsilon\ge2.$$

It follows that

$$\begin{aligned}&\sum_{p\ge1}^{\lambda>M_\varepsilon}|<u(.,t),X_p(.)>|^2\\&\quad\le M_\varepsilon^2\sum_{p\ge1}e^{2r(\lambda-M_\varepsilon)}\,|<u(x,t),X_p(.)>|^2\\&\quad\le M_\varepsilon^2F_r(t)\varepsilon^{\frac{r}{T}}.\end{aligned}\tag{98}$$

Thus (84) reduces to

$$\begin{aligned}&\left\|\Delta(u_\varepsilon-u)(.,t)\right\|_{L^2}^2\\&\quad\le\varepsilon^{\frac{t}{2T}}\sqrt{\frac{\ln(\frac{1}{\varepsilon})}{2TB_2}}+M_\varepsilon^2F_r(t)\varepsilon^{\frac{r}{T}}.\end{aligned}\tag{99}$$

Hence

$$\begin{aligned}&\left\|\Delta(u_\varepsilon-u)(.,t)\right\|_{L^2}\\&\quad\le\varepsilon^{\frac{T+t}{2T}}+M_\varepsilon\sqrt{F_r(t)}\varepsilon^{\frac{r}{2T}}\\&\quad=\varepsilon^{\frac{t}{2T}}\sqrt{\frac{\ln(\frac{1}{\varepsilon})}{2TB_2}}+\frac{\sqrt{F_r(t)}\ln(\varepsilon^{-1})}{2TB_2}\varepsilon^{\frac{r}{\cdot}2T}\end{aligned}\tag{100}$$

Using

$$\|w\|_{H^2} \le \|w\|_{L^2} + \|w\|_{H_0^1} + \| \Delta w \|_{L^2} \le 3 \| \Delta w \|_{L^2} . \quad (101)$$

we have

$$\begin{aligned} &\|u_\varepsilon(.,t) - u(.,t)\|_{H^2(\Omega)} \\ &\quad \le 3\|\Delta(u_\varepsilon - u)(.,t)\|_{L^2} \qquad (102) \\ &\quad \le 3\varepsilon^{\frac{t}{2T}} \sqrt{\frac{\ln(\frac{1}{\varepsilon})}{2TB_2}} + 3\frac{\sqrt{F_r(t)}\ln(\varepsilon^{-1})}{2TB_2}\varepsilon^{\frac{r}{2T}}. \end{aligned}$$

5 CONCLUSION

In this paper, we solved a backward in time problem of the heat equation with time-dependent coefficients and inhomogeneous source. We suggested two new methods; the truncation of high frequency and the quasi-boundary-type method. In the theoretical results, we obtained the error estimation of Hölder type in L^2, H^1, H^2 norms based on some assumptions of the exact solution. In the numerical results, it shows that both methods are stable and converged to the exact solutions at $t = 0$. In comparison between two regularized methods, the MQBV shows a better numerical performance in term of error estimation and convergence rate. However, it should be stated that the regularized solutions of our methods are based on the series expression of solution, which may lead a limitation of the methods for applications in a general domain where a solution is described by any physical meaning or interesting problem.

REFERENCES

Ames, K.A. & J. Epperson (1997). A kernel-based method for the approximate solution of backward parabolic problems. *SIAM J. Numer. Anal. 34, Vol. 8*, 127–145.

Atmadja, J. & A. Bagtzoglou (2003). Marching-jury backward beam equation and quasi-reversibility methods for hydrologic inversion: Application to contaminant plume spatial distribution recovery. *WRR 39*.

Berntsson, F. (1999). A spectral method for solving the sideways heat equation. *Inverse Problem 15*, 891–906.

Burmistrova, V. (2005). Regularization method for parabolic equation with variable operator. *J. Appl. Math. no. 4*, 382–392.

Clark G.W., & S.F. Oppenheimer (1994). Quasireversibility methods for non-well posed problems. *Elect. J. Diff. Eqns. 301*, 1–9.

Denche, M. & K. Bessila (2005). *A modified quasiboundary value method for ill-posed problems,* Volume 301. J. Math. Anal. Appl,.

Elden, L., F. Berntsson, T.R. (2000). Wavelet and fourier methods for solving the sideways heat equation. *SIAM J. Sci. Comput. 21(6)*, 2187–2205.

Evans, L.C. (1997). *Partial differential equation*. American Mathematical Society, Providence, Rhode Island 19.

Ewing, R. (1975). The approximation of certain parabolic equations backward in time by sobolev equation. *SIAM J. Math. Anal. 6,* 283–294.

Feng, X.L., L. Elden, C. (2010). Stability and regularization of a backward parabolic pde with variable coefficient. J. *Inverse and Ill-posed Problems 18,* 217–243.

Feng, X.L., L. Elden, C.F. (2008). Numerical approximation of solution of nonhomogeneous backward heat conduction problem in bounded region. *J. Math. Comp. Simulation 79, no. 2*, 177–188.

Feng, Xiao-Li; Elden, L.F. C.L. (2010). A quasiboundary-value method for the cauchy problem for elliptic equations with nonhomogeneous neumann data. *J. Inverse Ill-Posed Probl. 18*, 617–645.

Fu, C.L., X.X. Tuan, Z.Q. (2007). Fourier regularization for a backward heat equation. *J. Math. Anal. Apll. 331*, 472–480.

Isakov, V. (1998). Inverse problems for partial differential equation. *Springer-Verlag, New York*.

John, F. (1960). Continuous dependence on data for solutions of partial differential equations with a prescribed bound. *Comm. Pure Appl. Math (13)*, 551–585.

Lattès, R. & J.L. Lion (1967). Methode de quasireversibility et application. *Dunod, Paris.*

Lee, J., D.S. (2006). A parallel method for backward parabolic problem based on the laplace transformation. *SIAM J.Nummer. Anal. 44*.

Lee, J., D.S. (2009). F. John's stability conditions versus A. Carasso's SECB constraint for backward parabolic problems. *Inverse Problem*.

Melnikova, I.V., A.F. (1993a). The cauchy problem. Three approaches, monograph and surveys in pure and applied mathematics. *London - New York: Chapman and Hall 120*.

Melnikova, I.V., S.B. (1993b). I.v. melnikova, s.v. bochkareva. *Dok.Akad.Nauk. 329*, 270–273.

Miller, K. (1970). Least squares methods for ill-posed problems with a prescribed bound. *SIAM J. Math. Anal.*, 52–74.

Nam, P.T., D.D. Trong, N.T. (2010). The truncation method for a two-dimensional nonhomogeneous backward heat problem. *Appl. Math. Comput. 216*, 3423–3432.

Payne, L. (1973). *Some general remarks on improperly posed problems for partial differential equations. 1–30:* Symposium on Non-well Posed Problems and Logarithmic Convexity, Lecture Notes in Mathematics,.

Quan, Pham Hoang; Trong, D.D.T.L. M.T.N.H. (2011). A modified quasi-boundary value method for regularizing of a backward problem with time-dependent coefficient. *Inverse Probl. Sci. Eng. 19*, 409–423.

Schroter, T., U.T. (1996). On optimal regularization methods for the backward heat equation. *Z. Anal. Anw. 15*, 475–493.

Shidfar, A., A.Z. (2005). A numerical technique for backward inverse heat conduction problems in one – dimensional space. *Appl. Math. Comput. 171*, 1016–1024.

Showalter, R. (1974). The final value problem for evolution equations. *J. Math. Anal. Appl. 47*, 563–572.

Showalter, R. (1983). Cauchy problem for hyper -parabolic partial differential equations. *in Trends in the Theory and Practice of Non-Linear Analysis*.

Trong, D. & N. Tuan (2006). Regularization and error estimates for nonhomogeneous backward heat problem. pp. 1–10.

Trong, D. & N. Tuan (2008). A nonhomogeneous backward heat problem: Regularization and error estimates. *Electron. J. Diff. Eqns. 33*, 1–14.

Tuan, N. & D. Trong (2010). A nonlinear parabolic equation backward in time: regularization with new error estimates. *Nonlinear Anal. 73*, 1842–1852.

Tuan, N.H., P.H. Quan, D.T. & L. Triet (2013). On a backward heat problem with time-dependent coefficient: Regularization and error estimates. *Appl. Math. Comp. 219*, 6066–6073.

Yildiz, B., H. Yetis, A. (2003). A stability estimate on the regularized solution of the backward heat problem. *Appl. Math. Comp. 135*, 561–567.

Yildiz, B., M.O. (2000). Stability of the solution of backward heat equation on a weak conpactum. *Appl. Math. Comput. 111*, 1–6.

Advanced numerical methods

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

A study of Boundary Element Method for 3D homogeneous Helmholtz equation with Dirichlet boundary condition

M.-P. Tran & V.-K. Huynh
Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

ABSTRACT: In the paper, we study the numerical solution to 3D homogeneous Helmholtz equation with Dirichlet boundary condition. The discretization of the problem is considered using the Boundary Element Method (BEM). The problem in the whole domain is first established in terms of the interior and exterior boundary integral equation. Based on the Green formula, analytical solution to Helmholtz equation is represented in terms of the boundary data. At last, in this study, we also apply the Finite Difference Method (FDM) in order to give comparative results. Numerical performances are then proposed to indicate the validity of our method.

1 INTRODUCTION

The Helmholtz equation, carries the name of the physicist Hermann Ludwig Ferdinand von Helmholtz, was contributed to the mathematical acoustics and electromagnetics. It is an important equation to be solved in numerical electromagnetic problems, for instance, the waveguide problems in physical phenomena, acoustic rediation, heat conduction, wave propagation and electrolocation/echolocation problems. The research of Helmholtz equation in 2D and 3D are studied in many literatures, such as (Colton 1998), (E.A. Spence & Fokas 2009), (Olaf & Sergej 2007), (Burton & Miller 1971), (Ihlenburg & Babuska 1997), (Liu, Nakamura, & Potthast 2007), (Goldstein 1982), (Jan. 2009) and a lot of references therein. A variety of problems require solutions to the Helmholtz equation in both interior and exterior domains. In the case of no source, the three-dimensional (3D) time-independent linear Helmholtz equation is considered as:

$$\Delta u + \kappa^2 u = 0, \quad x \in \mathbb{R}^3, \tag{1.1}$$

where $\kappa > 0$ is the *wave number* given by:

$$\kappa = \frac{\omega}{c} = \frac{2\pi f}{c} = \frac{2\pi}{\lambda}. \tag{1.2}$$

We introduce here $\lambda = \frac{c}{f}$ the wavelength of plane waves of frequency f. The governing Dirichlet boundary condition is considered on the bounded Lipschitz domain $\Omega \subset \mathbb{R}^3$ as:

$$u = g, \quad \text{on} \quad \partial\Omega, \tag{1.3}$$

where g is a continuous function defined on $\mathbb{R}^3$.

The full 3D Helmholtz problem (1.1) is considered to divide into two subproblems, called *interior* and *exterior* problems accordingly.

The interior Helmholtz problem with Dirichlet boundary condition is presented as:

$$\begin{cases} \Delta u + \kappa^2 u & = 0, \quad \text{in } \Omega, \\ u & = g, \quad \text{on } \Gamma, \end{cases} \tag{1.4}$$

in which, let us note that κ^2 cannot be a Dirichlet eigenvalue for $-\Delta$ on Ω, for frequencies f in (1.2) small enough and therefore the continuous problem has a unique solution. Without this assumption, the Helmholtz operator is singular and there is either no solution or as infinite set of solution to (1.4). Thus, this is important to do sufficiently accurate discrete approximations.

The exterior Helmholtz problem, also called scattering problem, the domain of the solution to this problem is unbounded domain $\Omega^{ext} = \mathbb{R}^3 \setminus \overline{\Omega}$, with Dirichlet boundary condition and the additional Sommerfeld radiation condition holds at infinity as follows:

$$\begin{cases} \Delta u + \kappa^2 u & = 0, \quad \text{in } \Omega^{ext}, \\ u & = g, \quad \text{on } \Gamma, \\ \left| \dfrac{\partial}{\partial n_x} u(x) - i\kappa u(x) \right| & = \mathcal{O}\left(\dfrac{1}{\|x\|^2} \right), \|x\| \to \infty, \end{cases} \tag{1.5}$$

where $\Gamma = \partial\Omega$, and n_x represents the exterior normal vector. In the scattering problems, typically the function g is smooth restricted to the boundary. In addition, the extra condition Sommerfeld must be added to the condition that ensures the uniqueness of solution to the problem (1.5).

Numerical solution to the Helmholtz equation play a vital role in applications, likely mechanical, acoustics, electromagnetics etc. Numerical methods for solving the Helmholtz problem have been an active research in recent years, where the Finite Element Method (FEM) and FDM have been applied successfully. Reliable numerical methods lead to the 3D homogeneous Helmholtz equation (1.1) are discussed in (Jan. 2009) and several concerned references therein. In this paper, we also study the numerical solution to Helmholtz equation, and give some discrete schemes to the 3D problems. For instance, one proposes the boundary element method to the boundary integral equations and gives comparison with the classical finite difference method.

The rest of this paper is organized as follows. In Section 2, we study the solution representation formula of interior and exterior problems. In the next section, we develop the variational framework to propose boundary integral value problems (BIPs), with respected to single layer and double layer potentials. In Section 4, we employ some numerical methods applied to previous BIPs. For instance, both BEM and the classical FDM discretization are proposed therein. Some numerical experiments are then presented in Section 5. In the last section, some relevant conclusions are drawn, we also discuss on some remarks, open questions and future work.

2 SOLUTION REPRESENTATION FORMULA

In this section, we give a brief review of the solution representation formula to both interior and exterior Helmholtz equations (1.4), (1.5), follows the results given in (Jan. 2009), respectively.

Let us consider the fundamental solution to the Helmholtz equation. The study of fundamental solution is used to formulate solution to our problem.

Let the function $G_\kappa : \mathbb{R}^3 \times \mathbb{R}^3 \to C$ be the fundamental solution to the Helmholtz equation in $\mathbb{R}^3$, is defined by:

$$(\Delta + \kappa^2)G_\kappa = \delta_0, \tag{2.1}$$

where δ_0 represents the Dirac delta distribution. One also has:

$$G_\kappa(x,y) = \frac{1}{4\pi}\frac{e^{i\kappa\|x-y\|}}{\|x-y\|}, \quad x,y \in \mathbb{R}^3, \quad x \neq y. \tag{2.2}$$

Two following theorems allow the construction of the boundary element method, whose formula design the solution calculation to the boundary value problems in the next section.

2.1 *Representation formula to the interior problem*

Theorem 2.1. *(Representation Formula for Bounded Domains), (Jan. 2009) Let $\Omega \subset \mathbb{R}^3$ be a bounded in C^1 domain and the boundary $\Gamma = \partial\Omega$ is assumed to be smooth enough so that the integration by part formula (in multi-dimensional space) holds. Let G_κ denote the fundamental solution for the Helmholtz equation in $\mathbb{R}^3$ and let n denote the outward normal vector to Γ. Then for $u \in C^2(\overline{\Omega})$ we have the representation formula:*

$$\begin{aligned} u(x) = &\int_\Gamma \frac{\partial u}{\partial n}(y)G_\kappa(x,y)ds(y) \\ &- \int_\Gamma u(y)\frac{\partial G_\kappa(x,y)}{\partial n_y}ds(y)\text{for } x \in \Omega. \end{aligned} \tag{2.3}$$

2.2 *Representation formula to the exterior problem*

Theorem 2.2. *(Representation Formula for Unbounded Domains), (Jan. 2009) Let $\Omega \subset \mathbb{R}^3$ be a bounded in C^1 domain and the boundary $\Gamma = \partial\Omega$ is assumed to be smooth enough so that the integration by part formula (in multi dimensional space) holds. Let G_κ denote the fundamental solution for the Helmholtz equation in $\mathbb{R}^3$ and let n denote the outward normal vector to Γ. Let us define $\Omega^{ext} = \mathbb{R}^3 \setminus \overline{\Omega}$. Then for $u \in C^2(\overline{\Omega}^{ext})$ satisfying*

$$\Delta u + \kappa^2 u = 0, \quad \text{in } \Omega^{ext},$$

and the Sommerfeld radiation condition

$$\left|\frac{\partial}{\partial n_x}u(x) - i\kappa u(x)\right| = \mathcal{O}\left(\frac{1}{\|x\|^2}\right), \quad \|x\| \to \infty.$$

Then, we have the representation formula:

$$\begin{aligned} u(x) = &\int_\Gamma u(y)\frac{\partial G_\kappa(x,y)}{\partial n_y}ds(y) \\ &- \int_\Gamma \frac{\partial u}{\partial n}(y)G_\kappa(x,y)ds(y), \quad \textit{for } x \in \Omega^{ext}. \end{aligned} \tag{2.4}$$

3 BOUNDARY INTEGRAL EQUATIONS

This section aims at providing the boundary integral equations derived from representation formula in previous section. It allows us to find the solution to (1.1) in the form of single layer potential and double layer potential. This leads to the indirect boundary element method, in resolving boundary value problems.

Let us introduce the integral operators:

$$\tilde{S}_\kappa : H^{-1/2}(\Gamma) \to H^1_{loc}(\Omega),$$

such that for $x \in \mathbb{R}^3 \setminus \Gamma$,

$$(\tilde{S}_\kappa \phi)(x) = \int_\Gamma G_\kappa(x,y)\phi(y)ds(y), \tag{3.1}$$

and

$$D_\kappa : H^{1/2}(\Gamma) \to H^1_{loc}(\Omega),$$

such that for $x \in \mathbb{R}^3 \setminus \Gamma$,

$$(D_\kappa \phi)(x) = \int_\Gamma \frac{G_\kappa(x,y)}{\partial n_y}\phi(y)ds(y), \tag{3.2}$$

where $\phi : \Gamma \to \mathbb{R}$ are density functions. Then, the K_κ denotes the double layer potential operator

$$(K_\kappa \phi)(x) = \int_\Gamma \frac{\partial G_\kappa(x,y)}{\partial n_y}\phi(y)ds(y), \text{for } x \in \Gamma, \tag{3.3}$$

The adjoint double layer potential operator K^*_κ is defined by:

$$(K^*_\kappa \phi)(x) = \int_\Gamma \frac{\partial G_\kappa(x,y)}{\partial n_x}\phi(y)ds(y), \text{for } x \in \Gamma. \tag{3.4}$$

As a result, let us recall the definitions of Dirichlet and Neumann trace operators γ_0, γ_1, were proposed in (Jan. 2009). The Dirichlet trace operator γ_0 is:

$$\gamma_0 : H^1_{\text{loc}}(\Omega) \to H^{1/2}(\Gamma).$$

Combine the operator γ_0 with $\tilde{S}_\kappa$ one has the single layer potential:

$$S_\kappa : H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma), \quad S_\kappa = \gamma_0 \tilde{S}_\kappa. \tag{3.5}$$

And the Neumann trace operator is given as:

$$\gamma_1 : H^1_{\text{loc}}(\Omega) \to H^{-1/2}(\Gamma).$$

Combine the operator γ_1 with the single layer potential one has the linear continuous mapping:

$$\gamma_1 \tilde{S}_\kappa : H^{-1/2}(\Gamma) \to H^{-1/2}(\Gamma).$$

Theorem 3.1. *Let $\phi \in C(\Gamma)$ the space of continuous functions. Then the representation formula in* (2.3), (2.4) *can be rewritten as:*

$$u(x) = \tilde{S}_\kappa \gamma_1^{int} u(x) - D_\kappa \gamma_0^{int} u(x), \quad x \in \Omega, \tag{3.6}$$

$$u(x) = -\tilde{S}_\kappa \gamma_1^{ext} u(x) + D_\kappa \gamma_0^{ext} u(x), \quad x \in \Omega^{ext}, \tag{3.7}$$

where $\gamma_0^{int}, \gamma_0^{ext}$ are interior and exterior Dirichlet trace operators, defined as:

$$\gamma_0^{int} : H^1_{loc}(\Omega) \to H^{1/2}(\Gamma),$$
$$\gamma_0^{ext} : H^1_{loc}(\Omega) \to H^{1/2}(\Gamma),$$

such that

$$\begin{aligned} &\gamma_0^{int} v = v|_\Gamma \ \textit{for } v \in C^\infty(\overline{\Omega}), \\ &\gamma_0^{ext} v = v|_\Gamma \ \textit{for } v \in C^\infty(\overline{\Omega}^{ext}), \end{aligned} \tag{3.8}$$

and the $\gamma_1^{int}, \gamma_1^{ext}$ are interior and exterior Neumann trace operators are defined respectively as:

$$\gamma_1^{int} : H^1_{loc}(\Omega) \to H^{-1/2}(\Gamma),$$
$$\gamma_1^{ext} : H^1_{loc}(\Omega) \to H^{-1/2}(\Gamma),$$

such that

$$\begin{aligned} &\gamma_1^{int} v = \left.\frac{\partial v}{\partial n}\right|_\Gamma \ \textit{for } v \in C^\infty(\overline{\Omega}), \\ &\gamma_1^{ext} v = \left.\frac{\partial v}{\partial n}\right|_\Gamma \ \textit{for } v \in C^\infty(\overline{\Omega}^{ext}). \end{aligned} \tag{3.9}$$

Theorem 3.2. *Let γ_0 and γ_1 be the Dirichlet and Neumann trace operators. Then we have* $\forall \phi \in H^{-1/2}(\Gamma), \varphi \in H^{1/2}(\Gamma)$:

$$\begin{aligned} \gamma_0 \tilde{S}_\kappa \phi &= \gamma_0^{ext} \tilde{S}_\kappa \phi - \gamma_0^{int} \tilde{S}_\kappa \phi = 0, \\ \gamma_1 \tilde{S}_\kappa \phi &= \gamma_1^{ext} \tilde{S}_\kappa \phi - \gamma_1^{int} \tilde{S}_\kappa \phi = -\phi, \\ \gamma_0 D_\kappa \varphi &= \gamma_0^{ext} D_\kappa \varphi - \gamma_0^{int} D_\kappa \varphi = \varphi, \\ \gamma_1 D_\kappa \varphi &= \gamma_1^{ext} D_\kappa \varphi - \gamma_1^{int} D_\kappa \varphi = 0, \end{aligned}$$

Theorem 3.3. *(Jan. 2009) For $\phi \in H^{-1/2}(\Gamma)$ there holds*

$$\gamma_1^{int}(\tilde{S}_k \phi)(x) = \frac{1}{2}\phi(x) + (K^*_\kappa \phi)(x), \quad x \in \Gamma \tag{3.10}$$

$$\gamma_1^{ext}(\tilde{S}_k \phi)(x) = -\frac{1}{2}\phi(x) + (K^*_\kappa \phi)(x), \quad x \in \Gamma \tag{3.11}$$

In addition, let us also introduce here the hypersingular integral operator, which is defined as the negative Neumann trace of the double layer potential K_κ, denoted as E_κ.

$$E_\kappa : H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma), \tag{3.12}$$

such that, for a smooth density function ϕ one has

$$(E_\kappa \phi)(x) = -\gamma_1(D_\kappa \phi)(x), \tag{3.13}$$

where the Neumann trace of the double layer potential term $\gamma_1(D_\kappa \phi)$ is given on the boundary as:

$$\gamma_1(D_\kappa \phi) = \gamma_1^{\text{ext}}(D_\kappa \phi) - \gamma_1^{\text{int}}(D_\kappa \phi), \forall \phi \in H^{1/2}(\Gamma). \tag{3.14}$$

3.1 *Interior boundary value problem*

Theorem 3.4. *If u is a solution to the interior Dirichlet Helmholtz problem:*

$$\begin{cases} \Delta u + \kappa^2 u & = 0, \quad \textit{in } \Omega, \\ \gamma_0^{int} u & = g, \quad \textit{on } \Gamma, \end{cases} \tag{3.15}$$

with a bounded Lipschitz domain Ω *and Dirichlet boundary condition* $g \in H^{1/2}(\Gamma)$*, then the Neumann trace* $\gamma_1^{int} u$ *satisfies the boundary integral equation*

$$\left(S_\kappa \gamma_1^{int}\right)(u(x)) = \frac{1}{2} g(x) + (K_\kappa g)(x), x \in \Gamma. \tag{3.16}$$

and u has the representation formula (3.6).

Conversely, if $\gamma_1^{int} u$ *satisfies the boundary integral equation* (3.16)*, then the representation formula* (3.6) *defines a solution u to the interior Dirichlet problem* (3.15).

Proof. The solution to the equation (3.15) is given by the representation formula in (3.6):

$$u(x) = \tilde{S}_\kappa \gamma_1^{int} u(x) - D_\kappa \gamma_0^{int} u(x), \quad x \in \Omega,$$

with unknown Neumann trace data $\gamma_1^{int} u \in H^{-1/2}(\Gamma)$.

Apply the interior Dirichlet trace operator γ_0^{int} both sides of (3.6) for $x \in \Gamma$ and follow the Theorem 3.3, one gets the boundary integral equation for $x \in \Gamma$:

$$(\gamma_0^{int} u)(x) = \left(\frac{1}{2} I - K_\kappa\right)(\gamma_0^{int} u)(x) + (S_\kappa \gamma_1^{int} u)(x), \tag{3.17}$$

and since $u = g, \forall x \in \Gamma$ we get $\gamma_0^{int} u = g$, which implies the Fredholm boundary integral equation of the first kind:

$$\left(S_\kappa \gamma_1^{int}\right)(u(x)) = \frac{1}{2} g(x) + (K_\kappa g)(x), x \in \Gamma.$$

□

Respectively, apply the interior Neumann trace operator γ_1^{int} both sides of (3.6) gives:

$$(\gamma_1^{int} u)(x) = \left(\frac{1}{2} I + K_\kappa^*\right)(\gamma_1^{int} u)(x) + (E_\kappa \gamma_0^{int} u)(x), \tag{3.18}$$

obtains the Fredholm boundary integral equation of the second kind:

$$\frac{1}{2}(\gamma_1^{int} u)(x) - (K_\kappa^* \gamma_1^{int} u)(x) = (E_\kappa g)(x), x \in \Gamma. \tag{3.19}$$

According to these above equations it gives variational problems as

$$\left\langle S_\kappa \gamma_1^{int} u, \phi \right\rangle_\Gamma = \left\langle \left(\frac{1}{2} I + K_\kappa\right) g, \phi \right\rangle_\Gamma, \forall \phi \in H^{-1/2}(\Gamma), \tag{3.20}$$

and

$$\left\langle \left(\frac{1}{2} I - K_\kappa^*\right) \gamma_1^{int} u, \theta \right\rangle_\Gamma = \left\langle E_\kappa g, \theta \right\rangle_\Gamma, \forall \theta \in H^{1/2}(\Gamma). \tag{3.21}$$

3.2 *Exterior boundary value problem*

Theorem 3.5. *If u is a solution to the exterior Dirichlet problem:*

$$\begin{cases} \Delta u + \kappa^2 u & = 0, \quad \textit{in } \Omega^{ext}, \\ \gamma_0^{ext} u & = g, \quad \textit{on } \Gamma, \\ \left| \dfrac{\partial}{\partial n_x} u(x) - i\kappa u(x) \right| & = \mathcal{O}\left(\dfrac{1}{\|x\|^2}\right), \quad \|x\| \to \infty, \end{cases} \tag{3.22}$$

with a bounded Lipschitz domain Ω *and Dirichlet boundary condition* $g \in H^{1/2}(\Gamma)$*, then the Neumann trace* $\gamma_1^{int} u$ *satisfies the boundary integral equation:*

$$(\mathrm{S}_\kappa \gamma_1^{ext} u)(x) = -\frac{1}{2} g(x) + K_\kappa g(x), \quad \forall x \in \Gamma. \tag{3.23}$$

and u has the representation formula (3.7).

Conversely, if $\gamma_1^{int} u$ *satisfies the boundary integral equation* (3.23)*, then the representation formula* (3.7) *defines a solution u to the interior Dirichlet problem* (3.22).

Proof. The solution can be represented as in (3.7) as:

$$u(x) = -\tilde{S}_\kappa \gamma_1^{ext} u(x) + D_\kappa \gamma_0^{ext} u(x), \quad x \in \Omega^{ext},$$

with unknown Neumann trace data $\gamma_1^{ext} u$.

Apply the Dirichlet interior and exterior trace operators γ_0^{ext} and γ_1^{ext} both sides of (3.7) and also follow the theorem 3.3, moreover it uses the fact that $\gamma_0^{ext} u = g$ on $\Gamma = \partial\Omega$, we have the following Fredholm boundary integral equation of the first kind as:

$$(S_\kappa \gamma_1^{ext} u)(x) = -\frac{1}{2} g(x) + K_\kappa g(x), \quad \forall x \in \Gamma. \quad (3.24)$$

□

Similarly, take the exterior Neumann trace operator γ_1^{ext} both sides of (3.7) it obtains the second kind Fredholm boundary integral equation:

$$\frac{1}{2}(\gamma_1^{ext} u)(x) + (K_\kappa^* \gamma_1^{ext} u)(x) = -(E_\kappa g)(x), \forall x \in \Gamma. \quad (3.25)$$

Furthermore, boundary integral equations (3.23) and (3.25) are equivalent to variational problems:

$$\left\langle S_\kappa \gamma_1^{ext} u, \phi \right\rangle_\Gamma = \left\langle \left(-\frac{1}{2} I + K_\kappa \right) g, \phi \right\rangle_\Gamma, \forall \phi \in H^{-1/2}(\Gamma), \quad (3.26)$$

and

$$\left\langle \left(\frac{1}{2} I + K_\kappa^* \right) u, \theta \right\rangle_\Gamma = \langle -E_\kappa g, \theta \rangle_\Gamma, \forall \theta \in H^{1/2}(\Gamma). \quad (3.27)$$

4 NUMERICAL METHODS FOR HELMHOLTZ INTERIOR AND EXTERIOR EQUATIONS

4.1 *Boundary element method*

In this section, we describe the discretization of boundary integral equations from the previous section. In addition, we present a novel boundary element method for parametrization of triangular mesh in 3D. The solutions to boundary value problems (3.16) and (3.23) are then approximated by numerical techniques as below.

Let us assume here that Ω is a polyhedral bounded Lipschitz domain in $\mathbb{R}^3$. First, on decomposes the boundary $\Gamma \subset \mathbb{R}^3$ into a finite set of boundaries as:

$$\Gamma = \partial\Omega = \bigcup_{i=1}^{M} \Gamma_i, \quad (4.1)$$

where Γ_i represents the boundary element for $i = 1,2,...,M$ and M is large enough denotes the number of elements. Each element Γ_i in (4.1) represents the discretized triangular elements, has three vertices $X_1^k(x_1^k, y_1^k, z_1^k)$, $X_2^k(x_2^k, y_2^k, z_2^k)$, and $X_3^k(x_3^k, y_3^k, z_3^k)$. Let us define a triangular domain $\hat{\Gamma} \subset \mathbb{R}^2$ as following:

$$\hat{\Gamma} = \{\xi = (\xi_1, \xi_2) \in \mathbb{R}^2 : 0 < \xi_1 < 1; 0 < \xi_2 < 1 - \xi_1\}. \quad (4.2)$$

4.1.1 *Piecewise constant basis function*

For every elements Γ_k we define the piecewise constant function ψ_k as follows

$$\psi_k(X) = \begin{cases} 1 & \text{for } X \in \Gamma_k \\ 0 & \text{elsewhere} \end{cases}, \quad k = 1,2,...,M.$$

The function $\psi_k \in \Gamma_k$ can be identified as a function $\hat{\psi} \in \hat{\Gamma}$ as:

$$\hat{\psi}(\xi) = \begin{cases} 1 & \text{for } \xi \in \hat{\Gamma} \\ 0 & \text{elsewhere} \end{cases}.$$

Let $T_\psi(\Gamma)$ be the approximation of the Newmann data, i.e., of the normal derivatives on Γ. The linear space $T_\psi(\Gamma)$ is defined as

$$T_\psi(\Gamma) = \text{span}\{\psi_k\}_{k=1}^{M},$$

and for every complex-valued function $g_\psi \in T_\psi(\Gamma)$ can be represented

$$g_\psi = \sum_{k=1}^{M} g_k \psi_k,$$

Let N denote the number of nodes of a given triangular mesh. We define the family of functions $\{\varphi_l\}_{l=1}^{N}$ continuous over the whole discretized boundary as following

$$\varphi_l = \begin{cases} 1 & \text{for } x = x_l \\ 0 & \text{for } x \neq x_l \\ \text{piecewise affine} & \text{otherwise} \end{cases}.$$

Let us define the linear space $T_\varphi(\Gamma)$ as

$$T_\varphi(\Gamma) = \text{span}\{\varphi_l\}_{l=1}^{N}.$$

A function φ_k to an element $\Gamma_l \subset \text{supp}\,\varphi_k$ can be identified with one of the functions $\hat{\varphi}_1, \hat{\varphi}_2, \hat{\varphi}_3$ defined on the reference triangle as:

$$\hat{\varphi}_1 = 1 - \xi_1 - \xi_2; \quad \hat{\varphi}_2 = \xi_1, \quad \hat{\varphi}_3 = \xi_2, \quad \text{for } \xi \in \hat{\Gamma}.$$

For every function g_φ in $T_\varphi(\Gamma)$, it can be represented as

$$g_\varphi = \sum_{l=1}^{N} g_l \varphi_l.$$

The linear space $T_\varphi(\Gamma)$ is applied for the approximation of the Dirichlet data, i.e., the values of the solution on Γ .

4.1.2 *Discrete solution to interior problem*

Let us denote u_h is the approximated solution to the interior problem (1.4), and the discrete unknown Neumann data as:

$$W := \gamma_1^{int} u_h. \tag{4.3}$$

From (3.20), for all $\phi_h \in H_h$, we have

$$\langle S_\kappa W, \phi_h \rangle_\Gamma = \left\langle \left(\frac{1}{2} I + K_\kappa \right) g_h, \phi_h \right\rangle_\Gamma . \tag{4.4}$$

We find the approximate forms in term of functional bases as:

$$W \approx \sum_{k=1}^{M} W_k \psi_k \in T_\psi(\Gamma); \quad g_h \approx \sum_{l=1}^{N} g_l \varphi_l \in T_\varphi(\Gamma).$$

For every $i = 1,2,...,M$, from (4.4) we obtain the linear system to find the approximate solution W to:

$$V_\kappa W = \left(\frac{1}{2} R_\kappa + P_\kappa \right) g_h, \tag{4.5}$$

or simplicity in term:

$$CW = D, \tag{4.6}$$

where $\forall i = 1,...,M; \quad \forall j = 1,...,N$:

$$\begin{aligned} C_{ij} &= V_\kappa(i,j) = \int_{\Gamma_i} \int_{\Gamma_j} G_\kappa(x,y) ds(x) ds(y), \\ D_j &= \left(\frac{1}{2} R_\kappa + P_\kappa \right)_{jm} (g_h)_m, \end{aligned} \tag{4.7}$$

and

$$\begin{aligned} R_\kappa(j,m) &= \int_{\Gamma_j} \varphi_m(x) ds(x), \\ P_\kappa(j,m) &= \int_{\Gamma_j} \int_\Gamma \varphi_m(y) \frac{\partial G_\kappa(x,y)}{\partial n(y)} ds(y) ds(x). \end{aligned}$$

One gets the approximate solution u_h to the interior Dirichlet boundary value problem (1.4) is given by the discrete representation formula for $x \in \Omega$:

$$\begin{aligned} u(x) \approx u_h(x) = &\sum_{k=1}^{M} W_k \int_{\Gamma_k} G_\kappa(x,y) ds(y) \\ &- \sum_{l=1}^{N} (g_h)_l \int_\Gamma \varphi_l(y) \frac{\partial G_\kappa(x,y)}{\partial n(y)} ds(y). \end{aligned} \tag{4.8}$$

4.1.3 *Discrete solution to exterior problem*

Similarly, we denote u_h is the approximated solution to the exterior problem (1.5), and the unknown exterior Neumann data $\overline{W} = \gamma_1^{ext} u_h$. From (3.26), for all $\phi_h \in H_h$, we have

$$\langle S_\kappa \overline{W}, \phi_h \rangle_\Gamma = \left\langle \left(-\frac{1}{2} I + K_\kappa \right) g_h, \phi_h \right\rangle_\Gamma . \tag{4.9}$$

We find the approximate forms in term of functional bases as:

$$\overline{W} \approx \sum_{k=1}^{M} \overline{W}_k \psi_k \in T_\psi(\Gamma); \quad g_h \approx \sum_{l=1}^{N} g_l \varphi_l \in T_\varphi(\Gamma).$$

For every $i = 1,2,...,M$, from (4.9) it gives discrete variational problem:

$$\sum_{k=1}^{M} \overline{W}_k \langle S_\kappa \psi_k, \psi_i \rangle_\Gamma = \sum_{l=1}^{N} g_l \left\langle \left(-\frac{1}{2} I + K_\kappa \right) \varphi_l, \psi_i \right\rangle_\Gamma . \tag{4.10}$$

Finally, we obtain the linear system to find the approximation of Neumann data $\overline{W}$:

$$V_\kappa \overline{W} = \left(-\frac{1}{2} R_\kappa + P_\kappa \right) g_h, \tag{4.11}$$

where V_κ, R_κ, P_κ and g_h are discretized as in the previous section.

One gets the approximate solution u_h to the exterior Dirichlet boundary value problem (1.5) is given by the discrete representation formula for $x \in \Omega^{ext}$:

$$\begin{aligned} u(x) \approx u_h(x) = &\sum_{k=1}^{M} \overline{W}_k \int_{\Gamma_k} G_\kappa(x,y) ds(y) \\ &- \sum_{l=1}^{N} (g_h)_l \int_\Gamma \varphi_l(y) \frac{\partial G_\kappa(x,y)}{\partial n(y)} ds(y). \end{aligned} \tag{4.12}$$

4.2 *Algorithm*

We finally present the algorithm for BEM that follows description in Section 4. The algorithm is first proposed to the interior problem (1.4). Then, the scheme can be easily applied to the exterior problem (1.5) in 3D.

The BEM algorithm is described by following steps:

1. Set the initial data: g, κ (not too large), Γ, number of nodes and elements;
2. Initialize basis functions ψ_k, φ_l for the Neumann and Dirichlet data, respectively;
3. Calculate matrices C and D following (4.7);
4. Solve the sparse linear system (4.6);
5. Calculate numerical integrals on boundaries in (4.8) and numerical solution is then obtained.

It is also remarkable that in the discretization integral calculations, one can use some recent numerical quadrature rules, specially in case of evaluating singular integral along surfaces in three dimensions. One of effective methods is Quadrature By Expansion method (QBX), it was proposed in (O'Neil, Klockner, Barnett, & Greengard 2013).

4.3 *Finite Difference Method (FDM)*

In this section, we describe the simplest approximation of interior Helmholtz equation (1.4) by using finite difference method. In this contribution, we look for the solution in the 3D rectangular domain Ω, where the preliminary values $u(x,y,z)$ are known in the points of domain Ω.

Let us discretize the computational domain Ω with a 3D uniform grid size with the spatial step $h_1 = \Delta x, h_2 = \Delta y, h_3 = \Delta z$, respectively. Using centered finite difference approximations for the partial derivatives and Laplace's operator in (1.4), the following second order accurate system of simultaneous equations is obtained for all interior nodes:

$$\begin{aligned} u(x,y,z) = \frac{1}{6(1+\kappa^2)} & [u(x+h_1,y,z)+u(x-h_1,y,z) \\ & +u(x,y+h_2,z)+u(x,y-h_2,z) \\ & +u(x,y,z+h_3)+u(x,y,z-h_3)] \\ & +\mathcal{O}(h_1^2,h_2^2,h_3^2). \end{aligned} \tag{4.13}$$

Special equations are typically required for the boundary nodes depending of the boundary condition (1.3), it refers to (Hegedus 2009). The result can be written as a large sparse equation linear system:

$$Au = b, \quad A \in \mathbb{C}^{N\times N\times N}, u,b \in \mathbb{C}^N, \tag{4.14}$$

where A is a complex matrix as the boundary condition contains complex values, $N = N_x N_y N_z$ the total number of matrix u in Ω, N_x (N_y, N_z) are number of discretized points along x (y, z) direction.

Following the discretized FDM scheme, the approximate solution to the homogeneous model is then obtained.

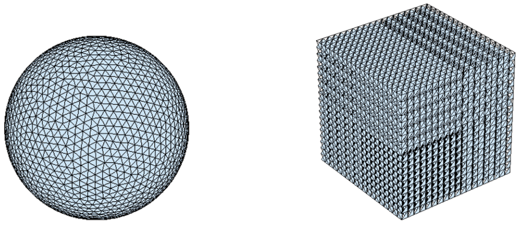
The convergence theory and the order of convergence to BEM and FDM are studied in (Sauter & Schwab 2010), (Erlangga 2005).

5 NUMERICAL EXPERIMENTS

In this section, some numerical performances are presented to confirm the efficiency of the BEM to the interior Boundary Value Problem (BVP) in a comparison with the classical FDM in Section 4.3. We consider the interior homogeneous Dirichlet boundary Helmholtz problem on the domain Ω, such as a unit sphere (center at the origin, radius 1) and a cube as in Figure 1. In this contribution, the numerical algorithm has been developed that effectively solves (1.4).

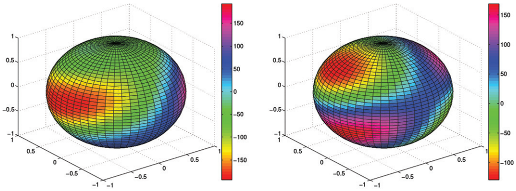
These following numerical simulations are performed entirely within the Matlab environment. It has been noticed that in the application of the piecewise constant basis functions ϕ, ψ as in Section 4.1.1 and even coarser triangular meshes, the large amount of RAM and time on computer facilities are needed.

Discrete solution is implemented to solve interior problem (1.4) with $\kappa = 2$. On the sphere, discrete solution by FDM and BEM are presented in Figures 2 and 3, respectively. Our results are validated by numerical examples for sphere. It is noticed that, differs from the FDM mesh, in triangular mesh one uses the specific function to display numerical results. Moreover, since it is difficult to display the results in four-dimensions, in these figures we fixed one of three directions (x or y or z). The numerical solutions



(a) A sphere. (b) A cube.

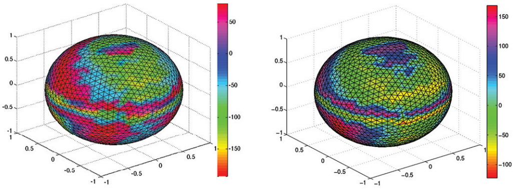
Figure 1. Domain Ω for numerical experiments.



(a) Real part.

(b) Imaginary part.

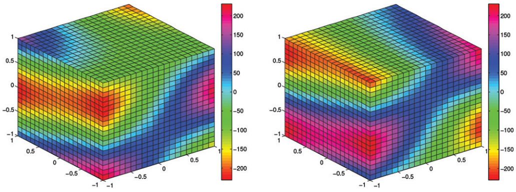
Figure 2. Solution to the interior problem by FDM on the sphere.



(a) Real part.

(b) Imaginary part.

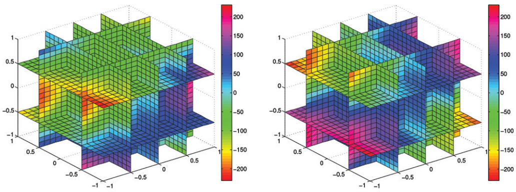
Figure 3. Solution to the interior BVP by BEM on the sphere.



(a) Real part.

(b) Imaginary part.

Figure 4. Solution to the interior problem by FDM on the cube.



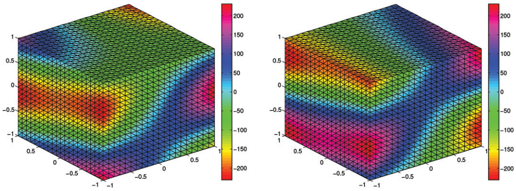
(a) Real part.

(b) Imaginary part.

Figure 5. Solution to the interior problem by FDM inside the cube.

$u(x,y,z)$ are then displayed in a three dimensional coordinate system.

Otherwise, on the cube, Figures 4 and 6 are the simulations results for the numerical solutions.



(a) Real part.

(b) Imaginary part.

Figure 6. Solution to the interior BVP by BEM on the cube.

We also present the solution inside the cube following Figure 5. It can be seen that the obtained waveform in these figures is very clear and it is easy to observe the values of solution under the color bar.

6 CONCLUSION

In this paper we focus on the studying of numerical solution to the homogeneous Helmholtz equation in 3D under the Dirichlet boundary condition. The boundary element method is presented. For instance, we first give the representation formula with respect to internal problem in finite domain and external problem in infinite domain, where the solution should satisfy radiation condition at infinity. Then, these problems can be reformulated as the boundary integral equations in form of singular and double layer potentials. We then propose the discrete formulation of variational Dirichlet boundary value problems. Using polynomial piecewise constant basic functions for approximation of solution we obtain a sparse system of equations. The discretization of the problem is solved by the BEM iterative method. In addition, the classical FDM scheme is also given to give comparative results. Many numerical computational examples on standard domains are implemented which validate the correctness and effectiveness of the algorithm. This work gives an idea to solving the 3D non-homogeneous Helmholtz equation with different boundary conditions numerically, that will be analysed in future research.

REFERENCES

Burton, A.J. & G.F. Miller (1971). The application of integral equation methods to the numerical solution of some exterior boundary-value problems. *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences 323*(1553), 201–210.

Colton, D. (1998). Inverse acoustic and electromagnetic scattering theory, second edition.

Erlangga, A. (2005). *A Robust and Efficient Iterative Method for the Numerical Solution of the Helmholtz Equation*.

Goldstein, C.I. (1982). A fem for solving helmholtz type equations in waveguides and other unbounded domains. *Mathematics of Computation 39*(160), 309–324.

Hegedus, G. (2009). The numerical solution of the three dimensional helmholtz equation with sommerfield boundary conditions. *PIERS Proceedings, Moscow, Russia*.

Ihlenburg, F. & I. Babuska (1997). Solution of Helmholtz problems by knowledge-based fem. *Computer Assisted Mechanics and Engineering Sciences 4*, 397–415.

Jan., Z. (2009). The boundary element method for dirichlet-newmann boundary value problem. *Ostrava*, 31.

Liu, J., G. Nakamura, & R. Potthast (2007). A new approach and error analysis for reconstructing the scattered wave by the point source method. *Journal of Computational Mathematics 25*(2), 113–130.

O'Neil, M., A. Klockner, A. Barnett, & L. Greengard (2013). Quadrature by expansion: A new method for the evaluation of layer potentials. *J. Comput. Phys 252*, 332–349.

Olaf, S. & R. Sergej (2007). The fast solution of boundary integral equations. *New York: Springer-Verlag*, 279.

Sauter, S. & C. Schwab (2010). Boundary element methods. Springer-Verlag, Heidelberg.

Spence, E.A., & A. Fokas (2009). A new transform method I: domain-dependent funcdamental solutions and integral representations. *Proceeding of the Royal Society, series A*, 1–23.

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

A study of stochastic FEM method for porous media flow problem

R. Blaheta
Institute of Geonics of the CAS, Ostrava, Czech Republic

M. Béreš & S. Domesová
VŠB—Technical University of Ostrava, Ostrava-Poruba, Czech Republic
Institute of Geonics of the CAS, Ostrava, Czech Republic

ABSTRACT: The paper provides an overview of the stochastic Finite Element Method (FEM) for the investigation of the flow in heterogeneous porous materials with a microstructure being a Gaussian random field. Quantities characterizing the flow are random variables and the aim is to estimate their probability distribution. The integral mean of the velocity over the domain is one of these quantities, which is numerically analyzed for a described model problem. The estimation of those quantities is realized using the standard Monte Carlo method and the multilevel Monte Carlo method. The paper also concerns the use of the mixed finite element method for the solution of the Darcy flow and efficient assembling and solving of the arising linear systems.

1 INTRODUCTION

Many natural materials, like geomaterials and biomaterials, possess a high level of heterogeneity which has to be properly treated for understanding and reliable modelling of processes in these materials. As a special case, we shall consider groundwater flow, which is important in many applications, as e.g. filtration and waste isolation. The groundwater flow can be further completed by transport of chemicals and pollutants or connected with deformation of the porous matrix.

The groundwater flow can be described by the boundary value problem

$$\begin{aligned} -\mathrm{div}(k\nabla p) &= 0 && \text{in } \Omega \\ p &= \hat{p} && \text{on } \Gamma_D \\ (-k\nabla p)\cdot n &= 0 && \text{on } \Gamma_N, \end{aligned} \tag{1}$$

where p is the pore (water) pressure, k is permeability, $u = -k\nabla p$ is the Darcy's velocity, $\hat{p}$ is a given Dirichlet type boundary condition on $\Gamma_D \subset \partial\Omega$ and no flow is assumed as the Neumann type boundary condition on $\Gamma_D \subset \partial\Omega$, n is the unit outer normal to $\partial\Omega$.

We shall consider a two field form of the above boundary value problem with two basic variables $p:\Omega \to R^1$ and $u:\Omega \to R^n$

$$\begin{aligned} \left.\begin{aligned} k^{-1}u + \nabla p &= g \\ \mathrm{div}(u) &= f \end{aligned}\right\} & \quad \text{in } \Omega \\ p = \hat{p} & \quad \text{on } \Gamma_D \\ u \times n = u_n = 0 & \quad \text{on } \Gamma_N. \end{aligned} \tag{2}$$

We will assume that $k = k(x,\omega)$ is a random variable, $x \in \Omega$ and $\omega \in S$. Here S is a sample space equipped by a suitable probability model with given parameters. Then the model outputs as p, u and another quantities $J(p,u)$, e.g. the averages

$$\langle \nabla p \rangle = \frac{1}{|\Omega|}\int_\Omega \nabla p \tag{3}$$

and

$$\langle u \rangle = \frac{1}{|\Omega|}\int_\Omega -k\nabla p u \tag{4}$$

will be also random variables and we will be interested in their characteristics as the mean (expectation) $\mathbb{E}$ and variance $\mathbb{V}$.

2 STOCHASTIC MICROSTRUCTURE

The permeability $k = k(x,\omega)$ can be considered as a random field in the domain Ω or in selected points within Ω. Especially, we shall assume that

$$\ln(k(x,\cdot)) = c_1\phi,\ \phi \in N(\mu,\sigma^2), \tag{5}$$

where $N(\mu,\sigma^2)$ denote the normal distribution with the mean μ and variance σ^2. This lognormal character of permeability is supported by experimental tests on rock as well as experimentally found logarithmic relation between permeability and porosity, see (Nelson et al. 1994, Freeze 1975). Thus ϕ in (5) could be interpreted as the porosity which gives to (5) a physical meaning.

If $X \in R^n$ is a random field, such that $X_i \in N(0,1)$, then the random field k connected with selected points $x^{(i)} \in \Omega, i = 1,\ldots,n$, can be generated as

$$\ln(k) = c_1(\sigma X + \mu), \tag{6}$$

i.e. $k = e^{c_1\mu}e^{c_1\sigma X}$. For numerical experiments we shall use $c_1 = 1, \mu = 0$, i.e.

$$k = e^{\sigma X}. \tag{7}$$

In this case the components of k have lognormal distribution with the mean $e^{\sigma^2/2}$ and variance $\left(e^{\sigma^2} - 1\right)e^{\sigma^2}$.

The random field X can be further smoothed by correlation, which provides the correlated random field X^c. The correlation is frequently described as an exponential expression involving a correlation length λ, e.g.

$$\begin{aligned} c(x,y) &= c(X^c(x), X^c(y)) \\ &= \sigma^2 exp\left(-\|x-y\|/\lambda\right). \end{aligned} \tag{8}$$

Different methods can be used to generate the correlated random field. The Choleski factorization of the correlation matrix C is probably the most straightforward one and will be used within the experiments in this paper. Further methods as a technique based on the discrete Fourier transform can be found in the literature, see e.g. (Lord, Powell, & Shardlow 2014, Powell 2014).

Given the set of selected points $x^{(i)} \in \Omega, i = 1,\ldots,n$, we can define the correlation matrix C by

$$\begin{aligned} C &= \mathbb{E}((X - \mathbb{E}(X))(X - \mathbb{E}(X))^T) \\ &= \mathbb{E}(XX^T) - \mathbb{E}(X)\mathbb{E}(X^T) \end{aligned} \tag{9}$$

In the case of $\mathbb{E}(X) = 0$, it provides

$$C = \mathbb{E}(XX^T), \; C_{ij} = c(x^{(i)}, x^{(j)}). \tag{10}$$

Theorem 1. *(Generation of the correlated random field). Let $C = LL^T$ be the Choleski factorization of C, X be an uncorrelated random field, $X_i \in N(0,1)$. Then $X^c = LX$ is the correlated random field with correlation matrix C. We can write $X^c \in N(0,C)$.*

Proof. If X_i are uncorrelated and have zero mean and unit variance for any *i*, then $\mathbb{E}(X_iX_j) = \delta_{ij}$ and therefore $\mathbb{E}(XX^T) = I$. The correlation matrix is SPD and therefore the Choleski factorization exists. For $X^c = LX$, where L is the Choleski factor, it holds

$$\begin{aligned} \mathbb{E}(X^c(X^c)^T) &= \mathbb{E}(LXX^TL^T) \\ &= L\mathbb{E}(XX^T)L^T \\ &= LL^T = C \end{aligned} \tag{11}$$

Note that the identity $\mathbb{E}(LXX^TL^T) = L\mathbb{E}(XX^T)L^T$ follows from the linearity of the expectation operator $\mathbb{E}$. □

2.1 *Model problem*

As a model problem, we shall consider the groundwater flow given by equation (1) on the unit square $\Omega = \langle 0,1 \rangle \times \langle 0,1 \rangle$ with the specific boundary conditions—the pressure difference in x_1 direction, see Figure 1.

We shall be interested in different quantities as e.g.

- $u(0.5,0.5)$,
- $k_{eff} = \int_0^1 u(1,x_2)dx_2$,
- $\langle u \rangle = \frac{1}{|\Omega|}\int_\Omega u = \frac{-1}{|\Omega|}\int_\Omega k\nabla p^{(i)}$.

For the realization of this calculations the mixed FEM method can be used, see section 4. If the permeability k will be a random field in Ω, then these quantities will be also random variables and we shall compute their characteristics like the expectation and variance.

2.2 *Visualization of the generated fields*

For numerical experiments with the model problem, we use different values $\sigma \in \{1,2,4\}$ and $\lambda \in \{0.3, 0.1\}$.

The following figures show the visualization of the generated random field k for six combinations of parameters σ and λ value. All of the Gaussian

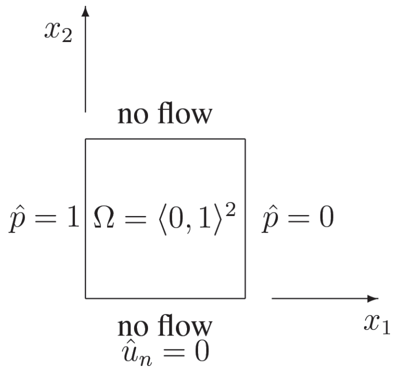
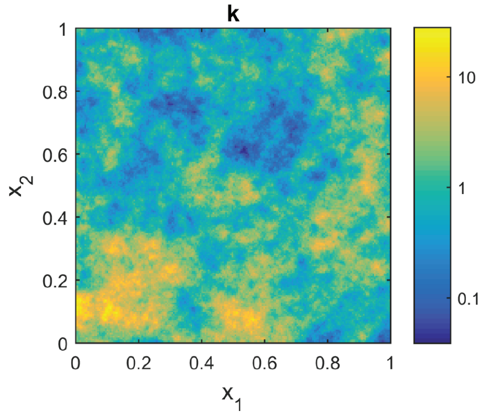
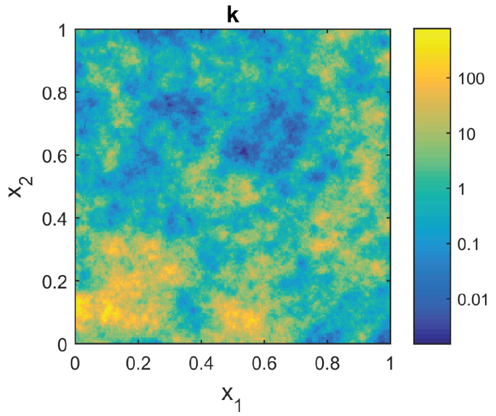


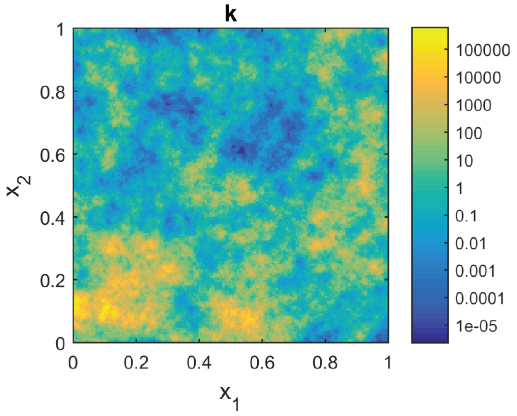
Figure 1. Test problem with pressure difference in x_1 direction.



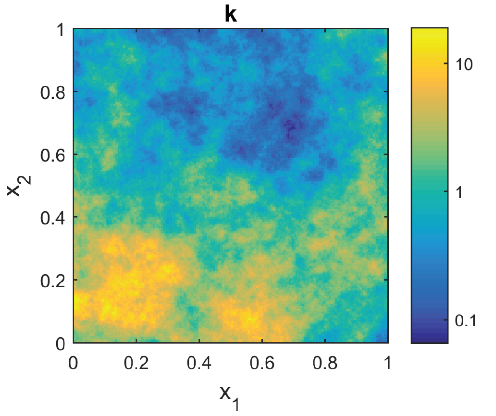
Figures 2. Random field for parameters: $\sigma = 1$, $\lambda = 0.1$.



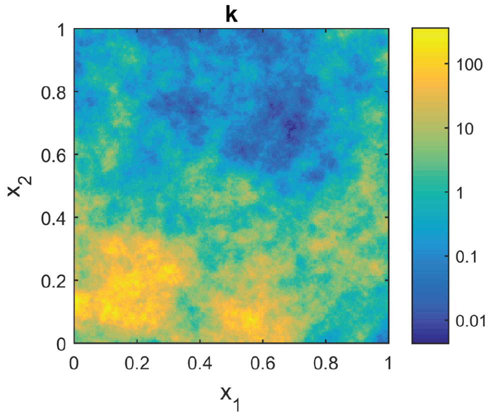
Figures 3. Random field for parameters: $\sigma = 2$, $\lambda = 0.1$.



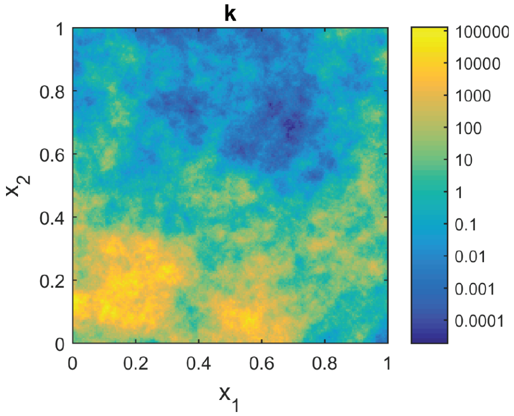
Figures 4. Random field for parameters: $\sigma = 4$, $\lambda = 0.1$.



Figures 5. Random field for parameters: $\sigma = 1$, $\lambda = 0.3$.



Figures 6. Random field for parameters: $\sigma = 2$, $\lambda = 0.3$.



Figures 7. Random field for parameters: $\sigma = 4$, $\lambda = 0.3$.

random fields were created from the same random vector X, where $X_i \sim N(0,1)$, so we can observe the effect of the parameters σ and λ changes on the material microstructure.

The Figures 2, 3, 4, 5, 6, 7 show that the changes of the parameter σ affects only the logarithmic scale of the values, which is caused by the linear relation between $\log k$ and σ^2. The influence of the parameter λ can be observed in a smoother material with growing λ.

3 MONTE CARLO METHODS

Consider the Darcy flow model problem. We are interested in the estimation of the quantities

$$u(0.5,0.5),\, k_{eff} \text{ and } \langle u \rangle. \tag{12}$$

In the case of the Monte Carlo (MC) simulations, we consider this quantities as random variables.

3.1 *Standard Monte Carlo method*

Using the standard MC method, the expectation $\mathbb{E}(\phi)$ of a random variable ϕ is estimated as a sample average

$$\frac{1}{N}\sum_{n=1}^{N} \phi^{(n)}, \tag{13}$$

where $\phi^{(n)}$ for $n \in \{1,\ldots,N\}$ are random samples of ϕ. The estimated probability distribution of the random variables is also described by the sample standard deviation, the estimated probability density function (pdf) and cumulative distribution function (cdf).

The variance of the MC estimator is calculated as

$$V_{MC} = \frac{1}{N}s^2, \tag{14}$$

where s is the sample standard deviation.

The experiments were performed with the following parameters: grid size: 200 × 200, $\sigma \in \{1,2,4\}$, $\lambda \in \{0.1, 0.3\}$, number of experiments: $2 \cdot 10^4$.

The following tables show the estimated sample average and sample standard deviation for the random variables $u_{x_1}(0.5,0.5), u_{x_2}(0.5,0.5), \langle u \rangle_{x_1}$ and $\langle u \rangle_{x_2}$. The values after the ± symbol correspond to the 95% confidence interval for the estimated value. For the random variable k_{eff} the same estimation as for $\langle u \rangle_{x_1}$ was obtained. The graphs in Figure 8 show the pdf and cdf estimation for the random variable $\langle u \rangle_{x_1}$.

3.2 *Multilevel Monte Carlo method*

For the mean value $\mathbb{E}(\phi_L)$ of a random variable $\phi = \phi_L$ we can write

$$\mathbb{E}(\phi_L) = \mathbb{E}(\phi_0) + \sum_{l=1}^{L} \mathbb{E}(\phi_l - \phi_{l-1}). \tag{15}$$

This leads to the multilevel Monte Carlo (MLMC) estimator

$$\mathbb{E}(\phi_L) \approx \frac{1}{N_0}\sum_{n=1}^{N_0} \phi_0^{(n)} + \sum_{l=1}^{L} \frac{1}{N_l}\sum_{n=1}^{N_l}\left(\phi_l^{(n)} - \phi_{l-1}^{(n)}\right), \tag{16}$$

see (Cliffe, Giles, Scheichl, & Teckentrup 2011, Barth, Schwab, & Zollinger 2011). For different levels $l \in \{1,\ldots,L\}$ the values $\phi_l^{(n)} - \phi_{l-1}^{(n)}$ are independent. However the values $\phi_l^{(n)}$ and $\phi_{l-1}^{(n)}$ for specific $n \in \{1,\ldots,N\}$ are correlated.

The variance of the MLMC estimator can be calculated as

$$V_{MLMC} = \sum_{l=0}^{L} \frac{1}{N_l} s_l^2, \tag{17}$$

where s_l is the sample standard deviation on the level l.

This approach was applied to the model problem with the grid size $d \times d$. We were interested in the random variable $\phi = \phi_L = \langle u \rangle_{x_1}^{(d)}$, i.e. the integral mean of the velocity over the $\langle 0,1 \rangle \times \langle 0,1 \rangle$ domain calculated for the grid size $d \times d$. Samples of the random variable $\phi_{L-1} = \langle u \rangle_{x_1}^{(d/2)}$ are calculated as the integral mean of the velocity for the grid $\frac{d}{2} \times \frac{d}{2}$, etc.

There are different ways of calculating the coarse grid approximation ϕ_{l-1} of ϕ_l in order to achieve strong correlation between this two random variables (high correlation between ϕ_{l-1} and ϕ_l leads to low variance on the MLMC level l). In this paper we describe two possible procedures for the coarse grid approximation.

Procedure 1: Coarse grid approximation preserving the Gaussian random field distribution

The samples $\phi_l^{(n)}$ and $\phi_{l-1}^{(n)}$ should be correlated, therefore it is necessary to determine the way of $\phi_{l-1}^{(n)}$ calculation. The value of $\phi_l^{(n)}$ corresponds to a specific sample $k^{(d)}$ of the Gaussian random field, which was obtained for a random vector X, where $X_i \sim N(0,1), i \in \{1,\ldots,d^2\}$. To obtain the value $\varphi_{l-1}^{(n)}$ we first create a coarse material $k^{(d/2)}$ from a

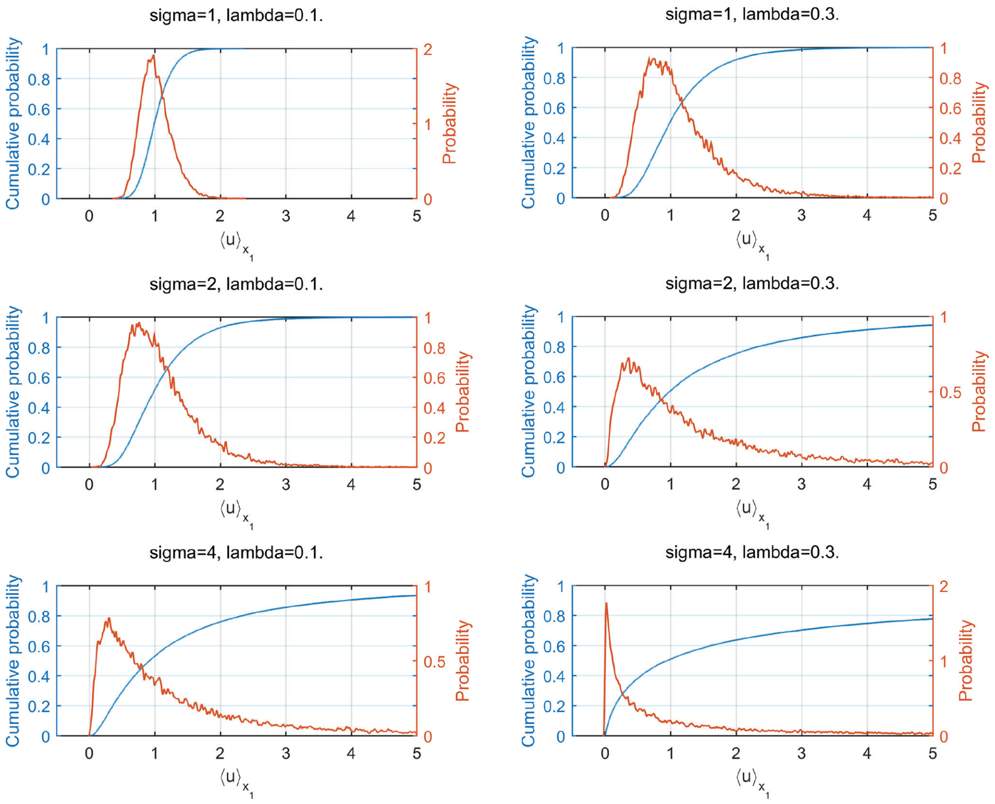


Figure 8. Estimated pdf and cdf of $\langle u \rangle_{x_1}$ for $\lambda = 0.1$ (left) and $\lambda = 0.1$ (right).

Table 1. Sample average of $u_{x_1}(0.5, 0.5)$.

	$\lambda = 0.3$	$\lambda = 0.1$
$\sigma = 1$	1.135 ± 0.0132	1.0409 ± 0.0096
$\sigma = 2$	1.7244 ± 0.055	1.1649 ± 0.0251
$\sigma = 4$	11.9402 ± 1.8383	1.9334 ± 0.1447

Table 2. Sample standard deviation of $u_{x_1}(0.5, 0.5)$.

	$\lambda = 0.3$	$\lambda = 0.1$
$\sigma = 1$	0.9504 ± 0.0093	0.6902 ± 0.0068
$\sigma = 2$	3.9656 ± 0.0389	1.8104 ± 0.0177
$\sigma = 4$	132.6336 ± 1.2999	10.4422 ± 0.1023

Table 3. Sample average of $u_{x_2}(0.5, 0.5)$.

	$\lambda = 0.3$	$\lambda = 0.1$
$\sigma = 1$	0.0008 ± 0.0062	-0.0006 ± 0.0058
$\sigma = 2$	-0.0276 ± 0.03	0.002 ± 0.0169
$\sigma = 4$	0.512 ± 1.1175	0.0173 ± 0.0968

Table 4. Sample standard deviation of $u_{x_2}(0.5, 0.5)$.

	$\lambda = 0.3$	$\lambda = 0.1$
$\sigma = 1$	0.4492 ± 0.0044	0.4201 ± 0.0041
$\sigma = 2$	2.1625 ± 0.0212	1.2215 ± 0.012
$\sigma = 4$	80.627 ± 0.7902	6.9851 ± 0.0685

Table 5. Sample average of $\langle u \rangle_{x_1}$.

	$\lambda = 0.3$	$\lambda = 0.1$
$\sigma = 1$	1.1228 ± 0.0083	1.018 ± 0.0032
$\sigma = 2$	1.6821 ± 0.0324	1.0983 ± 0.0079
$\sigma = 4$	10.0822 ± 0.9339	1.7447 ± 0.0409

vector Y of length $\frac{1}{4}d^2$, which is calculated from the vector X values. For example

$$Y_1 = \frac{1}{2}\left(X_1 + X_2 + X_{d+1} + X_{d+2}\right) \tag{18}$$

Table 6. Sample standard deviation of $\langle u \rangle_{x_1}$.

	$\lambda = 0.3$	$\lambda = 0.1$
$\sigma = 1$	0.6024 ± 0.0059	0.2335 ± 0.0023
$\sigma = 2$	2.34 ± 0.0229	0.5695 ± 0.0056
$\sigma = 4$	67.3806 ± 0.6604	2.9495 ± 0.0289

Table 7. Sample average of $\langle u \rangle_{x_2}$.

	$\lambda = 0.3$	$\lambda = 0.1$
$\sigma = 1$	0.0005 ± 0.0015	0.0001 ± 0.001
$\sigma = 2$	-0.003 ± 0.0062	-0.0002 ± 0.0022
$\sigma = 4$	-0.02 ± 0.1623	-0.0023 ± 0.0109

Table 8. Sample standard deviation of $\langle u \rangle_{x_2}$.

	$\lambda = 0.3$	$\lambda = 0.1$
$\sigma = 1$	0.1098 ± 0.0011	0.069 ± 0.0007
$\sigma = 2$	0.4499 ± 0.0044	0.1596 ± 0.0016
$\sigma = 4$	11.7075 ± 0.1147	0.7865 ± 0.0077

(weighted arithmetic mean), etc. This approach ensures that the values Y_i follow the $N(0,1)$ distribution, therefore the obtained material $k^{(d/2)}$ is also a Gaussian random field. The value of $\phi_{l-1}^{(n)}$ is then calculated on the coarse grid and remains correlated with the value of $\phi_l^{(n)}$.

The MLMC method was tested on the model problem with grid size 200 × 200, therefore is was possible to use three coarser grids of dimensions 100 × 100, 50 × 50 and 25 × 25. The numbers of samples N_l to be performed on specific levels were calculated from a preliminary simulation run. In this run the same number of samples was performed on each level and then the values of computation time T_l and sample standard deviation s_l were estimated for each level. The values of N_l were then calculated according to (Cliffe, Giles, Scheichl, & Teckentrup 2011) as $N \cdot \sqrt{s_l^2 / T_l}$, where N is a constant common to all the levels.

The Table 9 presents the results of the MLMC method that can be compared with the MC method results (Table 5).

The MLMC results were calculated with different number of samples (i.e. different computation time) than the MC results, therefore we propose the following indicator for comparison of the efficiency. The efficiency of the MLMC estimator in comparison to the MC estimator will be calculated as

$$\frac{V_{MC}}{V_{MLMC}} \cdot \frac{T_{MC}}{T_{MLMC}}, \tag{19}$$

where T_{MC} is the total time of the MC simulation and T_{MLMC} time of the MLMC simulation, see the Table 10.

The value 1 for $\sigma = 4$ and $\lambda = 0.3$ in Table 10 is caused by the fact that in this case it was evaluated in the preliminary run, that only one level should be used, i.e. it is the standard MC method. In the remaining cases all of the 4 levels were used.

The Table 11 shows the values of s_l^2 on each of the levels $l \in \{1,\ldots,4\}$ calculated in the preliminary run (level l = 1 corresponds to the coarsest grid, while the remaining values present the difference between the fine and coarse grid on the given level). We used these values to calculate the numbers of samples to be executed on each of the MLMC levels.

The following table shows the ratio of the numbers of samples, that were used on different levels.

Procedure 2: Coarse grid approximation as arithmetic mean of correlated random field

In this case we use a similar approach as in the procedure 1, but the key difference is that the smoothing is applied to the correlated values,

$$Y_1^c = \frac{1}{4}\left(X_1^c + X_2^c + X_{d+1}^c + X_{d+2}^c\right) \tag{20}$$

(arithmetic mean). A disadvantage is that this coarse grid approximation is not the same random field, so in the lower MLMC levels we always need to construct a new covariance matrix and its Choleski factorization. The new covariance matrix is created by averaging of elements of the fine grid covariance matrix according to the fine grid to coarse grid elements mapping, this con-

Table 9. MLMC method results for $\langle u \rangle_{x_1}$.

	$\lambda = 0.3$	$\lambda = 0.1$
$\sigma = 1$	1.1302 ± 0.0039	1.0189 ± 0.0007
$\sigma = 2$	1.6744 ± 0.0189	1.1003 ± 0.0021
$\sigma = 4$	9.6647 ± 0.5259	1.745 ± 0.0152

Table 10. MLMC/MC efficiency calculated via (19).

	$\lambda = 0.3$	$\lambda = 0.1$
$\sigma = 1$	1.9382	7.7148
$\sigma = 2$	1.1841	5.7974
$\sigma = 4$	1	3.0011

struction comes from the linearity of the covariance. This disadvantage is compensated by very high correlation between fine grid and coarse grid approximation.

In the Figure 9 we show an example of coarse grid approximations for both procedures.

The Table 13 presents the results obtained for $\langle u \rangle_{x_1}$ (including 95% confidence interval). The calculated efficiency compared to the MC estimator via formula (19) can be seen in the Table 14.

The Table 15 shows the values of s_l^2 on each of the levels $l \in \{1, \ldots, 4\}$ calculated in the preliminary run.

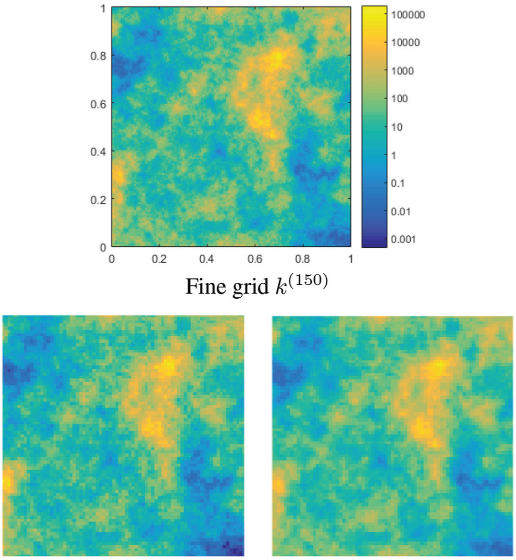
The following table shows the ratio of the numbers of samples, that were used on different levels. In all the six cases at least three levels were used.

Table 11. Variance on each MLMC level.

σ	λ	$l=4$	$l=3$	$l=2$	$l=1$
1	0.3	$4.4{\cdot}10^{-2}$	$4.1{\cdot}10^{-2}$	$3.6{\cdot}10^{-2}$	0.35
	0.1	$1.4{\cdot}10^{-3}$	$1.2{\cdot}10^{-3}$	$1{\cdot}10^{-3}$	$5.4{\cdot}10^{-2}$
2	0.3	1.4	0.84	1	4.5
	0.1	$1.1{\cdot}10^{-2}$	$1{\cdot}10^{-2}$	$1.1{\cdot}10^{-2}$	0.29
4	0.3	$1{\cdot}10^{5}$	$5.1{\cdot}10^{3}$	$9.3{\cdot}10^{3}$	$2.3{\cdot}10^{4}$
	0.1	0.63	0.76	0.91	4.5

Table 12. Ratios of N_l/N_4 values for the six combinations of parameters.

σ	λ	N_4	N_3	N_2	N_1
1	0.3	1	2.23	4.70	33.86
	0.1	1	2.12	4.53	75.15
2	0.3	1	1.73	4.35	20.97
	0.1	1	2.19	5.04	60.55
4	0.3	1	-	-	-
	0.1	1	2.51	6.21	32.24



Fine grid $k^{(150)}$

Coarse grids $k^{(75)}$ (left procedure 1; right procedure 2)

Figure 9. Comparison of coarse grid approximations.

Table 13. MLMC method results for $\langle u \rangle_{x_1}$.

	$\lambda = 0.3$	$\lambda = 0.1$
$\sigma = 1$	1.1298 ± 0.0011	1.0189 ± 0.0004
$\sigma = 2$	1.6945 ± 0.0044	1.1001 ± 0.0012
$\sigma = 4$	10.3715 ± 0.2517	1.7434 ± 0.0087

Table 14. MLMC efficiency calculated via (19).

	$\lambda = 0.3$	$\lambda = 0.1$
$\sigma = 1$	101.2564	90.7969
$\sigma = 2$	87.3758	72.5988
$\sigma = 4$	23.1090	38.6652

Table 15. Variance on each MLMC level.

σ	λ	$l=4$	$l=3$	$l=2$	$l=1$
1	0.3	$4.4{\cdot}10^{-8}$	$1.3{\cdot}10^{-6}$	$6.2{\cdot}10^{-6}$	0.36
	0.1	$1.1{\cdot}10^{-7}$	$1.9{\cdot}10^{-6}$	$9.6{\cdot}10^{-6}$	$5.4{\cdot}10^{-2}$
2	0.3	$7.5{\cdot}10^{-6}$	$2.0{\cdot}10^{-4}$	$1.0{\cdot}10^{-3}$	5.5
	0.1	$5.1{\cdot}10^{-6}$	$9.9{\cdot}10^{-5}$	$4.1{\cdot}10^{-4}$	$3.1{\cdot}10^{-1}$
4	0.3	$1.4{\cdot}10^{-1}$	1.7	$6.2{\cdot}10^{1}$	$1.1{\cdot}10^{4}$
	0.1	$2.1{\cdot}10^{-3}$	$2.0{\cdot}10^{-2}$	$1.2{\cdot}10^{-1}$	7.2

Table 16. Ratios of N_l/N_4 values for the six combinations of parameters.

σ	λ	N_4	N_3	N_2	N_1
1	0.3	1	12.20	60.45	31410.66
	0.1	1	9.66	49.56	9127.92
2	0.3	1	11.56	47.65	10186.95
	0.1	1	9.66	47.36	2942.09
4	0.3	1	11.32	41.83	2550.90
	0.1	1	8.04	38.23	906.28

4 MIXED FEM DISCRETIZATION AND SOLUTION

The groundwater flow (1) can be implemented by the mixed finite element method, e.g. in the way described in (Cliffe, Graham, Scheichl, & Stals 2000, Blaheta, Hasal, Domesová, & Béreš 2014). The first advantage of the mixed formulation is in more accurate approximation of both pressures and velocities. The random permeability field sampling then requires repeated assembling and solving of the mixed FEM system, which has the following saddle point structure

$$\begin{array}{lcl} Mu \quad +B^T p & = & G \\ Bu & = & F \end{array} \tag{21}$$

Note that only the velocity mass matrix M depends on realization $\omega \in S$,

$$M_{ij} = M_{ij}(\omega) = \int_\Omega k(\omega)^{-1} \Phi_j \Phi_i \, d\Omega, \tag{22}$$

where Φ_j, Φ_i are basis functions in the lowest order Raviart-Thomas space. The repeated assembling of the matrix

$$A = \begin{bmatrix} M & B^T \\ B & 0 \end{bmatrix} \tag{23}$$

is therefore restricted to the pivot block. A fast assembling of both M and B is implemented in the RT1 code, see (Blaheta, Hasal, Domesová, & Béreš 2014).

As a solution of the system, the discretized pressure p and velocity u is obtained. The following graphs at Figures 10, 11 and 12 show the visualization of the solution for an example given by the Gaussian random field 3.

When repeatedly solving the system (21) by a direct method, the benefit of B not dependent on sampling is not exploited. The use of an iterative solution method, such as MINRES or GMRES, with block preconditioner, provides the chance to save some effort as only the block corresponding to M is changing. It is the case the following preconditioners

$$P_1 = \begin{bmatrix} \tilde{M} + B^T W^{-1} B & \zeta B^T \\ 0 & W \end{bmatrix}, \tag{24}$$

with $\tilde{M}$ being a suitable approximation to M and W being a block independent on sampling, e.g. $W = \frac{1}{r} I$, where r is a (large) regularization parameter, $\zeta \in \{0,1,2\}$. Special cases are $\tilde{M}$ being a mass matrix for the mean value of the permeability k, $\tilde{M} = \frac{\mathrm{trace}(M)}{\mathrm{trace}(I),} I$ and $W = BB^T$, when $B^T W^{-1} B$ becomes a projection.

Other possibilities are preconditioners for the transformed system with the matrix

$$\mathcal{A} = \begin{bmatrix} M & B^T \\ -B & 0 \end{bmatrix} \tag{25}$$

as the HSS preconditioner

$$\mathcal{P}_2 = \begin{bmatrix} M + \alpha I & 0 \\ 0 & \alpha I \end{bmatrix} \begin{bmatrix} \alpha I & B^T \\ -B & \alpha I \end{bmatrix} \tag{26}$$

or relaxed HSS preconditioner

$$\mathcal{P}_3 = \begin{bmatrix} M & 0 \\ 0 & \alpha I \end{bmatrix} \begin{bmatrix} \alpha I & B^T \\ -B & 0 \end{bmatrix} \tag{27}$$

with a suitable parameter α.

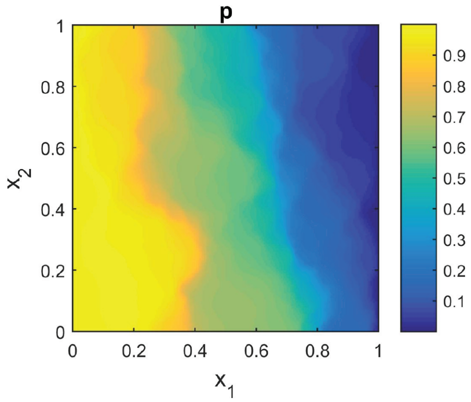


Figure 10. Discretized pressure p.

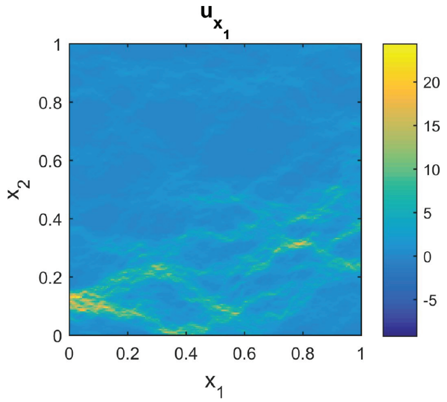


Figure 11. Discretized velocity u (first coordinate).

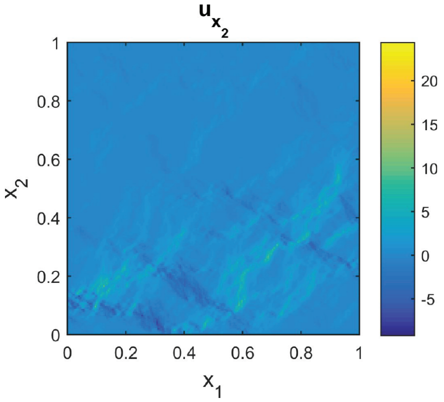


Figure 12. Discretized velocity *u* (second coordinate).

5 CONCLUSIONS

The article presents the first results of the authors in the field of the stochastic partial differential equations (PDEs) or stochastic FEM methods. The simple and multilevel Monte Carlo methods are used as tools for stochastic simulations.

We study the mixed FEM calculation of the Darcy flow problem with stochastic material coefficients. We focused on the characterizations of the velocity, especially on the integral average of velocity over the domain and the velocity in the middle of the the domain.

The MC approach was used for the estimation of the expected value, variance and distribution of the studied random variables.

The MLMC method was used for the more efficient estimation of the expected value of the random variable $\langle u \rangle_{x_1}$. We presented two approaches to the coarse grid approximation, the first one is straightforward and preserves the Gaussian random distribution on the coarse grid, but was inefficient due to low correlation between the fine and coarse grid approximation. The second one suffers from the more difficult sample generation on the coarse grids. Nevertheless the second approach was more efficient than the first one, according to Tables 10 and 14. Depending on the problem parameters λ and σ we achieved variance reduction from about 23 × to 101 ×.

The work is in progress, we plan to use a different approach to the Gaussian random field generation based on the Karhunen-Loève (K-L) decomposition. This will allow us to solve the problem on larger grids and as well it provides a different way of using the MLMC method (MLMC levels will correspond to the levels of the K-L decomposition). The K-L decomposition also provides a different approach to the stochastic PDEs solving by e.g. the collocation method or the stochastic Galerkin method.

ACKNOWLEDGEMENT

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project IT4Innovations excellence in science—LQ1602“.

REFERENCES

Barth, A., C. Schwab, & N. Zollinger (2011). Multi-level monte carlo finite element method for elliptic pdes with stochastic coefficients. *Numerische Mathematik 119*(1), 123–161.

Blaheta, R., M. Hasal, S. Domesová, & M. Béreš (2014). Rt1-code: A mixed rt0-p0 raviartthomas finite element implementation. http://www.ugn.cas.cz/publish/software/RT1-code/RT1-code.zip.

Cliffe, K., M. Giles, R. Scheichl, & A.L. Teckentrup (2011). Multilevel monte carlo methods and applications to elliptic pdes with random coefficients. *Computing and Visualization in Science 14*(1), 3–15.

Cliffe, K., I.G. Graham, R. Scheichl, & L. Stals (2000). Parallel computation of flow in heterogeneous media modelled by mixed finite elements. *Journal of Computational Physics 164*(2), 258–282.

Freeze, R.A. (1975). A stochastic-conceptual analysis of one-dimensional groundwater flow in nonuniform homogeneous media. *Water Resources Research 11*(5), 725–741.

Lord, G.J., C.E. Powell, & T. Shardlow (2014). *An Introduction to Computational Stochastic PDEs*. Cambridge University Press.

Nelson, P.H. et al. (1994). Permeability-porosity relationships in sedimentary rocks. *The log analyst 35*(03), 38–62.

Powell, C.E. (2014). Generating realisations of stationary gaussian random fields by circulant embedding. https://www.nag.co.uk/doc/techrep/pdf/tr1_14.pdf.

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Parallel resolution of the Schur interface problem using the Conjugate gradient method

M.-P. Tran
Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

ABSTRACT: In numerical analysis, the Schur complement is the heart of domain decomposition method. A lot of promising results have been derived to present its mathematical basis. In this paper, we propose a numerical method of solving Schur interface problem using the conjugate gradient method. The parallel STD algorithm is also described to give comparable results with the proposed method. Some numerical experiments have been performed to show a good parallel efficiency and convergence of our method. Efficient parallel computation requires a few minutes to be completed, and they are much less coupled than the direct solvers. In the rest of this paper, some numerical results have been carried out to show its convergence properties and some open problems are also discussed.

1 INTRODUCTION

Domain decomposition methods are important techniques for the numerical simulation. Domain decomposition can be used in the framework of several discretization method for Partial Differential Equations (PDEs) to get more efficient solution on parallel computing. The basic idea in domain decomposition methods is to split the domain of study into non-overlapping subdomains, on that one has discretized problems are simple and convenient to be solved. Many variants of domain decomposition method have been proposed and investigated in (Milyukova 2001, B. Smith 1996) and references therein. In numerical analysis, the Schur complement method is the basic non-overlapping domain decomposition method, one of the most popular linear solvers. The Schur complement is a directed parallel method, that can be applied to solve any sparse linear equation system. For instance, the parallel Schur complement method was followed by (Mansfield 1990, S Kocak 2010) and a lot of recent literatures. In many practical applications, the preconditioned conjugate gradient method is used because of its simplicity, one can refer to (Meyer 1990). Therefore, it is a convenient framework for the solution to our sparse matrix systems.

In this paper, we present a simple Schur complement using conjugate gradient method for solving one-dimensional Poisson's equation. We only consider the classical PDE in this study to claim the parallel efficiency of the proposed method for a simple linear equations system. The idea of solving other large equations systems using proposed method turns out to be very successful in the same way.

Let $\Omega \subset \mathbb{R}^d$ be an open bounded domain with the boundary $\Gamma = \partial\Omega$. Suppose that we want to solve the following Poisson's equation:

$$-\Delta u = f \quad \text{in } \Omega, \tag{1.1}$$

with Dirichlet boundary condition:

$$u = g, \quad \text{on } \Gamma. \tag{1.2}$$

The Schur complement method splits up the linear system into subproblems. To do so, let us divide Ω into p subdomains $\Omega_1, \Omega_2, ..., \Omega_p$ with share interfaces $\Gamma_1, \Gamma_2, ...$. One divides the entire problem into smaller non-overlapping subdomain problems, then solves the subdomain problems to form interface problem and solves it. This paper will discuss the Schur complement as proposing the parallel implementations for general sparse linear system. One considers the parallel solution to one dimensional case (1D), where the Schur complement system on subdomain interfaces is solved by conjugate gradient method.

The problem (1) and (2) are discretized to get the system:

$$A.U = F, \tag{1.3}$$

where the stiffness matrix A, the load vector f and approximate solution U can be decomposed into p groups, corresponding to subdomains $\Omega_1, \Omega_2, ..., \Omega_p$. In this study, we have just treated the following 1D problem. For general elliptic problem, the Schur complement is more complicated so that it is difficult to find approximate solution by parallel computation.

Suppose that we need to solve numerically the one dimensional PDE with inhomogeneous Dirichlet boundary condition as following:

$$\begin{cases} u_{xx} &= -F(x), \text{ on } \Omega = (a_1,a_2), \\ u(a_1) = \beta, \\ u(a_2) = \gamma. \end{cases} \tag{1.4}$$

Let $v(x) = u((1-x)a_1 + xa_2)$ and make substitution to (1.4), one obtains:

$$\begin{cases} v_{xx} &= -G(x), \text{ on } \Omega' = (0,1), \\ v(0) = \beta, \\ v(1) \ = \gamma, \end{cases} \tag{1.5}$$

where $\mathrm{G(x) = (a_1} - a_2)^2 F((1-x)a_1 + xa_2)$.

Let $\varpi(x) = v(x) - (\beta(1-x) + \gamma x)$, it gives the 1D problem with homogeneous boundary condition as follows:

$$\begin{cases} \varpi_{xx} &= -G(x), \quad \text{on } \Omega' = (0,1) \\ \varpi(0) = \varpi(1) = 0. \end{cases} \tag{1.6}$$

Therefore, without loss of generality, in this paper we only study the following problem with Dirichlet homogeneous boundary condition, as the same as (1.6):

$$\begin{cases} u_{xx} &= -F(x), \quad \text{on } \Omega' = (0,1), \\ u(0) = u(1) = 0. \end{cases} \tag{1.7}$$

where u is unknown, F represents a continuous source term. This problem (1.7) can be rewritten in term of the linear system:

$$A_u = B, \tag{1.8}$$

where A is the sparse matrix. By using the Schur method as proposed as in the paper, the Schur complement matrix S is introduced and the system (1.8) is rewritten in the abbreviated form:

$$S\phi = \psi, \tag{1.9}$$

where the matrix S in the interface problem is related to the entire problem (1.8). This paper also presents a proposed method allows Matlab users to take advantages of the Message Passing Interface (MPI) to design parallel performance. In particular, only the basic send and receive operations: MPI Send and MPI Recv are both blocking calls, respectively. These calls are used to implement programs across multiple processors for parallel computation.

The rest of this paper is organized as follows. In Section 2, we consider the Schur complement method for solving the problem (1.7) and propose to solve problem using parallel preconditioned conjugate gradient method, which is currently one of the most popular domain decomposition methods. The mathematical description is then established by a simple coding example with Matlab MPI (Message Passing Interface) standard where the programs are implemented on multiple processors. In the next section, one presents the parallel STD method in order to compare with the previous Schur complement method, and some Matlab MPI calls are also provided. Section 4 indicates some numerical examples testing both parallel computational methods. Some conclusions are then discussed in the last section to give the validity of method.

2 THE SCHUR INTERFACE PROBLEM

In this section, we present the Schur interface complement method that is applied to solve the problem (1.7). One refers to (Mansfield 1990) the Schur complement method given in the following steps:

1. The domain Ω of Sproblem (1.7) is subdivided into non-overlapping subdomains using parallel graph partition,
2. Rewrite the stiffness matrix A in the linear system (8) in each subdomain and interface,
3. Solve the subdomain problem to calculate the Schur matrix S from each known submatrix,
4. Solve the Schur complement system $SU = G$,
5. Solve the subdomain system to obtain solution in the whole domain by parallel algorithm.

In (1.8), one subdivides the problem into p parts ($p \geq 2$). The vector solution U can be decomposed into p groups, that is $U = (U_1, U_2, ..., U_p)$, where $U_i (i = 1, 2, ..., p)$ are corresponding to domain $\Omega_1, \Omega_2, ..., \Omega_p$, respectively. It is important to notice that the decomposition of Ω into the subdomain Ω_i does not have any cross point. In this study, we also discuss two models of domain decomposition and propose the parallel schemes for solving problem (8) numerically, for the case $p = 2$ and for general $p \geq 2$.

Suppose that the approximation to the weak formulation results of (1.1) and (1.2) is of the form (1.8), the stiffness matrix A is given as the following sparse matrix:

$$A=\frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{bmatrix}. \tag{2.1}$$

2.1 *The case* $p=2$

The case of two subdomains, i.e. $p=2$, is first considered.

Let $k>0$, the domain $\Omega=(0,1)$ is then divided into $2k+2$ grid cells as follows:

$$0 \equiv x_0 \le x_1 \le \cdots \le x_k \le \cdots \le x_{2k+1} \equiv 1 \tag{2.2}$$

We partition the domain into three non-overlapping subdomains denoted by $\Omega_1,\Omega_2,\Omega_3$, respectively. Apparently, one has:

- $\Omega_1=\{x_1,x_2,\cdots,x_k\}$,
- $\Omega_2=\{x_{k+1}\}$,
- $\Omega_3=\{x_{k+2},x_{k+3},\cdots,x_{2k}\}$.

Let us rewrite the stiffness matrix A in (2.1) in term of the block matrix as follows:

$$A=\begin{bmatrix} K_{11} & K_{12} & 0 \\ K_{21} & K_{22} & K_{23} \\ 0 & K_{32} & K_{33} \end{bmatrix}, \tag{2.3}$$

where K_{ij} are defined for $i,j=1,2,3$ as:

- K_{11} and K_{33} are matrices of order k, one denotes $K_{11},K_{33}\in\mathbb{R}^{k\times k}$ as below:

$$K_{11}=K_{33}=\frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \\ 0 & 0 & 0 & \cdots & 2 \end{bmatrix}. \tag{2.4}$$

- K_{12} and $K_{32}\in\mathbb{R}^{k\times 1}$:

$$K_{12}=\frac{1}{h^2}\begin{bmatrix} 0 \\ 0 \\ \cdots \\ -1 \end{bmatrix}; \quad K_{32}=\frac{1}{h^2}\begin{bmatrix} -1 \\ 0 \\ \cdots \\ 0 \end{bmatrix}. \tag{2.5}$$

- K_{21} and $K_{23}\in\mathbb{R}^{1\times k}$:

$$\begin{aligned} K_{21}&=\frac{1}{h^2}\begin{bmatrix} 0 & 0 & \cdots & -1 \end{bmatrix}; \\ K_{23}&=\frac{1}{h^2}\begin{bmatrix} -1 & 0 & \cdots & 0 \end{bmatrix}. \end{aligned} \tag{2.6}$$

- $K_{22}=\frac{2}{h^2}$.

The problem (1.8) can be rewritten in term of block system:

$$\begin{bmatrix} K_{11} & K_{12} & 0 \\ K_{21} & K_{22} & K_{23} \\ 0 & K_{32} & K_{33} \end{bmatrix}\begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix}=\begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix}, \tag{2.7}$$

where U_1,U_3 and U_2 be the vector solution associated with each subdomain Ω_1,Ω_2 and with the interface:

$$\begin{aligned} &U_1=(x_1,\cdots,x_k); \quad U_2=x_{k+1}; \\ &U_3=(x_{k+2},\cdots,x_{2k}), \end{aligned} \tag{2.8}$$

and F_i are the components of the load vector in each region:

$$\begin{aligned} &F_1=F(1:k); \quad F_2=F(k+1); \\ &F_3=F(k+2:2k). \end{aligned} \tag{2.9}$$

Then, the original linear system (1.8) is divided into three subproblems given as:

$$K_{11}U_1+K_{12}U_2=F_1, \tag{2.10}$$

$$K_{21}U_1+K_{22}U_2+K_{23}U_3=F_2, \tag{2.11}$$

$$K_{32}U_2+K_{33}U_3=F_3. \tag{2.12}$$

From the first and third equation (2.10) and (2.12), one obtains that:

$$\begin{cases} U_1=K_{11}^{-1}\left(F_1-K_{12}U_2\right), \\ U_2=K_{33}^{-1}\left(F_3-K_{32}U_2\right). \end{cases} \tag{2.13}$$

Make substitution to the second equation (2.11), we arrive at the Schur complement equation:

$$SU_2=G, \tag{2.14}$$

where the introducing Schur complement matrix

$$S=K_{22}-K_{21}K_{11}^{-1}K_{12}-K_{23}K_{33}^{-1}K_{32} \tag{2.15}$$

and

$$G=F_2-K_{23}K_{33}^{-1}F_3-K_{21}K_{11}^{-1}F_1. \tag{2.16}$$

In the Schur complement method, the solution of linear system can be approximated by first solving the Schur system, and then solving the interior system. Here, the equation (2.14) above can be solved in parallel scheme with two processors ($p = 2$). It is noticed that the computation of the inverse terms in (2.15) and (2.16) can be done in parallel. Let us describe the parallel schemes in the Figures 1 and 2. In these figures, equal works are implemented on two processors to calculate vector G and S_x, respectively.

Then, the equation (2.14) is solved by the parallel preconditioned conjugate gradient method because of its simplicity and efficiency. The numerical scheme is presented as in Figure 3, in which the implemented program has been developed under the Matlab computing environment.

In the Schur complement domain decomposition method, each subdomain is handled by different processors. More precisely, the proposed parallel solver with two processors is introduced.

2.2 *The case for general* p

Similar to the previous section for $p = 2$, when the domain decomposition method is used, the problem domain Ω is to be divided into p subdomains Ω_j ($j = 1,2,3...,p$) as in Figure 4, in which the unknown qualities can be calculated simultaneously in parallel. For instance,

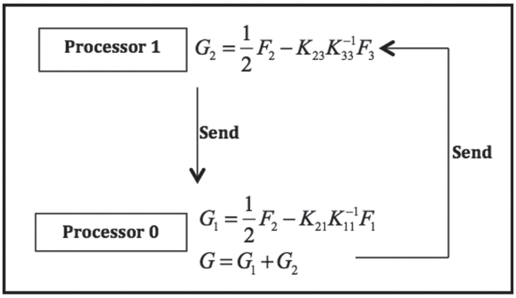

Processor 1
$G_2 = \frac{1}{2}F_2 - K_{23}K_{33}^{-1}F_3$
Send
Send
Processor 0
$G_1 = \frac{1}{2}F_2 - K_{21}K_{11}^{-1}F_1$
$G = G_1 + G_2$

Figure 1. Update G by the parallel scheme with two processors.

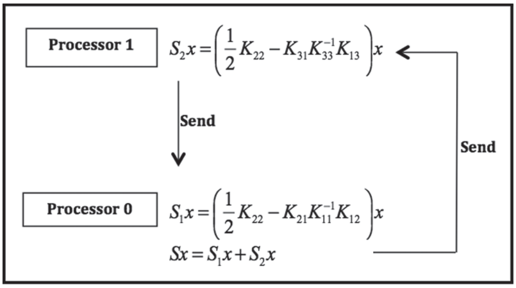
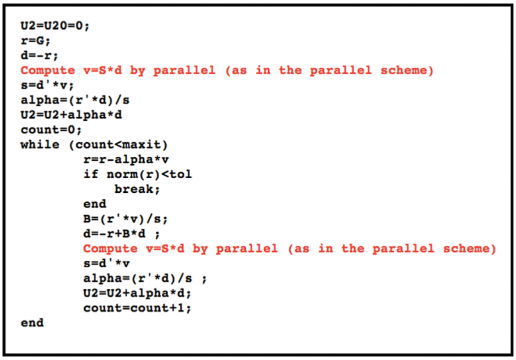

Processor 1
$S_2x = \left(\frac{1}{2}K_{22} - K_{31}K_{33}^{-1}K_{13}\right)x$
Send
Send
Processor 0
$S_1x = \left(\frac{1}{2}K_{22} - K_{21}K_{11}^{-1}K_{12}\right)x$
$Sx = S_1x + S_2x$

Figure 2. Update S_x by the parallel scheme.



```
U2=U20=0;
r=G;
d=-r;
Compute v=S*d by parallel (as in the parallel scheme)
s=d'*v;
alpha=(r'*d)/s
U2=U2+alpha*d
count=0;
while (count<maxit)
        r=r-alpha*v
        if norm(r)<tol
            break;
        end
        B=(r'*v)/s;
        d=-r+B*d ;
        Compute v=S*d by parallel (as in the parallel scheme)
        s=d'*v
        alpha=(r'*d)/s ;
        U2=U2+alpha*d;
        count=count+1;
end
```

Figure 3. Algorithm to solve the equation $Sx = G$ by conjugate gradient method.

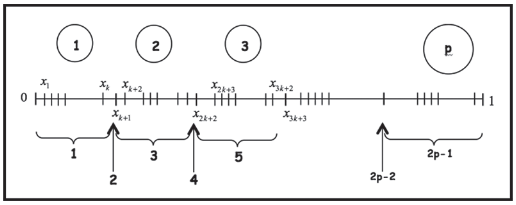


Figure 4. Ω is divided into p non-overlapping subdomains.

- $\Omega_1 = x_1, x_2, ..., x_k$,
- $\Omega_2 = x_{k+1}$,
- $\Omega_3 = x_{k+2}, x_{k+3}, ..., x_{2k+1}$,
- $\Omega_{2p-1} = x_{(p-1)k+p}, ..., x_{pk+(p-1)}$.

The general form of a linear algebraic problem (for general $p > 2$) defined on the domain Ω can also be rewritten in terms of (1.8), where the matrix A in the whole domain Ω is presented as:

$$A = \begin{bmatrix} K_{11} & K_{12} & 0 & \cdots & 0 \\ K_{21} & K_{22} & 0 & \cdots & 0 \\ 0 & K_{32} & K_{33} & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & K_{(2p-2)(2p-2)} & K_{(2p-2)(2p-1)} \\ 0 & \cdots & 0 & K_{(2p-1)(2p-2)} & K_{(2p-1)(2p-1)} \end{bmatrix}, \qquad (2.17)$$

where K_{ij} are defined as following:

- $K_{(2i-1)(2i-1)} \in \mathbb{R}^{k \times k}$ are given by:

$$K_{(2i-1)(2i-1)} = \frac{1}{h^2}K_2 = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \\ 0 & 0 & 0 & \cdots & 2 \end{bmatrix}. \qquad (2.18)$$

- $K_{(2i-1)(2i)}$ and $K_{(2i+1)(2i)} \in \mathbb{R}^{k\times 1}$:

$$K_{(2i-1)(2i)} = \frac{1}{h^2}K_3 = \frac{1}{h^2}\begin{bmatrix} 0 \\ 0 \\ \cdots \\ -1 \end{bmatrix}; \qquad (2.19)$$

$$K_{(2i+1)(2i)} = \frac{1}{h^2}K_5 = \frac{1}{h^2}\begin{bmatrix} -1 \\ 0 \\ \cdots \\ 0 \end{bmatrix}. \qquad (2.20)$$

- $K_{(2i)(2i-1)}$ and $K_{(2i)(2i+1)} \in \mathbb{R}^{1\times k}$:

$$\begin{aligned} K_{(2i)(2i-1)} &= \frac{1}{h^2}K_4 = \frac{1}{h^2}\begin{bmatrix} 0 & 0 & \cdots & -1 \end{bmatrix}; \\ K_{(2i)(2i+1)} &= \frac{1}{h^2}K_6 = \frac{1}{h^2}\begin{bmatrix} -1 & 0 & \cdots & 0 \end{bmatrix}. \end{aligned} \qquad (2.21)$$

- $K_{(2i)(2i)} = 2$.

It can be noticed that since the subdomain Ω_i is disconnected to each other (non-overlap), the corresponding block matrices K_{ij} is also disconnected to each other. This allows us to make an easy parallelization.

The linear sparse system (1.8) is split into several particular blocks:

$$K_{11}U_1 + K_{12}U_2 = F_1, \qquad (2.22)$$

$$K_{21}U_1 + K_{22}U_2 + K_{23}U_3 = F_2, \qquad (2.23)$$

$$K_{32}U_2 + K_{33}U_3 + K_{34}U_4 = F_3, \qquad (2.24)$$

$$\ldots$$

$$K_{(2p-1)(2p-2)}U_{2p-2} + K_{(2p-1)(2p-1)}U_{2p-1} = F_{2p-1}, \qquad (2.25)$$

Discrete solution is obtained in terms of:

$$\begin{aligned} U_1 &= K_{11}^{-1}(F_1 - K_{12}U_2), \\ U_{2p-1} &= K_{(2p-1)(2p-1)}^{-1}\left(F_{2p-1} - K_{(2p-1)(2p-2)}U_{2p-2}\right), \\ U_j &= K_{jj}^{-1}\left(F_j - K_{j(j-1)}U_{(j-1)} - K_{jj}U_{k+1}\right), \\ &\quad \text{for } j = (2i+1), i = 1,3,5... \end{aligned} \qquad (2.26)$$

Substitute this to the remain equations, we finally get linear equations system:

$$SU = G, \qquad (2.27)$$

where S is defined as:

$$S = \begin{bmatrix} A_{11} & A_{12} & 0 & \cdots & 0 \\ A_{21} & A_{22} & A_{23} & \cdots & 0 \\ 0 & A_{32} & A_{33} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & A_{(p-2)(p-2)} & A_{(p-2)(p-1)} \\ 0 & \cdots & \cdots & A_{(p-1)(p-2)} & A_{(p-1)(p-1)} \end{bmatrix}, \qquad (2.28)$$

where matrices A_{ij} are given:

$$\begin{aligned} A_{ii} &= K_{(2i)(2i)} - K_{(2i)(2i-1)}K_{(2i-1)(2i-1)}^{-1}K_{(2i-1)(2i)}, \\ &\quad -K_{(2i)(2i+1)}K_{(2i+1)(2i+1)}^{-1}K_{(2i+1)(2i)}, \end{aligned} \qquad (2.29)$$

$$A_{i(i+1)} = -K_{(2i)(2i+1)} - K_{(2i+1)(2i+1)}^{-1}K_{(2i+1)(2i+2)}, \qquad (2.30)$$

$$A_{(i+1)i} = -K_{(2i)(2i-1)} - K_{(2i-1)(2i-1)}^{-1}K_{(2i-1)(2i-2)}. \qquad (2.31)$$

For $i = 1,2,...,p-1$ one has

$$\begin{aligned} G_i &= F_{2i} - K_{(2i)(2i-1)}K_{(2i-1)(2i-1)}^{-1}F_{2i-1} \\ &\quad -K_{(2i)(2i+1)}^{-1}F_{2i+1}, \end{aligned} \qquad (2.32)$$

and $U = (U_2, U_4, \cdots, U_{2(p-1)})$.

The problem of solving (2.27) is called the Schur interface problem, and the assembly and solution of sub-matrices in (2.26) can be performed parallely by different processors. The parallel implementation of the Schur conjugate gradient method can be presented in three steps:

1. Step 1: Calculate the matrix G by the parallel scheme in Figure 5;
2. Step 2: Calculate Sx by parallel scheme as in Figure 6;
3. Step 3: Solve the equation $SU = G$ by the preconditioned conjugate gradient method. The similar procedure is applied to independent subproblems. Figure 3 also shows the pseudocode to solve it numerically.

3 PARALLEL STD

3.1 *Setting of the problem*

In this section, base on the idea of dividing a large system of equations into many small ones to solve them efficiently, the parallel STD is also introduced. This method allows parallelization MPI to distribute the systems of equations to solve each subproblem. In the world of parallel computing, the MPI is the standard implementing program on multiple processors. The MPI scheme consists of

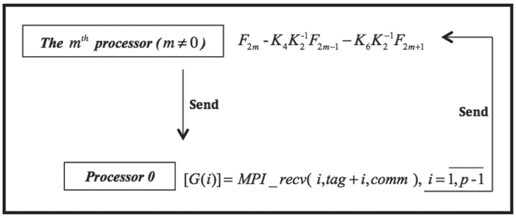

The m^{th} processor ($m \neq 0$)
$F_{2m} - K_4K_2^{-1}F_{2m-1} - K_6K_2^{-1}F_{2m+1}$
Send
Send
Processor 0
$[G(i)] = MPI\_recv(\ i, tag+i, comm\ ),\ i = \overline{1, p-1}$

Figure 5. Parallel scheme to calculate G.

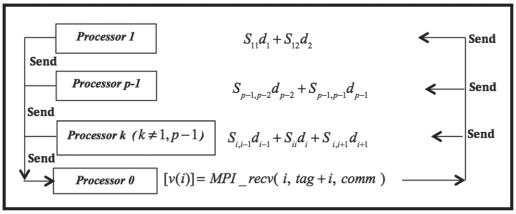

Processor 1
$S_{11}d_1 + S_{12}d_2$
Send
Send
Processor p-1
$S_{p-1,p-2}d_{p-2} + S_{p-1,p-1}d_{p-1}$
Send
Send
Processor k ($k \neq 1, p-1$)
$S_{i,i-1}d_{i-1} + S_{ii}d_i + S_{i,i+1}d_{i+1}$
Send
Send
Processor 0
$[v(i)] = MPI\_recv(\ i,\ tag+i,\ comm\ )$

Figure 6. Parallel scheme to calculate Sx.

several libraries with a set of routines to send and receive data messages (MPI_Send and MPI_Recv, respectively). MPI can be configured to execute one or several processes and run them in parallel. In addition, it is possible to implement MPI entirely within the Matlab environment, which is handled by the following algorithm.

Similar to the previous section, let us also divide domain Ω into p non-overlapping subdomains as in Figure 4. One notices that in this MPI implementation, each partition is assigned to one processor. The stiffness matrix A is rewritten as:

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & 0 & 0 \\ A_{21} & A_{22} & \cdots & 0 & 0 \\ 0 & A_{32} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & A_{(m-1)(m-1)} & A_{(m-1)m} \\ 0 & \cdots & \cdots & A_{m(m-1)} & A_{mm} \end{bmatrix}, \qquad (3.1)$$

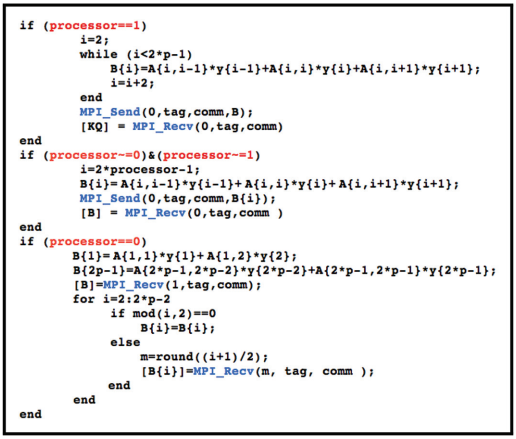
where $m = 2p-1$. Each subdomain $\Omega_3, \Omega_5, ..., \Omega_{2p-1}$ consists of k points, and they does not have any cross points as in previous section. Then, one has:

- $\Omega_1 = x_1, x_2, \cdots, x_k$,
- $\Omega_2 = x_{k+1}$,
- $\Omega_3 = x_{k+2}, x_{k+3}, \cdots, x_{2k+1}$,
- $\Omega_{2p-1} : x_{(p-1)k+p}, \cdots, x_{pk+(p-1)}$.

This yields the linear algebraic equations $A.U = B$, which can be rewritten in block form as follows:

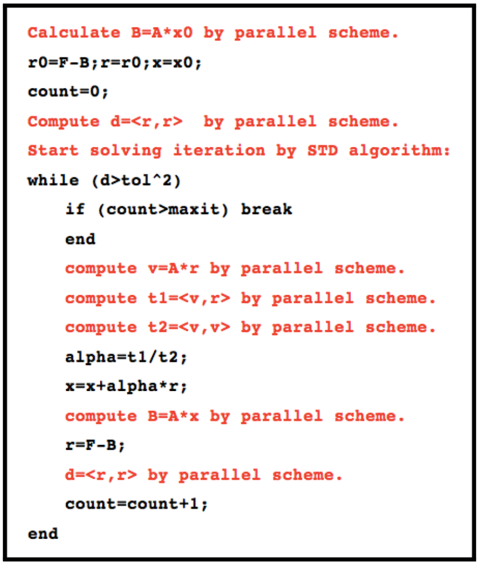
$$A.U = \begin{bmatrix} A_{11}U_1 + A_{12}U_2 \\ A_{21}U_1 + A_{22}U_2 + A_{23}U_3 \\ \cdots \\ A_{(2p-1)(2p-2)}U_{2p-2} + A_{(2p-1)(2p-1)}U_{2p-1} \end{bmatrix}, \qquad (3.2)$$

where $U = \left(U_1, U_2, \cdots, U_{2p-1}\right)$. Suppose that we have comm $= p$, that means, it shares works for p processors. Computation of quantities in the system in each subdomain can be done by parallel scheme in



```
if (processor==1)
        i=2;
        while (i<2*p-1)
            B{i}=A{i,i-1}*y{i-1}+A{i,i}*y{i}+A{i,i+1}*y{i+1};
            i=i+2;
        end
        MPI_Send(0,tag,comm,B);
        [KQ] = MPI_Recv(0,tag,comm)
end
if (processor~=0)&(processor~=1)
        i=2*processor-1;
        B{i}= A{i,i-1}*y{i-1}+ A{i,i}*y{i}+ A{i,i+1}*y{i+1};
        MPI_Send(0,tag,comm,B{i});
        [B] = MPI_Recv(0,tag,comm )
end
if (processor==0)
       B{1}= A{1,1}*y{1}+ A{1,2}*y{2};
       B{2p-1}=A{2*p-1,2*p-2}*y{2*p-2}+A{2*p-1,2*p-1}*y{2*p-1};
       [B]=MPI_Recv(1,tag,comm);
       for i=2:2*p-2
            if mod(i,2)==0
                B{i}=B{i};
            else
                m=round((i+1)/2);
                [B{i}]=MPI_Recv(m, tag, comm );
            end
       end
end
```

Figure 7. Parallel scheme to calculate AU.



```
Calculate B=A*x0 by parallel scheme.
r0=F-B;r=r0;x=x0;
count=0;
Compute d=<r,r>  by parallel scheme.
Start solving iteration by STD algorithm:
while (d>tol^2)
    if (count>maxit) break
    end
    compute v=A*r by parallel scheme.
    compute t1=<v,r> by parallel scheme.
    compute t2=<v,v> by parallel scheme.
    alpha=t1/t2;
    x=x+alpha*r;
    compute B=A*x by parallel scheme.
    r=F-B;
    d=<r,r> by parallel scheme.
    count=count+1;
end
```

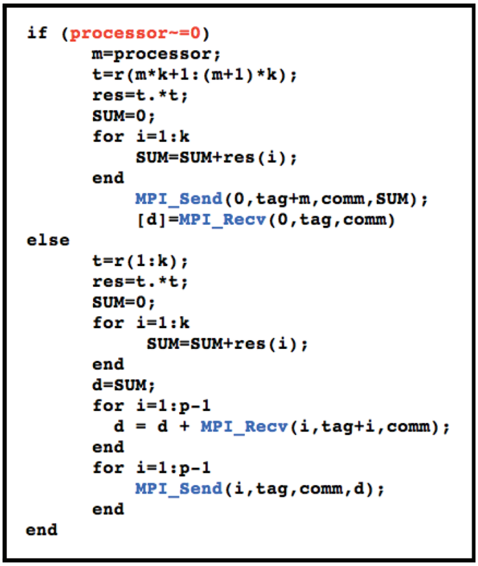
Figure 8. Parallel scheme to solve the problem.

Figure 7. It also remarks that in parallel STD, all communications were handled with MPI.

Once we get the values $B\{i\}$, the processor 0 will send to the others. And then, it is sufficient to calculate p terms in linear system (1.8) on each subdomain by parallel strategy.

3.2. *Solve the problem by STD algorithm*

The parallel STD algorithm is then given in Figure 8 using the Matlab codes, where the parallel algorithm to calculate the inner product $d = \langle r, r \rangle$ is presented in Figure 9.



```
if (processor~=0)
      m=processor;
      t=r(m*k+1:(m+1)*k);
      res=t.*t;
      SUM=0;
      for i=1:k
          SUM=SUM+res(i);
      end
          MPI_Send(0,tag+m,comm,SUM);
          [d]=MPI_Recv(0,tag,comm)
else
      t=r(1:k);
      res=t.*t;
      SUM=0;
      for i=1:k
           SUM=SUM+res(i);
      end
      d=SUM;
      for i=1:p-1
        d = d + MPI_Recv(i,tag+i,comm);
      end
      for i=1:p-1
          MPI_Send(i,tag,comm,d);
      end
end
```

Figure 9. Parallel scheme to calculate $d = \langle r, r \rangle$.

4 NUMERICAL EXAMPLES

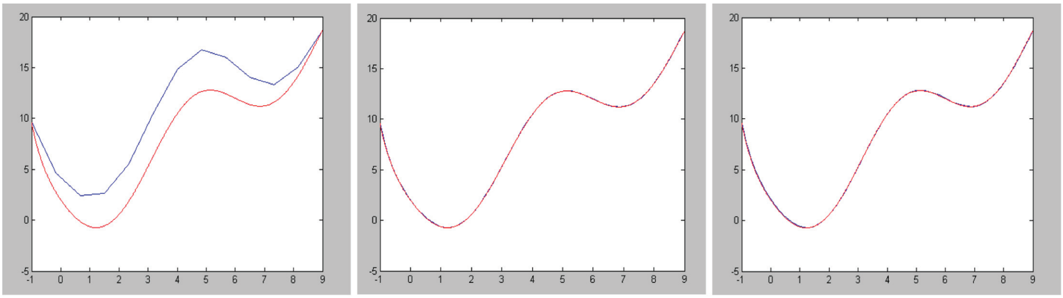
In this section, some numerical examples are provided in order to demonstrate a good performance of the proposed method. It is also important that such results are inherently valid for parallel computing. We survey our recent research on the parallel solvers to one-dimensional problem (1.4). In future the problem in higher dimensions could be considered and we will focus on a generic parallel implementation framework of thousands of processors.

4.1 *Example 1*

Let us consider the function

$$u(x) = e^{-2x} - 4\sin\left(\frac{\pi}{3}x\right) + \cos\left(\frac{\pi}{4}x\right) + 2x, \qquad (4.1)$$

the analytical solution to the equation (1.4). Then, the Schur interface by conjugate gradient algorithm and the parallel STD scheme are applied to get approximate solution to this problem. The numerical results have been tested for the specified error tolerance tol = 10^{-6} and the maxit = 10^{2} the number of iterations. The approximate solutions by the Schur complement conjugate gradient method, where the spatial domain of the problem is decomposed into p = 2, 3 subdomains. Otherwise, the parallel STD scheme is also applied together with p processors ($p > 2$). Some numerical results represented in Figure 10 demonstrate the effective use of proposed Schur conjugate gradient method rather than parallel STD algorithm. One can see that the Schur conjugate gradient method gives good convergent solution (is very closed to the exact solution) with both cases p =2 and p = 3 within two first iterations, meanwhile the STD does not give the convergence after maxit iterations.

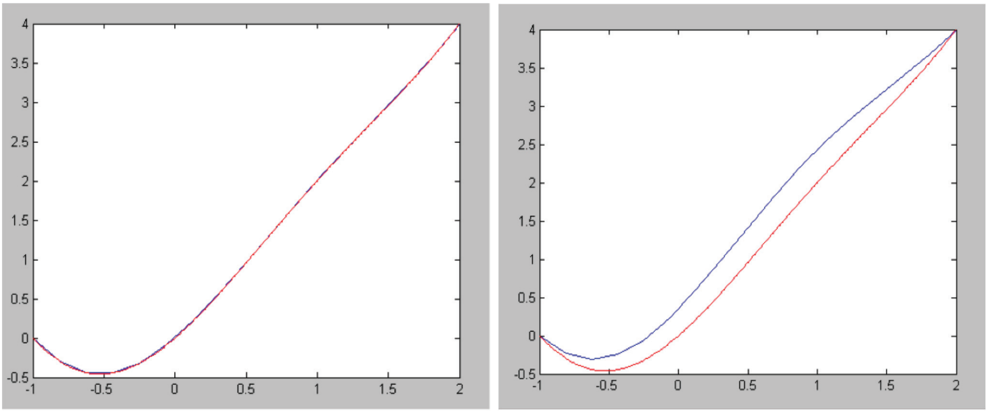


(a) For $p = 3$ and $k = 3$.

(b) For $p = 3$ and $k = 3$.

(c) For $p = 2$ and $k = 20$.

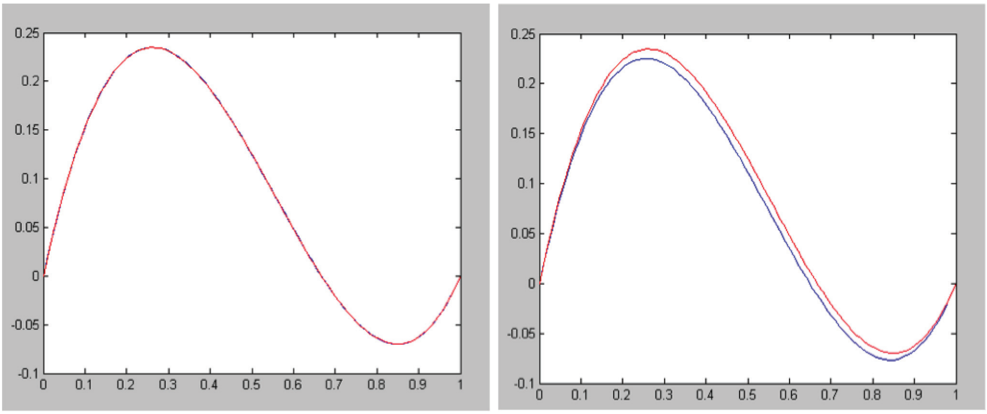
Figure 10. Numerical solution to Example 1, for different cases of p and k. (a): By STD algorithm, (b) and (c): By Schur conjugate gradient method. Red: the exact solution. Blue: The approximate solution.



(a) Conjugate Gradient method.

(b) STD algorithm.

Figure 11. Numerical solution to Example 2, for case $p = 4$ and $k = 3$. (a) By Conjugate Gradient method, (b) By STD algorithm. Red: the exact solution. Blue: The approximate solution.



(a) Conjugate Gradient method.

(b) STD algorithm.

Figure 12. Numerical solution for case $p = 3$ and $k = 30$. (a) By conjugate gradient method, (b) By STD algorithm. Red: the exact solution. Blue: The approximate solution.

4.2 *Example 2*

Let us consider the function:

$$u(x) = \sin\left(\frac{\pi}{2}x\right) + x^2, \tag{4.2}$$

is the exact solution to (1.4). The numerical results have been implemented also for the tolerance tol = 10^{-6} and the iterations maxit = 200. The approximate solution by conjugate gradient method and the parallel STD implementation can be represented in Figure 11. One remarks that in this case, four processors have been used, that is, $p = 4$.

As in the Example 1, the numerical behaviour of parallel Schur conjugate gradient gives convergent solution in only two first iterations. Nevertheless, in Figures 10 and 11, we give numerical evidence that by the STD algorithm, the convergence is still not achieved after maxit = 200 iterations. Let us consider additional example, where the exact solution $u(x) = 3x^3 - 5x^2 + 2x$ to (1.4), the numerical simulation of both parallel schemes are also presented in Figure 12 to give a comparison.

5 CONCLUSION

In this study, we have presented the non-overlapping Schur complement method, in which the linear system is calculated using parallel conjugate gradient method. The basic idea is to split a large system of equations into smaller systems that can be solved independently in paralleled processors. The parallel STD algorithm is also described to give comparable results with the proposed method. Some numerical experiments have been performed to show a good parallel efficiency and convergence of our method. At this point, it should be noted that in our computational program, MPI libraries can be called under Matlab environment. From this result, one can recognize that this is a promising method that could be well adapted to solve a large sparse matrix systems under the parallel implementation. Furthermore, it can be seen that the computational time is minimal and required memory is optimal when subdomains are used. The method also works well in more general settings of problems in many applications. Nevertheless, a lot of works are still open from this study. Some other important topics in the field of domain decomposition method development and applications to higher dimensional problems will be analyzed in future research.

REFERENCES

Kocak, S., H.A. (2010). Parallel schur complement method for large-scale systems on distributed memory computers. *Applied Mathematical Modelling 25*(10), 873–886.

Mansfield, L. (1990). On the conjugate gradient solution of the Schur complement system obtained from domain decomposition. *SIAM journal os Numerical Analysis 27*(6), 1612–1620.

Meyer, A. (1990). A parallel preconditioned conjugate gradient method using domain decomposition and inexact solvers on each subdomain. *Computing 45*(3), 217–234.

Milyukova, O.Y. (2001). Parallel approximate factorization method for solving discrete elliptic equations. *Parallel Computing 27*(10), 1365–1379.

Smith, B., P. Bjorstad, W.G. (1996). *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*. Cambridge: Cambridge University Press.

This page intentionally left blank

Statistics and applied statistics

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Determining results of the Phadiatop test using logistic regression and contingency tables based on a patient's anamnesis

P. Kuráňová
VŠB—Technical University of Ostrava, Ostrava, Czech Republic

ABSTRACT: This article models the Phadiatop test. This study was created with support from the Clinic of Occupational and Preventive Medicine to try to avoid unnecessary and costly testing. This estimation used statistical methods, specifically logistic regression to predict a patient into a particular group, and contingency tables to verify other dependences between the patient's other characteristics. Patients were categorized only on the basis of their personal and family anamnesis, age and sex. Patients were put into the correct group (healthy or sick) with 64% probability. Also a testing based on age groups of the patients was done using this database. The presence of the positive Phadiatop test was the most common for people born between 1972 and 1981, where the genetic predispositions for a positive Phadiatop test results are about 55%.

1 INTRODUCTION

The knowledge of the results of the Phadiatop test is very important especially for diagnosis of allergic dermatitis and also for the professional medical care for travellers (Hajduková et al. 2005, 2009, Williamas et al. 2001). The Phadiatop test is used as a measure of atopy. The atopy rate of inhabitants of the Czech Republic is increasing. Atopy could be understood as a personal or family predisposition to become, mostly in childhood or adolescence, hyper-sensible to normal exposure of allergens, usually proteins. These individuals are more sensitive to typical symptoms of asthma, eczema, etc.

According to disease severity, results of the Phadiatop test are divided into the six following groups: Groups 0 and I indicate none or weak form of atopy and the remaining groups (II, III, IV, V, and VI.) indicate increasing severe forms of atopic symptoms. Unfortunately, the Phadiatop test is expensive, so we try to predict the results of the test on the basis of a detailed family and personal anamnesis (Wüthrich et al. 1995, 1996, Sigurs 2010).

Information obtained from personal and family anamneses of each patient were used for detecting the presence of asthma, allergic rhinitis, eczema or other forms of allergy (contact allergy, food, etc.). Family and personal anamnesis of each patient were evaluated by medical expert. Furthermore, other characteristics were available for each patient: age and sex. Then, we created and verified a mathematical model for the accurate classification of patients into one of two groups of the Phadiatop test.

In this paper we discuss the logistic regression approaches for obtaining the results of Phadiatop test based only on family, personal anamnesis and other characteristics. Besides, it also examines the mutual relationship between a positive Phadiatop test result, sex and age of the patient. In this paper we deal with the prediction of each patient into one of two groups of the Phadiatop test based logistic regression (Kuráňová & Hajduková 2013, Briš et al. 2015). Next, we describe the connections of genetic predispositions for the atopy according to the age group of inhabitants. Database of patients comes from 2010–2012.

2 THE USED OF METHODS

2.1 *Logistic regression as a toll for discrimitation*

The logistic regression was not originally created for the purpose of discrimination, but it can be successfully applied for this kind of analysis (Hosmer, & Lemeshow 2004, Menard 2009, Miner 2009). A logistic regression model, which is modified for the purpose of discrimination, is defined as follows. Let $Y_1,\dots,\ Y_n$ is a sequence of independent random variables with alternative distributions, whose parameters satisfy:

$$\begin{aligned} P\left(Y_i = 1 \middle| X_i = x_i\right) &= \frac{e^{\beta_0+\beta' x}}{e^{\beta_0+\beta' x}+1}, \\ P\left(Y_i = 0 \middle| X_i = x_i\right) &= \frac{1}{e^{\beta_0+\beta' x}+1}, \end{aligned} \tag{1}$$

for $i = 1, \ldots, n$ where $\beta' = (\beta_1, \ldots, \beta_n)'$, is unknown p-dimensional parameter and $X_1, \ldots, X_n$, are (p+1) - dimensional random vectors $(\beta_1, \ldots, \beta_n)$. This model can be called a learning phase, in which both values X_i and Y_i are known for each object (i.e. it is known to which group each object belongs to). Based on this knowledge, we try to predict parameters $\beta_1,\ldots,\beta_n$ and thus we try to estimate function $\pi(x)$, where

$$\pi(x) = P(Y = 1 | X = x) = \frac{e^{\beta_0 + \beta' x}}{e^{\beta_0 + \beta' x} + 1}. \tag{2}$$

Another object for which the classification is unknown is assigned to one of two groups according to the value of decision function $\pi(x)$.

The object will be included in the first group if $\pi(x)>0.5$. Otherwise, the object will be included in the second group. The main advantage of this model is that it does not require conditions for distributions of random vectors $X_1, \ldots, X_n$. However, the model assumes a very specific form of probability $P(Y=1|X=x)$ and we should verify the significance of the relationship

$$\pi(x) = P(Y = 1 | X = x) = \frac{e^{\beta_0 + \beta' x}}{e^{\beta_0 + \beta' x} + 1} \tag{3}$$

2.2 *Variable dependence analysis in a contingency table for 3 variables*

Independence of two dichotomous variables contingent on another categorical variable can be tested by using Cohran or Mantel-Haenszel statistics. In the case of Cohran statistics, we used the formula

$$Q_C = \frac{\left(\sum_{l=1}^{L} n_{l11} - \sum_{l=1}^{L} m_{l11}\right)^2}{\sum_{l=1}^{L} v_{l11}}, \tag{4}$$

where L represents the number of categories of a variable, the total number of subjects included in the *lth* table (l = 1,2, …, L) as n_l, joint frequency as n_{lij}, row marginal frequency as n_{li+} and column marginal frequency as n_{l+j}. If a null hypothesis about the dichotomous variable independence is true, the expected frequency (average frequency) in the cell in the *lth* table, *ith* row and *jth* column is given by the relation

$$m_{lij} = \frac{n_{li+} n_{l+j}}{n_l}. \tag{5}$$

And the variance of this frequency by the relation

$$v_{lij} = \frac{n_{l1+} n_{l2+} n_{l+1} n_{l2+2}}{n_l^3}. \tag{6}$$

To determine the rate of association, it is possible to use Mantel-Haenszel common odds ratio estimate for L fourfold tables, which is given by the relation

$$\psi_{MH} = \frac{\sum_{l=1}^{L} \frac{n_{l11} n_{l22}}{n_l}}{\sum_{l=1}^{L} \frac{n_{l12} n_{l21}}{n_l}}. \tag{7}$$

This rate takes the value 1 in case of independence; the independence testing is based on the natural logarithm of the calculated value (Agresti 2003, Simonoff, 2003).

3 OBTAINED RESULTS

Our database comes from the University Hospital of Ostrava, the Department of Work and preventive medicine. The database includes a total of 1,132 records of patients who underwent the Phadiatop test examination. For the purposes of our comparison, we consider only the patients in the control group, who filled the records completely (personal anamnesis, family anamnesis, gender and year of birth). The control group is a group of patients who were on preventive examinations without previously known diseases or travellers. The number of complete records is 274.

The database contained these pieces of information about individual patients: the Phadiatop test result Group 0 have Phadiatop test 0 or I (no visible symptoms), so no treatment was necessary. The remaining patients with Phadiatop test II–VI are members of Group 1. Medical treatment is necessary for these patients.

Logistic regression does not have any requirements for the data arrangement, but we need a specific format of the data for the logistic regression. For this particular case we have one dependent variable Y, Phadiatop (*Ph*), which depends on two independent variables of Personal Anamnesis (*PA*) and Family Anamnesis (*FA*). Variable Y can be 0 or 1, according to the membership of a patient to Group 0 or Group 1, respectively. The illnesses which, according to doctors, influence the Phadiatop test result the most were established as independent variables. These are asthma, allergic rhinitis, eczema and others. The category “Others” represents the score of various kinds of allergies (food allergies, etc). Each patient’s family and personal anamnesis was examined for all these illnesses.

For testing purposes, we define a dependent variable Phadiatop test result and a total of 10 independent variables (4 variables of personal anamnesis, 4 for family anamnesis, year of birth and gender). An example of database is shown in Table 1.

3.1 *Results obtained by Logistic regression*

For testing using logistic regression, we thought of all 10 independent variables and one dependent variable Phadiatop test, coded as 0 for healthy patients (test result 0 and I) and 1 for patients with a disease (the result of testing II to VI). Results of statistical significance of the individual independent variables are given in Table 2.

On the basis of Wald's test and test of statistical significance, we see that a statistically significant variable in this case appears only PA_allerdic rhitis, PA_ekzema and PA_Asthma. Other variables are statistically insignificant and could be excluded from the model.

The predictive qualities of the logistic model are shown in Table 3. Here, it is obvious that the model predicts better into group 0, thus, a group of

Table 1. Evaluation and verification of the independent variables for Logistic regression.

Number of patient	*5*	*16*	*35*	*40*	*41*
Sex	m	m	m	m	w
Year of birth	1973	1970	1974	1986	1991
PA_asthma	0	1	1	0	0
PA_allergic rhinitis	1	1	0	0	0
PA_eczema	1	0	0	0	0
PA_others	0	0	0	1	0
FA_asthma	0	1	0	0	0
FA_allergic rhinitis	1	0	0	0	0
FA_eczema	1	0	0	0	0
FA_others	1	0	0	1	0

Table 2. Evaluation and verification of the independent variables for Logistic regression.

Independent variables	Estimate	Wald's	Significant
PA_asthma	0.428	2.480	0.115
PA_allergic rhinitis	1.111	15.780	0.000
PA_eczema	0.993	8.774	0.003
PA_others	0.422	2.338	0.126
FA_asthma	0.351	1.373	0.241
FA_allergic rhinitis	–0.208	0.398	0.528
FA_eczema	–0.411	1.429	0.232
FA_others	0.070	0.043	0.836
Sex	0.006	0.000	0.984
Year of birth	–0.002	0.018	0.895
Constant	1.651	0.005	0.943

Table 3. Classification table using by Logistic regression.

	Predicted value		
Observed value	Phadiatop 0	Phadiatop 1	Percentage correct
Phadiatop 0	136	31	81.4
Phadiatop 1	68	39	36.4
			63.9

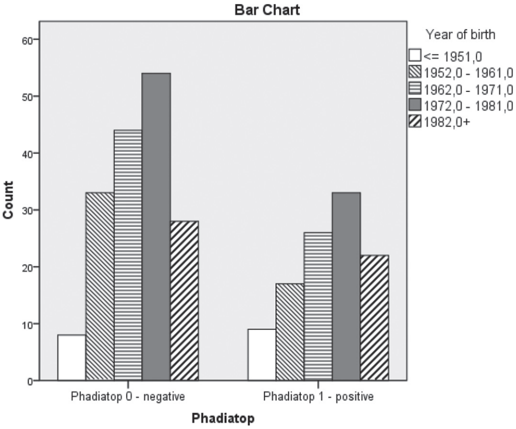


Figure 1. Depiction of age groups contingent on the Phadiatop test result.

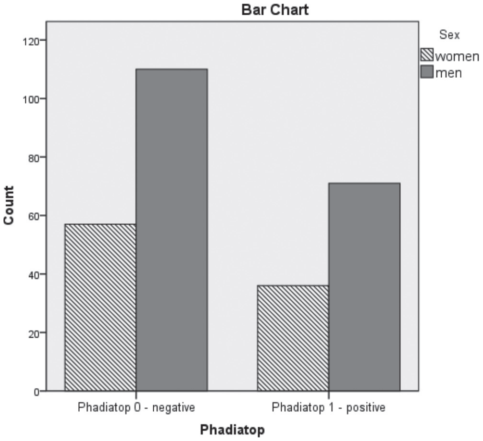


Figure 2. Depiction of patients' sex contingent on the Phadiatop test result.

healthy patients. The total value of the prediction model is 63.9%.

3.2 *In depth evaluation of the Phadiatop test dependence on the sex and age variables*

Based on the previous testing where several variables proved to be statistically insignificant, it seems necessary to analyse the sex and age variables separately. The variables regarding a patient's anamneses are Personal Anamneses (PA) and Family Anamneses (FA). The other variables, Age and Sex, are variables which are not related to a patient's anamnesis; therefore, their further analysis in relation to the Phadiatop test result is important.

Table 4. Phadiatop test results by sex and year of birth of the patients.

			Phadiatop		
Year of birth			Negative	Positive	Total
<= 1951	Sex	Woman	3	4	7
		Man	5	5	10
	Total		8	9	17
1952–1961	Sex	Woman	9	6	15
		Man	24	11	35
	Total		33	17	50
1962–1971	Sex	Woman	18	11	29
		Man	26	15	41
	Total		44	26	70
1972–1981	Sex	Woman	18	7	25
		Man	36	26	62
	Total		54	33	87
1982 +	Sex	Woman	9	8	17
		Man	19	14	33
	Total		28	22	50

The age of the patients ranges from 17 to 69 years of age (year of birth 1943–1995). For the purpose of a clearer analysis, we will divide the age variable into 4 groups. The representation of patients in each of the groups by sex and age is stated in Figures 1 and 2. We are examining dependence for groups of patients who differ in age. Table 4 contains the overall summary of the patients in the groups. Based on the stated data, we can see that the number of patients is larger in some groups than in others.

Based on the data stated in Table 4, we can see that the most examined patients were born in the years 1972 to 1981, 87 in total. Next is the group of patients born between the years 1962 and 1971, the total of 70 patients who underwent the examination. The most patients with a positive Phadiatop test result are in the age group 1972–1981, 33 patients in total, of which 26 are men.

3.3 *Resulting obtained by testing contingency tables*

Based on the results stated in Table 4, we can compare three groups of results based on the Pearson

Table 5. Results of the Chi-Square Test by patient year of birth groups.

Year of birth		Value	df	Asymp. Sig. (2-sided)
<= 1951	Pearson Chi-Square	0.084	1	0.772
	Continuity Correction	0.000	1	1.000
	Likelihood Ratio	0.084	1	0.771
	N of Valid Cases	17		
1952–1961	Pearson Chi-Square	0.344	1	0.558
	Continuity Correction	0.068	1	0.794
	Likelihood Ratio	0.339	1	0.560
	N of Valid Cases	50		
1962–1971	Pearson Chi-Square	0.013	1	0.909
	Continuity Correction	0.000	1	1.000
	Likelihood Ratio	0.013	1	0.909
	N of Valid Cases	70		
1972–1981	Pearson Chi-Square	1.470	1	0.225
	Continuity Correction	0.937	1	0.333
	Likelihood Ratio	1.510	1	0.219
	N of Valid Cases	87		
1982 +	Pearson Chi-Square	0.098	1	0.754
	Continuity Correction	0.000	1	0.990
	Likelihood Ratio	0.098	1	0.755
	N of Valid Cases	50		

Table 6. Conditional dependence tests.

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	0.020	1	0.886
Mantel-Haenszel	0.000	1	0.992

Chi-Square statistics, Table 5. If we are examining dependence for groups of patients by their year of birth, we can see from the stated values that in any of the groups we do not reject the null hypothesis on independence of the Phadiatop test result on sex at a 5% level of significance.

To determine a more accurate result of the Phadiatop test testing contingent on a patient's sex, depending on whether we consider a patient's age as well, we will perform a conditional dependence test by means of Cohran and Mantel-Haenszel statistics, Table 6.

Based on the performed tests, we do not reject the null hypothesis on independence of the Phadiatop result on sex at a 5% level of significance.

According to the above stated calculated statistics it was verified that a patient's sex and age have no influence on the Phadiatop test result. The illnesses influencing the Phadiatop test result are Asthma, Allergic rhinitis and Eczema in a patient's personal anamnesis.

4 THE PHADIATOP TEST RESULTS EVALUATION BASED ON THE AGE AND GENETICAL PREDISPOSITIONS

Based on the above stated information about a patient's age and categorization of patients into age groups, we will perform the search for genetic predispositions to a positive Phadiatop test. The positive (II to VI) and the negative (0 and I) Phadiatop test results are stated.

Based on the mentioned age groups and the positive and negative Phadiatop tests division it is clear that a proportional representation of the diseased patients is more or less equal. The only group that differs and has the most positive patients is the age group born 1981–1972, thus, the young patients. According to the available information, the genetic predisposition for atopy (positive Phadiatop test II to VI) should be about 30% for the inhabitants of the Czech Republic. For this analysis, a patient with genetic predispositions was that one who filled in the statistically significant positive diseases of the Phadiatop test results into his family anamnesis. Thus, a patient with genetic predispositions included into his/her family anamnesis:

Table 7. Age proportional representation of the inhabitants based on the Phadiatop test.

	Phadiatop		Predisposition		
Year of birth	0 + I	II to VI	Genetic	For the positive	Ratio of the positive
To 1982	28	22	26	11	0.21
1981 to 1972	54	33	45	18	0.31
1962 to 1971	44	26	27	11	0.24
1952 to 1961	33	17	23	9	0.16
Before 1952	8	9	9	4	0.08
Sum	167	107	130	53	1

- Positive asthma or allergic rhinitis and other records then did not need to be taken into account
- Positive eczema and other diseases at once, if the asthma and rhinitis were negative.

Based on Table 7, it is obvious that the most patients with a genetic predisposition are from the age group born between 1981 and 1972, 45 patients out of the total of 130 patients who were found to have a genetic predisposition to a positive result of the Phadiatop test, Table 7 (the column "Predisposition—Genetic"). The proportional representation of the patients with genetic predisposition is (130/274 = 0.47).

This proves the division of the data file, where there are 107 positive patients out of 274, about 39%. Based on the family predispositions, 47% of all patients should be in the positive Phadiatop test group. Nevertheless, there are (53/107 = 0.49) genetically predisposed patients with the positive Phadiatop test (107 records).

5 CONCLUSION

Knowledge of the Phadiatop test is very significant, for both patient examination in offices of occupational and preventive medicine and for correct patient care (e.g. Travellers). Unfortunately performing the Phadiatop test is expensive; therefore, there is an effort to model its result as accurately as possible using characteristics that are easy to discover, such as a patient's personal and family anamnesis, and also e.g. age, sex etc.

The tested database came from the years 2010–2012 form the University Hospital of Ostrava; there were 274 patient entries available for the so called control group of patients, i.e. patients who have no specific illness, and the test is performed preventively (e.g. Travellers etc.). The testing used logistic regression; on the basis of the performed tests, the characteristics influencing the Phadiatop test result were identified to be Asthma, Allergic rhinitis and Eczema in a patient's personal anamnesis. Family anamnesis proved to be statistically insignificant. Characteristics which are not related to any illnesses, in our case a patient's age and sex, were tested separately by means of contingency tables. Even here, it was confirmed that these characteristics do not influence the Phadiatop test result in any way. A model constructed on the basis of logistic regression categorizes a patient into the correct group (healthy or sick) with 63.9% reliability. This means that every fourth or third patient is classified incorrectly. The model works better for patients who belong to Group 0, healthy patients, where the model predicts with 81% reliability.

Another interesting result is for testing for genetic predispositions to a positive Phadiatop test. The group which was the most predisposed to a positive Phadiatop test was the one for the year of birth from 1981 to 1972; here 31% of patients have a positive test result. In contrast, patients born before 1961 have a lower genetic predisposition. Our conclusion proves the presumption that about 30% of population has the genetic predisposition for the positive Phadiatop test. Based on our calculation, the proportional representation is about 47%.

ACKNOWLEDGMENT

This paper was done thanks to cooperation with The University Hospital of Ostrava, the Department of Clinic of Occupation and Preventive medicine.

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science—LQ1602" and Operational Programme Education for Competitiveness—Project No. CZ.1.07/2.3.00/20.0296.

REFERENCES

Agresti, A. (2003). Logit models for multinomial responses. *Categorical Data Analysis, Second Edition*, 267–313.

Bris, R., Majernik, J., Pancerz, K., & Zaitseva, E. (2015). Applications of Computational Intelligence in Biomedical Technology.

Hajduková, Z., Pólová, J., & Kosek, V. (2005). The importance of Atopy Investigation in the Department of Travel Medicine. *ALERGIE-PRAHA-*, *7*(2), 109.

Hajduková, Z., Vantuchová, Y., Klimková, P., Makhoul, M., & Hromádka, R. (2009). Atopy in patients with allergic contact dermatitis. *Journal of Czech Physicians: Occupational therapy*, (2), 69–73.

Hosmer Jr, D.W., & Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons.

Kuráňová, P., & Hajduková, Z. (2013, May). The use of logistic and ordinal regression for the prediction of the phadiatop test results. In *Digital Technologies (DT), 2013 International Conference on* (pp. 111–115). IEEE.

Menard, S. (2009). *Logistic regression: From introductory to advanced concepts and applications.* Sage Publications.

Miner, G., Nisbet, R., & Elder IV, J. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press.

Sigurs, N., Aljassim, F., Kjellman, B., Robinson, P.D., Sigurbergsson, F., Bjarnason, R., & Gustafsson, P.M. (2010). Asthma and allergy patterns over 18 years after severe RSV bronchiolitis in the first year of life. *Thorax*, *65*(12), 1045–1052.

Simonoff, J.S. (2003). *Analyzing Categorical Data*. Springer Science & Business Media.

Williams, P.B., Siegel, C., & Portnoy, J. (2001). Efficacy of a single diagnostic test for sensitization to common inhalant allergens. *Annals of Allergy, Asthma & Immunology*, *86*(2), 196–202.

Wüthrich, B., Schindler, C., Leuenberger, P., & Ackermann-Liebrich, U. (1995). Prevalence of atopy and pollinosis in the adult population of Switzerland (SAPALDIA study). *International archives of allergy and immunology*, *106*(2), 149–156.

Wüthrich, B., Schindler, C., Medici, T.C., Zellweger, J.P., Leuenberger, P.H., & Team, S. (1996). IgE levels, atopy markers and hay fever in relation to age, sex and smoking status in a normal adult Swiss population. *Internationalarchives of allergy and immunology*, *111*(4), 396–402.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

On the distortion risk measure using copulas

S. Ly
Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

U.H. Pham
Faculty of Economic Mathematics, University of Economics and Laws, Ho Chi Minh City, Vietnam

R. Briš
VSB—Technical University of Ostrava, Ostrava, Czech Republic

ABSTRACT: Distortion risk measure is a very effective tool for quantifying losses in finance and insurance while copulas play an important role in modeling dependence structure of random vectors. In this paper, we propose a new method to estimate distortion risk measures and use copulas to find the distribution of a linear combination of two dependent continuous random variables. As a result, partial risks as well as aggregate risk are definitely estimated via distortion risk measure using copulas approach.

1 INTRODUCTION

Suppose that we have a portfolio Y consisting of two assets X_1 and X_2 as follows:

$$Y = w_1 X_1 + w_2 X_2, \tag{1}$$

where, w_i denotes the weight of asset $i, i = 1,2$.

Let F_{X_1}, F_{X_2} and F_Y be distribution functions of X_1, X_2 and Y, respectively, where X_1 and X_2 are not independent. Here, our goal is to calculate the risk of the portfolio Y under distortion risk measure, see Wang (2000), given by

$$R_g[Y] = \int_0^{\infty} g\left(\overline{F}_Y(y)\right)dy + \int_{-\infty}^{0}\left[g\left(\overline{F}_Y(y)\right) - 1\right]dy, \tag{2}$$

where, g is a distortion function and $\overline{F}_Y(y) = 1 - F_Y(y)$ is a survival function of Y.

The risk measure R_g is formed using Choquet integral, see Wang (2000). In some cases, Y denotes non-negative loss, then the distortion risk measure only has the first part in (2). As we can see, the important thing is that we have to derive the distribution of Y.

Recall that if $Y = X_1 + X_2$ and X_1, X_2 are independent, then it is well-known that the solution can be solved through convolution product of two density functions f_{X_1} and f_{X_2} , given by

$$f_Y(y) = f_{X_1} * f_{X_2}(y) = \int_{-\infty}^{\infty} f_{X_1}(x) f_{X_2}(y - x)dx. \tag{3}$$

In Cherubini et al. (2011), the authors consider the case X_1 and X_2 are not independent. In their approach, copula is used to define a C-convolution given by

$$\begin{aligned} F_{X_1+X_2}(y) &= F_{X_1} \overset{C}{*} F_{X_2}(y) \\ &= \int_0^1 \frac{\partial}{\partial u} C\left(u, F_{X_2}\left(y - F_{X_1}^{-1}(u)\right)\right)du, \end{aligned} \tag{4}$$

where, C is a copula capturing dependence structure of X_1 and X_2.

In this article, we consider a more general case in the sense that using copula to find the distribution of the porfolio Y as a combination of two continuous variables. After that, we will conduct an estimation for the risk of Y using distortion risk measure.

The paper is organized as follows. The introduction is presented in section 1. The preliminaries about distortion risk measures and copulas are briefly recalled in section 2 and section 3. After that, in section 4, we propose a new formula for estimating the risk. In section 5, a copula-based method for finding the distribution of a linear combination of random variables is established. Next, we show the applications in section 6 and the conclusions are stated in the last section.

2 DISTORTION RISK MEASURE

Suppose Y is a non-negative loss random variable with distribution function F_Y. Then, it is well-known that the expectation of Y can be written in the form:

$$E(Y)=\int_0^\infty (1-F_Y(y))dy.$$

However, the expectation is not used as a risk measure. Instead of using this quantity, ones prefer to transform it with a function g leading to distortion risk measure defined as

$$R_g[Y]=\int_0^\infty g(1-F_Y(y))dy, \tag{5}$$

where $g:[0;1]\to[0;1]$, such that $g(0)=0, g(1)=1$ and g is a non-decreasing function. Such, g is called *distortion function*.

A number of risks measures found in finance and insurance literature are special cases of the distortion risk measure, see Sereda et al. (2010).

i. VaR: $g_\alpha(u)=1_{\{1-\alpha;1\}}(u)$, for some $\alpha\in(0;1)$.
ii. ES (TVaR): $g_\alpha(u)=\min\{1,\frac{u}{1-\alpha}\}$, $\alpha\in(0;1)$.
iii. Proportional hazard transform: $g_\beta(u)=u^{1/\beta}$, for some $\beta>1$.
iv. Wang's transform: $g_\gamma(u)=\Phi(\Phi^{-1}(u)+\gamma)$. where, Φ is a standard normal distribution function and γ is often chosed by $\gamma=\Phi^{-1}(\alpha), 0<\alpha<1$.

For more class of distortion functions, one can see in Wang (1996).

3 COPULAS AND MEASURES OF DEPENDENCE

Let $I=[0;1]$ be the closed unit interval and $I^2=[0;1]\times[0;1]$ be the closed unit square.

Definition 1. ***(Copula)*** *A 2-copula (two dimensional copula) is a function C: $I^2\to I$ satisfying the conditions:*

i. $C(u,0)=C(v,0)=0$, *for any* $u,v\in I$.
ii. $C(u,1)=u$ *and* $C(1,v)=v$, *for any* $u,v\in I$.
iii. *For any* $u_1,u_2,v_1,v_2\in I$ *such that* $u_1\le u_2$ *and* $v_1\le v_2$,

$$C(u_2,v_2)-C(u_2,v_1)-C(u_1,v_2)+C(u_1,v_1)\ge 0.$$

The most important role in copula theory is from Sklar's theorem (1959). In fact, let X_1 and X_2 be random variables with continuous marginal distribution functions F_{X_1} and F_{X_2}, respectively, and a joint distribution function H, then by Sklar's theorem, see Nelsen (2006), there exists a unique copula C such that

$$H(x_1,x_2)=C\left(F_{X_1}(x),F_{X_2}(x_2)\right). \tag{6}$$

This copula C captures the dependent structure of X_1 and X_2 In particular, X_1 and X_2 are independent if and only if $C(u,v)=(u,v)=uv$; While X_1 and X_2 are comonotonic (i.e. $X_2=f(X_1)$ a.s., where f is strictly increasing) if and only if $C(u,v)=M(u,v)=\min(u,v)$ and X_1 and X_2 are countermonotonic (i.e. $X_2=f(X_1)$ a.s., where f is strictly decreasing) if and only if $C(u,v)=W(u,v)=\max(u+v-1,0)$.

Note (see Nelsen (2006)): for any copula C and for any $(u,v)\in I^2$, we have the bound property for copula:

$$W(u,v)\le C(u,v)\le M(u,v). \tag{7}$$

Since a copula can model the dependence structure of random variables, one can construct measures of dependence using copulas with suitable metrics. In fact, some well-known measures can be written in terms of copula, see Nelsen (2006).

The Kendall's $\tau(X_1,X_2)$, or $\tau(C)$ is

$$\tau(C)=4\iint_{I_2} C(u,v)dC(u,v)^{-1} \tag{8}$$

The Spearman's $\rho(X_1,X_2)$, or $\rho(C)$ is

$$\rho(c)=12\iint_{I^2} C(u,v)dudv-3. \tag{9}$$

The upper and lower tail dependence are

$$\lambda_U(C)=\lim_{t\to 1^-}\frac{1-C(t,t)}{1-t}, \tag{10}$$

$$\lambda_L(C)=\lim_{t\to 0^+}\frac{C(t,t)}{t}. \tag{11}$$

In Tran et al. (2015), we also proposed a new non-parametric measure of dependence for two continuous random variables X_1 and X_1 with copula C, is defined by

$$\lambda(C)=| C\|_S^2 -2\| C-M\|_S^2, \tag{12}$$

where, $\| C\|_S$ denotes a modified Sobolev norm for copula C, given by

$$\| C\|_S=\left(\iint_{I^2}\left[\frac{\partial C^2(u,v)}{\partial u}+\frac{\partial C^2(u,v)}{\partial v}\right]dudv\right)^{1/2}, \tag{13}$$

The measure $\lambda(C)$ could be used as a measure of monotone dependence because it attains its extreme values of 1 (or −1) if and only if X and Y are monotonic (or countermonotonic).

4 ESTIMATION OF DISTORTION RISK MEASURE

In this section, we are going to establish an expression for approximate distortion risk measure given by (5). Notice that we only consider Y as a nonnegative loss variable. It is because for $Y < 0$, we can definitely plus a constant number m (large enough) such that $Y + m = Y' \geq 0$. Then, the distortion risk measure $R_g[Y] = R_g[Y'] - m$.

Back then, to deal with an integral over infinite intervals, we firstly change variable to get a finite interval. In particular, one can take $y = \frac{t}{1-t}$ and the risk $R_g[Y]$ becomes

$$R_g[Y] = \int_0^1 g\left(1 - F_Y\left(\frac{t}{1-t}\right)\right)\frac{1}{(1-t)^2}dt.$$

Let $k(t) = g\left(1 - F_Y\left(\frac{t}{1-t}\right)\right)\frac{1}{(1-t)^2}$ and apply the composite trapezoidal rule, we have an approximation:

$$R_g[Y] = \int_0^1 k(t)dt \approx \frac{1}{n}\left(\frac{k(0)}{2} + \frac{k(1)}{2} + \sum_{i=1}^{n-1} k\left(\frac{i}{n}\right)\right).$$

It is straightforward to check that

$$\begin{aligned}
&k(0) = g(1) = 1,\\
&k(1) = \lim_{t\to 1} g\left(1 - F_Y\left(\frac{t}{1-t}\right)\right)\frac{1}{(1-t)^2} = 0,\\
&k\left(\frac{i}{n}\right) = g\left(1 - F_Y\left(\frac{i}{n-i}\right)\right)\frac{n^2}{(n-i)^2}.
\end{aligned}$$

Therefore, we obtain a formula for approximate the risk $R_g[Y]$ as follow:

$$R_g[Y] \approx \frac{1}{2n} + \sum_{i=1}^{n-1} g\left(1 - F_Y\left(\frac{i}{n-i}\right)\right)\frac{n}{(n-i)^2}. \tag{14}$$

5 DISTRIBUTIONS OF A SUM OF TWO DEPENDENCE RANDOM VARIABLES USING COPULAS

We now turn to the main theorem deriving distribution of a sum of random variables using copulas.

Theorem 1. *Suppose that (X_1, X_2) be a continuous random vector having the marginal distributions F_1 and F_2, respectively, and they are not independent. Let C be an absolutely continuous copula modeling dependence structure of a random vector (X_1, X_2)* and define Y as

$$Y = w_1X_1 + w_2X_2, \tag{15}$$

where, $w_1, w_2 \in \mathbb{R} \setminus \{0\}$.

Then, the density and distribution function of Y are defined as follows:

$$f_Y(y) = \frac{1}{|w_2|}\int_0^1 c\left(u, F_2\left(\frac{y - w_1F_1^{-1}(u)}{w_2}\right)\right) f_2\left(\frac{y - w_1F_1^{-1}(u)}{w_2}\right)du, \tag{16}$$

$$F_Y(y) = \operatorname{sgn}(w_2)\int_0^1 \frac{\partial}{\partial u}C\left(u, F_2\left(\frac{y - w_1F_1^{-1}(u)}{w_2}\right)\right)du, \tag{17}$$

where, c denotes the density of copula C and sgn (x) *is a sign function of x*,

$$\operatorname{sgn}(x) = \begin{cases} 1, & \text{if } x > 0,\\ -1, & \text{if } x < 0.\end{cases}$$

Proof. Firstly, we set up

$$\begin{cases} Y_1 = w_1X_1 + w_2X_2\\ \quad Y_2 = X_1.\end{cases} \tag{18}$$

Let F and f be the joint distribution and join density of (X_1, X_2) . Then, due to Sklar's theorem (1959), there exists a unique copula C such that

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)). \tag{19}$$

Or it may write in term of joint density function as

$$f(x_1, x_2) = c(F_1(x_1), F_2(x_2))f_1(x_1)f_2(x_2), \tag{20}$$

where, c denotes density of copula C given by

$$c(u_1, u_2) = \frac{\partial^2 C(u_1, u_2)}{\partial u_1 \partial u_2}. \tag{21}$$

From (18), the inverse transform is

$$\begin{cases} X_1 = Y_2, \\ X_2 = \dfrac{Y_1 - w_1 Y_2}{w_2}. \end{cases}$$

The Jacobian of the transform is

$$J = \begin{vmatrix} \dfrac{\partial X_1}{\partial Y_1} & \dfrac{\partial X_1}{\partial Y_2} \\ \dfrac{\partial X_2}{\partial Y_1} & \dfrac{\partial X_2}{\partial Y_2} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ \dfrac{1}{w_2} & -\dfrac{w_1}{w_2} \end{vmatrix} = -\frac{1}{w_2} \neq 0.$$

Then, substituting into (20), we get the joint density of (Y_1, Y_2) denoted by h as follows:

$$\begin{aligned} h(y_1, y_2) &= f\left(y_2, \frac{y_1 - w_1 y_2}{w_2} \right) |J| \\ &= \frac{1}{|w_2|} c\left(F_1(y_2), F_2\left(\frac{y_1 - w_1 y_2}{w_2} \right) \right) \\ &\quad f_1(y_2) f_2\left(\frac{y_1 - w_1 y_2}{w_2} \right). \end{aligned} \tag{22}$$

Therefore, one can derive the density of Y_1 in the following:

$$\begin{aligned} f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} h(y_1, y_2) dy_2 \\ &= \frac{1}{|w_2|} \int_{-\infty}^{\infty} c\left(F_1(y_2), F_2\left(\frac{y_1 - w_1 y_2}{w_2} \right) \right) \end{aligned} \tag{23}$$

$$\begin{aligned} &f_1(y_2) f_2\left(\frac{y_1 - w_1 y_2}{w_2} \right) dy_2 \\ &= \frac{1}{|w_2|} \int_0^1 c\left(u, F_2\left(\frac{y_1 - w_1 F_1^{-1}(u)}{w_2} \right) \right) \\ &f_2\left(\frac{y_1 - w_1 F_1^{-1}(u)}{w_2} \right) du \end{aligned} \tag{24}$$

Next, the computation of $Y_1's$ distribution is straightforward:

$$\begin{aligned} F_{Y_1}(t) &= \int_{-\infty}^{t} f_{Y_1}(y_1) dy_1 \\ &= \int_{-\infty}^{t} \frac{1}{|w_2|} \int_0^1 c\left(u, F_2\left(\frac{y_1 - w_1 F_1^{-1}(u)}{w_2} \right) \right) \\ &\quad f_2\left(\frac{y_1 - w_1 F_1^{-1}(u)}{w_2} \right) du dy_1 \\ &= \frac{1}{|w_2|} \int_0^1 \int_{-\infty}^{t} c\left(u, F_2\left(\frac{y_1 - w_1 F_1^{-1}(u)}{w_2} \right) \right) \\ &\quad f_2\left(\frac{y_1 - w_1 F_1^{-1}(u)}{w_2} \right) dy_1 du. \end{aligned} \tag{25}$$

By taking $v = F_2\left(\frac{y_1 - w_1 F_1^{-1}(u)}{w_2} \right)$, the formula (25) becomes as shown:

$$\begin{aligned} F_{Y_1}(t) &= \frac{w_2}{|w_2|} \int_0^1 \int_0^{F_2\left(\frac{t - w_1 F_1^{-1}(u)}{w_2} \right)} c(u, v)\, dvdu \\ &= \frac{w_2}{|w_2|} \int_0^1 \int_{-\infty}^{F_2\left(\frac{t - w_1 F_1^{-1}(u)}{w_2} \right)} \frac{\partial^2 C}{\partial u \partial v}(u, v)\, dvdu \\ &= \operatorname{sgn}(w_2) \int_0^1 \frac{\partial}{\partial u} C\left(u, F_2\left(\frac{t - w_1 F_1^{-1}(u)}{w_2} \right) \right) du. \end{aligned}$$

The proof is completed.

Remark: Due to the fact that they are exchangeable, it is totally possible to obtain other formulas as (16) and (17), given by

$$\begin{aligned} f_Y(y) &= \frac{1}{|w_1|} \int_0^1 c\left(F_1\left(\frac{y - w_2 F_2^{-1}(v)}{w_1} \right), v \right) \\ &\quad f_1\left(\frac{y - w_2 F_2^{-1}(v)}{w_1} \right) dv, \end{aligned} \tag{26}$$

$$F_Y(y) = \operatorname{sgn}(w_1) \int_0^1 \frac{\partial}{\partial v} C\left(F_1\left(\frac{y - w_2 F_2^{-1}(v)}{w_1} \right), v \right) dv, \tag{27}$$

Let us consider a special case, $w_1 = w_2 = 1$. Then, from (23), we obtain an expression

$$\begin{aligned} f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} c(F_1(y_2), F_2(y_1 - y_2)) \\ &\quad f_1(y_2) f_2(y_1 - y_2) dy_2. \end{aligned} \tag{28}$$

This formula can be seen as a general convolution product (called C-convolution) of two dependent density functions. In fact, when X_1 and X_2 are independent, their copula is $C(u,v) = uv$. Thus, its copula density is $c(u,v) = 1$. Again, we get the usual convolution product

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_1(y_2) f_2(y_1 - y_2) dy_2. \tag{29}$$

6 APPLICATIONS

Let us consider the portfolio consisting of two assets as follows:

$$Y = w_1 X_1 + w_2 X_2, \tag{30}$$

where, X_1 presents return of Exxon Mobil's stock; X_2 presents return of JP Morgan's stock and Y denotes return of the portfolio.

The data are selected from New York Stock Exchange during the year 2013 and 2014. Assume that the portfolio investment would start in 2014. To build an optimal portfolio, one can use Markowitz rule for two assets by computing the weights w_1 and w_2 as follows:

$$w_1 = \frac{E(X_1)\sigma_2^2 - E(X_2)\sigma_{12}}{E(X_1)\sigma_2^2 + E(X_2)\sigma_1^2 - \left[E(X_1)+E(X_2)\right]\sigma_{12}}$$
$$w_2 = 1 - w_1, \tag{31}$$

where, σ_1^2, σ_2^2 are the variances of X_1 and X_2, respectively, and their covariance is σ_{12}. Here, the weights are calculated without risk free.

Using the data 2013, we obtain sample means and variances given by $\bar{x}_1 = 0.0006, \bar{x}_2 = 0.012, \sigma_1^2 = \sigma_2^2 = 0.0001$ and $\sigma_{12} = 4.754297 * 10^{-5}$. Thus, the optimal weights are determined by

$$w_1 = 3\%, \text{ and } w_2 = 97\%.$$

First of all, the descriptive summary of the portfolio is shown in Figure 1 and Table 1:

The important thing is that one has to make a sketch of an association between the return X_1 and X_2 and this can be done by using scatter plot as shown in Figure 2. Clearly, they are not independent (the Pearson's correlation coefficient $r(X_1, X_2 = 0.44)$. In addition, one can find out that although the weight for Exxon Mobil's return is very small, $w_1 = 0.03$, it has a quite large effect on the portfolio values, $r(X_1, Y) = 0.48$. Also, X_2 and Y have a perfect linear dependence, $r(X_2, Y) = 1$.

Next step, we will construct a joint distribution as well as a copula C modeling the dependence structure of X_1 and X_2. Firstly, marginal distributions are estimated by using maximum likelihood estimation method (MLE). As a result, F_1, F_2 and F_Y are approximated normal cumulative distribution functions (CDF), see Figures 3, 4, and 5.

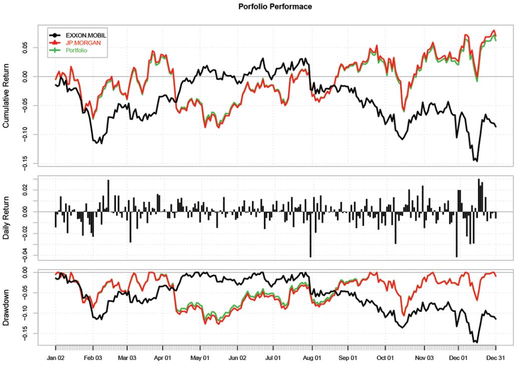


Figure 1. The Portfolio Performance.

Table 1. Descriptive statistics for the portfolio Y.

Statistics	Exxon	JP	Portfolio
Observations	260.00	260.00	260.00
Minimum	–0.0417	–0.0424	–0.0406
Quartile 1	–0.0056	–0.0047	–0.0049
Median	–0.0002	0.0005	0.0006
Arithmetic Mean	–0.0003	0.0003	0.0003
Geometric Mean	–0.0003	0.0003	0.0002
Quartile 3	0.0053	0.0069	0.0069
Maximum	0.0302	0.0352	0.0331
SE Mean	0.0006	0.0007	0.0007
LCL Mean (0.95)	–0.0015	–0.0010	–0.0010
UCL Mean (0.95)	0.0010	0.0017	0.0016
Variance	0.0001	0.0001	0.0001
Stdev	0.0102	0.0111	0.0108
Skewness	–0.4977	–0.2753	–0.2819
Kurtosis	2.3042	1.3998	1.3119

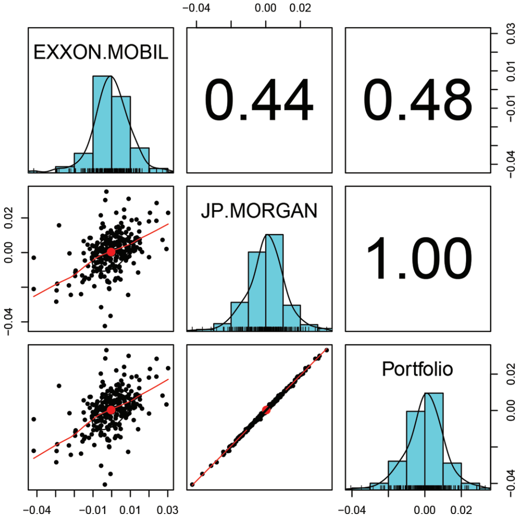


Figure 2. The relationship among X_1, X_2 and Y.

To estimate a copula $\hat{C}$ of X_1 and X_2, one can use the *copula* package from R, see Yan et al. (2007). As we can see in Table 2, Student copula (with parameter $\rho = 0.46$ and degree of freedom $\nu = 9$) could be the best fit for dependence structure of X_1 and X_2 due to the fact that the maximized log likelihood value is the highest among the common copulas such as family of normal, Student, Gumbel, Frank and Clayton, as shown in Table 2. To verify this fact, we apply goodness-of-fit test using Cramer-von Mises statistic and then

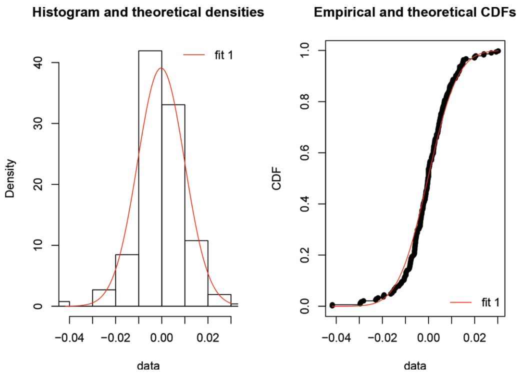


Figure 3. $\widehat{F_1}$ approximates normal CDF, $N\left(-0.03\%,1.02^2\%\right)$.

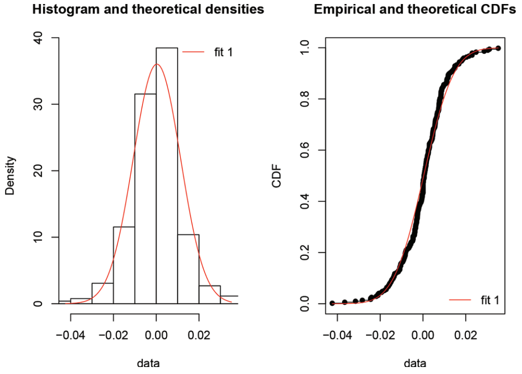


Figure 4. $\widehat{F_2}$ approximates normal CDF, $N\left(0.03\%,1.11^2\%\right)$.

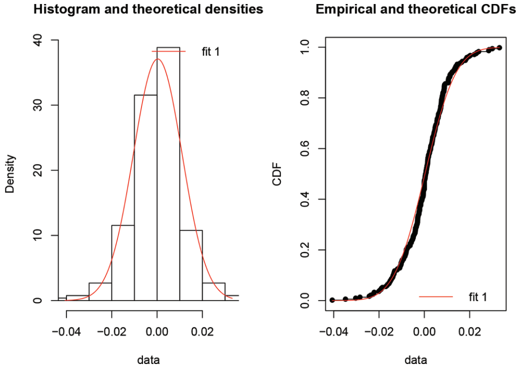


Figure 5. $\widehat{F_Y}$ approximates normal CDF, $N\left(0.02\%,1.08^2\%\right)$.

Table 2. Results for estimating copula $\widehat{C}$.

Copula	Param.	Std.	z value	P-value	Loglike
Normal	0.46	0.045	10.26	$2 * 10^{-16}$	28.9
Student	0.46; 9	0.053	8.68	$2 * 10^{-16}$	30.17
Gumbel	1.38	0.064	21.41	$2 * 10^{-16}$	25.70
Frank	2.89	0.426	6.79	$1.15 * 10^{-11}$	26.44
Clayton	0.70	0.094	7.39	$1.45 * 10^{-13}$	26.66

Table 3. Goodness-of-fit test for Student copula with $\nu=9$.

Copula	Statistic	Parameter	P-value
$\widehat{C}$	0.015919	0.45947	0.6658

Table 4. Measures of dependence for Student copula $\widehat{C}$.

Copula	Kendall's τ	Spearman's ρ	Tail Index	$\lambda(C)$
$\widehat{C}$	0.3043	0.4432	0.0834	0.4377

present the result in Table 3. It is clear that P-value is 0.6658 which is higher than the significant level α, say $\alpha=1\%$. Hence, there is enough evidence to conclude that the Student copula can be used to model dependence structure of the two returns and the degree of dependence is moderate, see Table 4. Note: the tail indices $\lambda_U=\lambda_L$.

Applying Sklar's theorem, one can definitely construct a join distribution for X_1 and X_2 as follow:

$$\widehat{H}(x_1,x_2)=\widehat{C}\left(\widehat{F}_1(x_1),\widehat{F}_2(x_2)\right), \tag{32}$$

where, $\widehat{F}_1$ and $\widehat{F}_2$ approximate normal distribution as shown in Figure 3 and 4; $\widehat{C}$ is a Student copula with the parameter $\rho=0.46$ and $\nu=9$, given by

$$\widehat{C}(u,v)=t_{\rho,\nu}\left(t_\nu^{-1}(u),t_\nu^{-1}(v)\right),$$

where, $t_{\rho,\nu}$ is the cumulative distribution function of a bivariate Student distribution, ρ is the correlation coefficient and ν is the degree of freedom.

In section 5, we have shown a new method to establish distribution of the return portfolio Y which is a linear combination of two dependent asset returns X_1 and X_2. In the above arguments, the dependence structure has been determined by

Student copula. Therefore, the density and distribution of Y will come out naturally from (16) and (17). The numerical results are plotted in Figure 6 and 7. Furthermore, the graphs seem to perform an approximately normal distribution that is consistent with the results using maximum likelihood estimation method as in Figure 5.

Finally, we can apply the formula (14) with several distortion functions to estimate risks for the portfolio Y, as shown in Table 5. Note: $n = 1000, \gamma_1 = \Phi^{-1}(0.05)$ and $\gamma_2 = \Phi^{-1}(0.01)$.

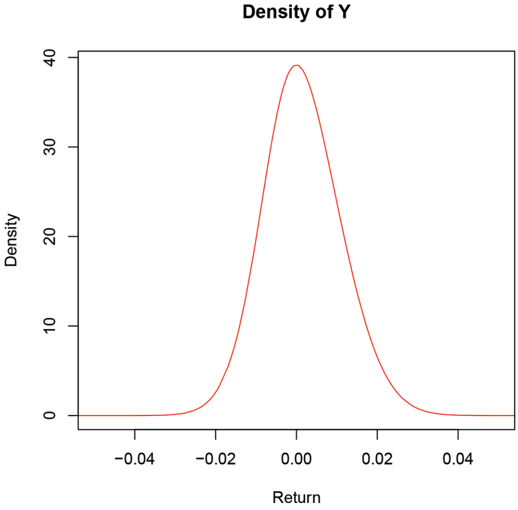


Figure 6. Density of Y using C-convolution.

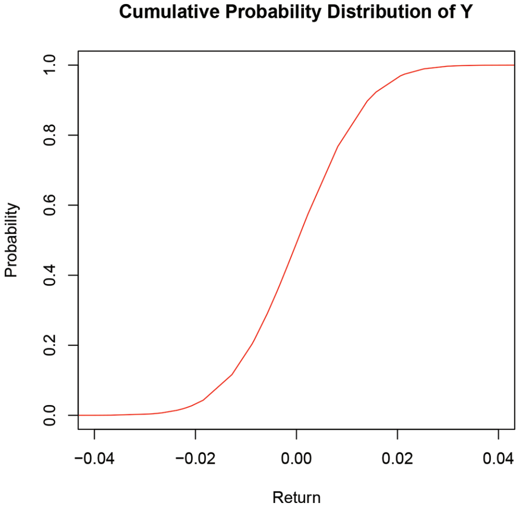


Figure 7. Cumulative distribution of Y using C-convolution.

Table 5. Risk Measures for X_1, X_2, Y.

Risk	Exxon Mobil	JP Morgan	Portfolio
VaR 5%	–1.70%	–1.79%	–1.74%
VaR 1%	–2.39%	–2.53%	–2.47%
ES 5%	–2.12%	–2.24%	–2.19%
ES 1%	–2.74%	–2.91%	–2.84%
Wang γ_1	–1.66%	–1.88%	–1.77%
Wang γ_2	–2.25%	–2.46%	–2.48%

7 CONCLUSIONS

We have proposed a new method to estimate distortion risk measures and use copula-based procedure to approach the distribution of a portfolio consisting of dependent assets. The latter is our main focus since all the information of dependence is used properly. For further research, we are going to study the optimization problem of a general portfolio using the copula's approach as well as its distributed behavior.

ACKNOWLEDGEMENT

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science—LQ1602".

REFERENCES

Aas, K. (2004). Modelling the dependence structure of financial assets: A survey of four copulas.

Balbás, A., J. Garrido, & S. Mayoral (2009). Properties of distortion risk measures. *Methodology and Computing in Applied Probability 11*(3), 385–399.

Cherubini, U., E. Luciano, &W. Vecchiato (2004). *Copula methods in finance*. John Wiley & Sons.

Cherubini, U., S. Mulinacci, & S. Romagnoli (2011). A copulabased model of speculative price dynamics in discrete time. *Journal of Multivariate Analysis 102*(6), 1047–1063.

Embrechts, P., F. Lindskog, & A. McNeil (2001). Modelling dependence with copulas. *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*.

Frees, E.W. & E.A. Valdez (1998). Understanding relationships using copulas. *North American actuarial journal 2*(1), 1–25.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

Kim, J.H. (2010). Bias correction for estimated distortion risk measure using the bootstrap. *Insurance: Mathematics and Economics 47*(2), 198–205.

Nelsen, R.B. (2006). *An introduction to copulas*. Springer Science & Business Media.

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

On the performance of sequential procedures for detecting a change, and Information Quality (InfoQ)

Ron S. Kenett

The KPA Group, Raanana, Israel, Department of Mathematics "G. Peano", University of Turin, Italy
Institute for Drug Development, The Hebrew University of Jerusalem, Israel
Faculty of Economics, Ljubljana, Slovenia, Center for Risk Engineering, NYU Tandon School of Engineering, New York, USA

ABSTRACT: The literature on statistical process control has focused on the Average Run Length (ARL) to an alarm, as a performance criterion of sequential schemes. When the process is in control, ARL_0 denotes the ARL to false alarm and represents the in-control operating characteristic of the procedure. The average run length from the occurrence of a change to its detection, typically denoted by ARL_1, represents the out-of-control operating characteristic. These indices however do not tell the whole story. The concept of Information Quality (InfoQ) is defined as the potential of a dataset to achieve a specific (scientific or practical) goal using a given empirical analysis method. InfoQ is derived from the Utility (U) of applying an analysis (f) to a data set (X) for a given purpose (g). Formally, the concept of Information Quality (InfoQ) is defined as: InfoQ(f, X, g) = U(f(X | g)). These four components are deconstructed into eight dimensions that help assess the information quality of empirical research in general. In this paper, we suggest the use of Probability of False Alarm (PFA) and Conditional Expected Delay (CED) as an alternative to ARL_0 and ARL_1 enhances the Information Quality (InfoQ) of statistical process control methods. We then review statistical process control methods from a perspective of the eight InfoQ dimensions. As an extension, we discuss the concept of a system for statistical process control.

1 INTRODUCTION

Change point detection and process control sequential methods are designed to detect change. This paper discusses how performance indicators such as Conditional Expected Delay (CED) and Probability of False Alarm (PFA) enhances the information quality of statistical process control methods. As an extension, a System for Statistical Process Control (SSPC), in the context of a life cycle view of statistics, is presented. A main point provided by this approach is that ARL_0 and ARL_1 are not sufficiently informative and therefore not adequate for determining or comparing performance of alternative process control sequential methods.

Kenett and Shmueli (2014, 2016) formulate the concept of information quality (InfoQ). InfoQ is derived from the Utility (U) of applying an analysis (f) to a data set (X) for a given purpose (g). Eight dimensions help assess the level of InfoQ of a study. These are: Data Resolution, Data Structure, Data Integration, Temporal Relevance, Generalizability, Chronology of Data and Goal, Operationalization, and Communication. These eight dimensions represent a deconstruction of the four InfoQ components: U, f, X and g. In this paper we review and discuss change point detection performance indicators using an InfoQ perspective. Section 2 is an introduction to InfoQ, Section 3, 4 and 5 are about detection of change, false alarms and detection delay. Section 6 is a detailed analysis of change detection using InfoQ, section 7 is presenting a system for statistical process control and section 8 presents a summary and discussion.

2 INFORMATION QUALITY

Information Quality (InfoQ) is the potential of a dataset to achieve a specific (scientific or practical) goal using a given empirical analysis method (Kenett and Shmueli, 2014, 2016). InfoQ is different from data quality and data analysis quality, but is dependent on these components and on the relationship between them. InfoQ is derived from the utility of applying an analysis (f) to a data set (X) for a given purpose (g). Formally the concept of Information Quality (InfoQ) is defined as:

$$\text{InfoQ}(f, X, g) = U(f(X \mid g))$$

InfoQ is therefore affected by the quality of its components g ("quality of goal definition"), X ("data quality"), f ("analysis quality"), and U ("utility measure") as well as by the relationships between X, f, g and U. Expanding on the four InfoQ components provides some additional insights.

Analysis Goal (g): Data analysis is used for various purposes. Three general classes of goals are causal explanations, predictions, and descriptions. Causal explanations include questions such as "Which factors cause the outcome?". Descriptive goals include quantifying and testing for population effects using data summaries, graphical visualizations, statistical models, and statistical tests. Prediction goals include forecasting future values of a time series and predicting the output value of new observations given a set of input variables.

Data (X): The term "data" includes any type of data to which empirical analysis can be applied. Data can arise from different collection tools such as surveys, laboratory tests, field and computer experiments, simulations, web searches, observational studies and more. "Data" can be univariate or multivariate and of any size. It can contain semantic, unstructured information in the form of text or images with or without a dynamic time dimension. Data is the foundation of any application of empirical analysis.

Data Analysis Method (f): The term data analysis refers to statistical analysis and data mining. This includes statistical models and methods (parametric, semi-parametric, non-parametric), data mining algorithms, and machine learning tools. Operations research methods, such as simplex optimization, where problems are modelled and parametrized, also fall into this category.

Utility (U): The extent to which the analysis goal is achieved, as measured by some performance measure or "utility". For example, in studies with a predictive goal, a popular performance measure is predictive accuracy. In descriptive studies, common utility measures are goodness-of-fit measures. In explanatory models, statistical power and goodness-of-fit measures are common utility measures.

Eight dimensions are used to deconstruct InfoQ and thereby provide an approach for assessing it. These are: Data Resolution, Data Structure, Data Integration, Temporal Relevance, Chronology of Data and Goal, Generalizability, Operationalization and Communication. We proceed with a description of these dimensions.

i. *Data Resolution*: Data resolution refers to the measurement scale and aggregation level of X. The measurement scale of the data needs to be carefully evaluated in terms of its suitability to the goal, the analysis methods to be used, and the required resolution of U. Given the original recorded scale, the researcher should evaluate its adequacy. It is usually easy to produce a more aggregated scale (e.g., two income categories instead of ten), but not a finer scale. Data might be recorded by multiple instruments or by multiple sources. To choose among the multiple measurements, supplemental information about the reliability and precision of the measuring devices or data sources is useful. A finer measurement scale is often associated with more noise; hence the choice of scale can affect the empirical analysis directly. The data aggregation level must also be evaluated in relation to the goal.

ii. *Data Structure*: Data structure relates to the type of data analysed and data characteristics such as corrupted and missing values due to the study design or data collection mechanism. Data types include structured numerical data in different forms (e.g., cross-sectional, time series, network data) as well as unstructured, non-numerical data (e.g., text, text with hyperlinks, audio, video, and semantic data). The InfoQ level of a certain data type depends on the goal at hand.

iii. *Data Integration*: With the variety of data sources and data types, there is often a need to integrate multiple sources and/or types. Often, the integration of multiple data types creates new knowledge regarding the goal at hand, thereby increasing InfoQ. For example, in online auction research, the integration of temporal bid sequences with cross-sectional auction and seller information leads to more precise predictions of final prices as well as to an ability to quantify the effects of different factors on the price process.

iv. *Temporal Relevance*: The process of deriving knowledge from data can be put on a time line that includes the data collection, data analysis, and study deployment periods as well as the temporal gaps between the data collection, the data analysis, and the study deployment stages. These different durations and gaps can each affect InfoQ. The data collection duration can increase or decrease InfoQ, depending on the study goal, e.g. studying longitudinal effects vs. a cross-sectional goal. Similarly, if the collection period includes uncontrollable transitions, this can be useful or disruptive, depending on the study goal.

v. *Chronology of Data and Goal*: The choice of variables to collect, the temporal relationship between them, and their meaning in the context of the goal at hand also affects InfoQ. For example, in the context of online auctions, classic auction theory dictates that the number of bidders is an important driver of auction price. Models based on this theory are useful for explaining the effect of the number of bidders on price. However, for the purpose of predicting the price of ongoing online auctions, where the number of bidders is unknown until the auction ends, the variable "number of bidders", even if available in the data, is useless. Hence, the level of InfoQ contained in “number of bidders” for models of auction price depends on the goal at hand.

vi. *Generalizability*: The utility of f(X|g) is dependent on the ability to generalize f to the appropriate population. There are two types of generalization, statistical and scientific generalizability. Statistical generalizability refers to inferring from a sample to a target population. Scientific generalizability refers to applying a model based on a particular target population to other populations. This can mean either generalizing an estimated population pattern/model f to other populations, or applying f estimated from one population to predict individual observations in other populations using domain specific knowledge.

vii. *Operationalization*: Operationalization relates to both construct operationalization and action operationalization. Constructs are that describe a phenomenon of theoretical interest. Measurable data is an operationalization of underlying constructs. The relationship between the underlying construct and its operationalization can vary, and its level relative to the goal is another important aspect of InfoQ. The role of construct operationalization dependents on the goal, and especially on abstractions whether the goal is explanatory, predictive, or descriptive. In explanatory models, based on underlying causal theories, multiple operationalizations might be acceptable for representing the construct of interest. As long as the data is assumed to measure the construct, the variable is considered adequate. In contrast, in a predictive task, where the goal is to create sufficiently accurate predictions of a certain measurable variable, the choice of operationalized variables is critical. Action operationalization is characterizing the practical implications of the information provided.

viii. *Communication*: Effective communication of the analysis and its utility directly impacts InfoQ. There are plenty of examples where miscommunication of valid results has led to disasters, such as the NASA shuttle Challenger disaster (Kenett and Thyregod, 2006). Communication media are visual, textual, and verbal in the form of presentations and reports. Within research environments, communication focuses on written publications and conference presentations. Research mentoring and the refereeing process are aimed at improving communication and InfoQ within the research community.

3 DETECTION OF CHANGE

Change happens and, invariably, its early detection is of importance. Usually we are not given advance notice of the occurrence of change, and detection must rely on observations made on the system being monitored. Generally, post-change observations differ stochastically from pre-change ones, and a single observation or a finite set of observations does not clearly differentiate the pre- and post-change regimes. Consequently, a trigger-happy detection scheme will give rise to many false alarms, whereas a conservative procedure will be too slow to react. For comparison between different methods, operating characteristics of a detection scheme must be formulated.

As an example, consider the 4 run charts in Figure 1. The series Y1-Y4 were generated with a change point at the 10th observation using MINITAB® version 16.0. The data is a realization of a normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 3$, with shifts in the mean after the 10th observation to 11.5, 13, 14.5 and 16 respectively.

Just as in other situations where statistical methods are applied, the approach to change point detection may be frequentist or Bayesian. The problem has the flavour of testing hypotheses. At each stage, one must decide whether a change is in effect (and raise an alarm) or whether the process is in control (and continue the monitoring). The frequentist approach calls for separate operating characteristics for the in-control and the out-of-control situations. In the Bayesian context, an operating characteristic combines the in- and out-of-control scenarios by means of the prior distribution on the change point, τ. A partial list of the vast literature on these topics includes Page (1954), Shiryaev (1963), Lorden (1971), Lucas (1976), Zacks (1981), Kenett and Pollak (1983, 2012), Pollak (1985), Yashchin (1985), Zacks and Kenett (1994), Kenett and Pollak (1996), Woodall and Montgomery (1999), Frisén (2003) and Box and Luceño (2006).

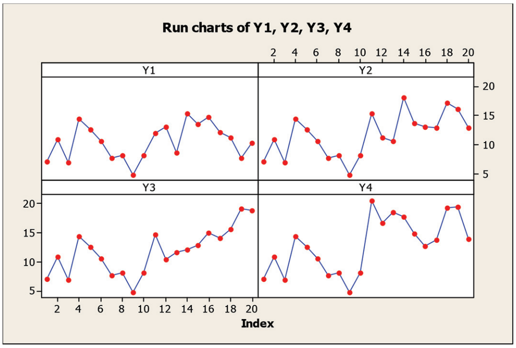


Figure 1. Four time series with change point at the 10th observation.

4 FALSE ALARMS

The popular index for false alarms, when viewing a problem from a frequentist point of view, is ARL_0, the Average Run Length to false alarm. Another quantity of interest is $P_{in\text{-}control}(T = n|T \geq n)$, the probability that a stopping time T will raise a false alarm at time n, conditional on its not having raised a false alarm previously. This may depend on n, so the worst-case scenario is indexed by $\sup_{n<\infty} P_{in\text{-}control}(T = n|T \geq n)$, and one may want to keep this quantity low.

The ARL to false alarm has been criticized for not fully considering the skewness of the distribution of the run length (Lai, 1995, Zacks, 2004, Mei, 2008, Frisén, 2003, 2006, 2011). It should be noted however that, very often, the run length to false alarm has approximately an exponential distribution and for discrete time, a geometric distribution (Gold, 1989). Since an exponential distribution is fully characterized by its mean, in such cases, the ARL to false alarm fully describes the false alarm behaviour of a procedure (Knoth, 2015). In surveillance applications, processes cannot be reset like in typical industrial applications where a machine can be stopped. In such cases an investigation is triggered by an alarm and the state of alarm can be persistent. Kenett and Pollak (1983) account for the frequency of false alarms when a process cannot be reset in an application of monitoring congenital malformations.

When considering the problem from a Bayesian point of view, where there is a prior on the change point τ, the quantity of interest for describing the possibility of a false alarm is the probability of false alarm, PFA = $P(T<\tau)$. A common constraint on false alarms is $P(T<\tau)\leq\alpha$.

Shiryaev (1961, 2010) considers that the error to be minimized over all stopping times of X is expressed as the linear combination of the probability of the false alarm and the expected detection delay. His change point scenario is defined as multi-cylic stationary and, in that context the Shyarev-Roberts (SR) procedure is optimal (Kenett and Pollak, 1986, 1996). The Shiryaev's multi-cyclic version of the change-point detection problem is equivalent to the generalized Bayesian setup so that the SR procedure is exactly optimal in the generalized Bayesian sense as well. It should be noted that neither the Cumulative Sum (CUSUM) nor the Exponentially Weighted Moving Average (EWMA) possess such strong optimality properties (Kenett and Pollak, 2012).

Related results, in the frequentist framework, are presented in Hillier (1969), which discusses the False Alarm Rate (FAR), and Chakraborti et al. (2008) which introduces the False Alarm Probability (FAP). The value of FAR is the probability of false alarm at every sampling stage (with every subgroup or with every measurement result), i.e. FAR = 1- Pr{LCL < Yi < UCL}, where Yi is the statistic being tracked for the ith sample (individual value X_i, or average value $\bar{X}$, or range R, or standard deviation S, etc.); LCL and UCL being the lower and upper control limits used for process monitoring. The FAP is the probability of at least one false alarm during the established period of time, i.e. FAP = 1- $[1\text{-} \Pr\{LCL < Yi < UCL\}]^m$ = $1\text{-}(1\text{-}FAR)^m$, where m is the number of subgroups included in the process capability analysis phase and the i.i.d. assumption is implied. The value of FAP is used to calculate the limits for verifying the stability of the process when establishing control limits with different limits. FAP is correcting for a multiple comparison effect during process capability analysis by applying a Bonferroni correction. Another approach, different from this family error wise consideration, is to apply a False Discovery Rate (FDR) correction which considers, not the number of comparisons (data points) but the ratio of false alarms relative to the number of alarms. For more on this topic see Kenett and Zacks (2014). Because of the dynamic aspect of process control, both PFA and FDR have adequate performance in retrospective data analysis (the process capability analysis phase) but limited relevance in future looking process monitoring.

5 DELAY TO DETECTION

From a frequentist point of view, the post-change ARL is often characterized by ARL_1, the ARL to detection assuming that the change is in effect at the very start. However, letting τ denote the serial number of the first post-change observation, the conditional expected delay of detection, conditional on a false alarm not having been raised before the (unknown) time of change τ, is $CED(\tau) = E(T - \tau + 1| T \geq \tau)$. CED may depend on τ, and there is no guarantee that CED, as a function of τ, is represented well by ARL_1. If one has no anticipation of the time of change, $\sup_{\tau<\infty} P_{out\text{-}of\text{-}control}\, CED(\tau)$ can be considered an appropriate index. In some sequential procedures, such as the Shewhart control chart, $\sup_{\tau<\infty} P_{out\text{-}of\text{-}control}(T = \tau|T \geq \tau) = (ARL_1)^{-1}$. Moreover, CED(τ) is constant for Shewhart charts. If there is a good chance that a change will be in effect right at the start, one may be interested in a fast initial response scheme, where CED(0) is made to be low, at the expense of a higher CED at later τ, so that ARL_1 does not tell the whole story (Lucas and Crosier, 1988). If a change is likely to take place in a distant future, then $\lim_{\tau\to\infty} CED(\tau)$, the conditional steady-state ARL, may be of interest. An alternative index is $P_{out\text{-}of\text{-}control}(T = \tau|T \geq \tau)$, and if one has no anticipation of the time of change, $\sup_{\tau<\infty}(T=\tau|T\geq\tau)$ can be considered an appropriate index.

In principle, the CED may not be the primary characteristic of interest. For example, consider the case of monitoring for the outbreak of an epidemic. For illustration's sake, suppose simplistically that each infected person infects k others (all within the next time unit). Thus, if the epidemic starts with one person, at the second time unit k+1 are infected, at the third time unit k^2+k+1 have been infected, etc.; after n time units the number of infected people adds up to $n(n+1)(2n+1)/6 \approx n^3/3$. Hence the primary object of interest would be $E_{\tau}((T-\tau+1)^3| T\geq\tau)$. Or, if each infected person subsequently infects one other person every time unit, n time units after the start of the epidemic the number of infected people will be 1+1+2+4+8+… $+2^{n-2} = 2^{n-1}$; hence the primary object of interest would be $E_{\tau}(2^{T-\tau} | T\geq\tau)$.

Even if the price for the delay in detection is linear in $(T - \tau +1)^+$, ARL_1 may not be a meaningful operating characteristic. For example, consider monitoring for a change of a mean μ to a mean δ, when the baseline μ and the post-change parameter δ are unknown. For example, suppose one wants to monitor a change in the average daily water flow in a river, where one has no historic data and only Bayesian priors. Obviously, if the change occurs at the onset, no surveillance system will be able to differentiate between pre-change and post-change, so that the expected delay to detection will equal the ARL to false alarm. Approximately, the same will happen if the change takes place within a few observations after the onset of surveillance. If the change occurs later on, the pre-change observations may constitute a learning sample of sufficient size to reduce the CED to the proportions of the CED of a procedure, like in a situation where the baseline parameters are known. Hence, ARL_1 is not a good index. Figure 2 shows simulated run lengths and the respective values of PFA, CED and ARL for the four process scenarios shown in Figure 1 when applying a two sided CUSUM procedure set up to detect the specific change in the scenario. This assumes the CUSUM is specified optimally, during process capability analysis with exact knowledge of the process state before and after change. The parameters used for the four simulations are:

Teta	Teta+	K+	h+	Teta-	K-	h-
10	11.5	10.75	17.9744	8.5	9.25	–17.9744
	13.0	11.50	8.9872	7.0	8.50	–8.9872
	14.5	12.25	5.9915	5.5	7.75	–5.9915
	16.0	13.00	4.4936	4.0	7.00	–4.4936

For the R code to run these simulations see Kenett and Zacks (2014) and the *mistat* R application available for download in https://cran.r-project.org/web/packages/mistat/index.html. For usability and meaning of measures for evaluating detection schemes and general R code for computing performance indicators of sequential methods, see Knoth (2006) and http://cran.r-project.org/web/packages/spc. For assessment of surveillance schemes see: http://economics.handels.gu.se/english/Units+and+Centra/statistical_research_unit/software.

The skewed run length distributions render interpretations of ARL low in information quality and PFA and CED with higher information quality, especially in terms of operationalization and communication.

For more literature on CED see Kenett and Pollak, 1983, 1986, 1996, 2012, Zacks and Kenett, 1994, Kenett and Zacks, 1998, 2014, Luceño and Cofiño, 2006 and Frisén, 2011.

Another situation where ARL_1 is not an appropriate index is when one is willing to tolerate many false alarms. As an example, consider checking for an intruder, where it is of utmost importance that the intrusion be detected even at the price of making many false alarms. Here, the characteristic of interest is the expected delay and, again, this is different from ARL_1.

When considering the problem from a Bayesian point of view, the quantity of interest for describing the possibility that a change is in effect is $P(\tau\leq n)$. Usually, the speed of detection of a method defined by a stopping time T is embodied by $E(T - \tau +1| T \geq \tau)$. Note that although this looks like the CED, there is a subtle difference: the CED regards τ as an unknown constant, whereas the Bayesian expression is, in effect, a weighted average of delay times. Considerations, as in the frequentist case, of $E((T - \tau +1)^3| T\geq\tau)$ or $E(2^{T-\tau} | T\geq\tau)$ apply here, too.

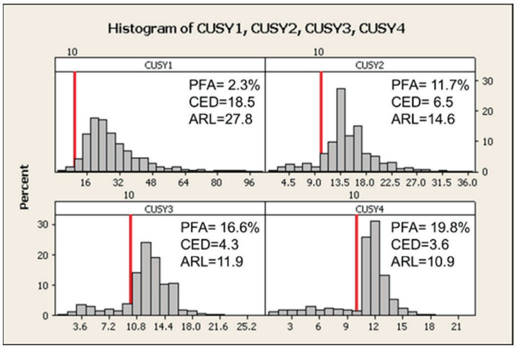


Figure 2. Run length distributions with CUSUM for the four series in Figure where changed occurred at the 10th observation.

6 INFOQ ASSESSMENT OF CHANGE DETECTION

In this section we review the above considerations from an InfoQ perspective. We begin with a discussion of the four InfoQ components. As introduced in Section 3, InfoQ is derived from the utility (U) of applying an analysis (f) to a data set (X) for a given goal (g).

The *goal* of change point detection is typically economically motivated. If we are able to design a process with acceptable capability, and we want to avoid reliance on mass inspection, it is essential to keep the process under control (AT&T, 1956). Statistical process control consists of an alarm triggering mechanism and proactive management actions that trigger corrective actions. This is an economically optimal combination. In some cases, the control limits can be determined by considering various cost elements (Kenett and Zacks, 2014), but in most cases the specific economic costs are not used to set up the process control system. In identifying the goal of a change point detection method it is critical to distinguish between processes that can be reset, such as industrial machines, and situation where an alarm triggers an investigation with delayed impact, such as in surveillance of health related epidemics.

The *utility* of a change point detection method is assessed by various performance indicators such as those discussed above. Traditionally these are ARL_0 and ARL_1, however, as suggested, it appear that PFA and CED are more informative, see also Kenett and Pollak 2012.

The *data* used in process control is univariate or multivariate. In some cases that data is grouped in rational samples that represent inherent local variability. Such local variability is used to determine control limits for ongoing process monitoring.

The *analysis* of process control data is based on a conceptual framework that is different from the classical hypothesis testing framework. In fact, Shewhart's view on statistical control, presented in his 1931 book, is connected to predictability (Di Bucchianico and Van Heuvel, 2015). Note that Shewhart's view, in his 1939 book, is described in terms of exchangeability and has an almost Bayesian flavour. The process control perspective is that the data analysed is generated by a process under investigation. The objective of change point detection is to identify a change from an underlying condition which was used to determine the process capability. Unlike classical statistical modelling, if a condition of change is detected, especially in processes that can be reset, the process is changed. The implication being that if the data does not fit the model, you do not fit a new model to the data but only observe if the control intervention has been successful and the process is back under control. This aspect of process control is not alwfays appreciated and many textbooks and papers consider process control as a standard application of hypothesis testing. Part of this confusion might be due to the relatively recent nomenclature of Phase I, where control limits are set and Phase II where monitoring is performed. In fact, Phase I and Phase II are typically iterated. A preferred convention would be to name this phases process capability analysis and monitoring phases. For more on this fundamental difference see AT&T (1956) and Hawkins et al. (2003).

As mentioned in Section 2, eight dimensions are used to assess the level of InfoQ of a study. We proceed to review these dimensions in the context of change point detection procedures in process control.

i. *Data resolution* is related to the concept of rational samples. The frequency and extent of the data sample used to control a process is a reflection of the process characteristics. For example, multi stream processes require a sample with representations of individual streams. Generally stable processes do not require data at the microsecond level and can probably achieve proper control with quarterly or even hourly data.
ii. *Data structure* is about the available data types such as time series, cross-sectional, panel data, geographic, spatial, network, text, audio, video, semantic, structured, semi or non-structured data. These can include output quality or process data, including video images or textual inputs by operators.
iii. *Data integration* considers how process monitoring data from different sources is integrated. Such methods include Extract-Transform-Load (ETL) methods, Bayesian networks, data fusion and general machine learning methods (Goeb, 2006, Weeze et al, 2015, Kenett, 2016).
iv. *Temporal relevance* is relevant to both the data used for the process capability analysis stage and data used for ongoing monitoring.
v. *Chronology of Data and Goal* is the dimension determining effectiveness of sequential methods. The signals should be produced in a timely and informative manner. This is why Conditional Expected Delay (CED) is such an essential performance measure.
vi. *Generalizability* is at the core of statistical process control. The information generated from the process control procedure should be used by operators, engineers and managers in a broader context than the specific sample points. The first generalization is statistical in scope, deriving insights on the process from

the rational samples. Further generalization consists of considering other similar processes, impact of raw materials or management effects, like shifts or training methods. The point here is to generalize the change point detection signals and data driven statistics to various application domains (for more on generalizability see Chapter 11 in Kenett and Shmueli, 2016).

vii. *Operationalization* is again a critical dimension of process control. Control charts that are ignored, or looked at retrospectively with considerable time delays, are not informative.

viii. *Communication*. The simple display of data over time, with an annotation scheme pointing out alarms based on diverse triggering mechanisms, is an essential element of process control. Combining the display with mathematical calculations has made these methods so popular.

The next section discusses an expanded view of statistical process control, adding a system perspective that integrates elements included in the InfoQ perspective.

7 A SYSTEM FOR STATISTICAL PROCESS CONTROL

A System for Statistical Process Control (SSPC) is an infrastructure, mostly technological, that enhances the impact of change point detection methods. In terms of functionality, an SSPC provides features for data acquisition, data integration, reporting, filtering and visualisation. An outline of such a system is presented in Figure 3. Such systems integrate with ERP systems so that data from work orders is automatically linked to the statistical process control procedures, including a meta-tagging of critical parameters and their specification limits. An additional feature of SSPC is its ability to handle data with high volume, velocity and variety, so called "big data". Integrating structured and unstructured data leads to improved diagnostic and troubleshooting capabilities, for example using Bayesian networks (Kenett, 2016).

In designing, or evaluating, an SSPC, one can apply the eight InfoQ dimensions: Data Resolution, Data Structure, Data Integration, Temporal Relevance, Chronology of Data and Goal, Generalizability, Operationalization and Communication. These dimensions will help the system designers and implementer cover the scope of functionalities needed by an SSPC. Specifically, the data collected by the system needs to have the right resolution, structure and temporal relevance. This includes, for example, the on line and off line measurements of process outputs, in-process parameters, work order requirements, traceability of parts across revisions, trouble shooting and corrective actions information in the form of text and images etc etc. All this data needs of course to be properly integrated and analyzed.

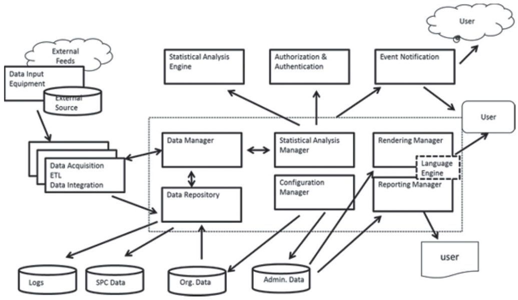


Figure 3. High level design of an SSPC.

An additional capability of SSPC is that application of multivariate statistical process control methods and advanced root cause analysis tools involving machine learning algorithm. In general, the application of multivariate process control has been delayed because of the complexity in deployment. Within SSPC, methods proposed in the past, such a multivariate tolerance regions (Fuchs and Kenett, 1987), can now be easily implemented. For more on multivariate process control and machine learning methods see Goeb 2006, Kenett and Zacks, 2014 and Weeze et al, 2015. For an example of an SSPC see www.spclive365.com.

8 SUMMARY AND DISCUSSION

Process control is a major elements in the body of knowledge analysed and developed in industrial statistics and applied statistics in general. In this paper we review several considerations of process control methods and discussions in the literature from a perspective of Information Quality (InfoQ). The main point is that an InfoQ approach broadens the scope of much of the work on the subject, as currently presented in books and journal articles. With similar considerations, Dennis Lindsey writes about the question *what is meant by statistics* by referring to who he considers as the founding fathers: Harold Jeffreys, Bruno de Finetti, Frank Ramsey and Jimmie Savage: "Both Jeffreys and de Finetti developed probability as the coherent appreciation of uncertainty, but Ramsey and Savage looked at the world rather differently. Their starting point was not the concept of uncertainty but rather decision-making in the face of uncertainty. They thought in terms of action, rather than in the passive contemplation of the uncertain world. Coherence for them was not so much a matter of how your beliefs hung together but of whether your several actions, considered collectively, make sense.....If one looks today at a typical statistical paper that uses the Bayesian method, copious use will be made of probability, but utility, or maximum expected utility, will rarely get a mention..... When I look at statistics today, I am astonished at the almost complete failure to use utility....Probability is there but not utility. This failure has to be my major criticism of current statistics; we are abandoning our task half-way, producing the inference but declining to explain to others how to act on that inference. The lack of papers that provide discussions on utility is another omission from our publications." (Lindsey, 2004). The four InfoQ components and the eight InfoQ dimensions are proposed as an antidote to the issues raised by Lindsey. We focus here on an evaluation of process control methods and change point detection procedures. By taking an InfoQ perspective we re-evaluate the performance indicators used in the literature to compare procedures and emphasize the application of CED and PFA, instead of the commonly used ARL_0 and ARL_1. By considering the growing role and impact of technology on analytic methods, we describe a System for Statistical Process Control (SSPC) that can help enhance the impact and relevance of statistical process control in modern business and industry. We also show how InfoQ dimensions can be used to design and evaluate such systems. The challenges we describe are not specific to process control and are relevant to modern applied statistics and quality management systems in general. In this sense, SSPC is a special case of integrated quality management systems envisioned by Juran in the 1950s (Godfrey and Kenett, 2007).

ACKNOWLEDGEMENTS

In preparing this paper several people contributed comments, suggestions and perspectives. These helped enhance the scope and completeness of the paper. I would like to particularly thank Alessandro di Bucchianico, Murat Testik, Sven Knoth, Stefan Steiner and Rainer Goeb for such contributions.

REFERENCES

AT&T (1956). *Statistical Quality Control Handbook*, Western Electric Company.

Box, G.E.P. and Luceno, A. (1997). *Statistical Control by Monitoring and Feedback Adjustment*, John Wiley and Sons, New York.

Chakraborti S., Humanb, S. and Graham, M. (2008). Phase I Statistical Process Control Charts: An Overview and Some Results. *Quality Engineering*; 21:52–62

Champ, C.W., and Woodall, W.H. (1987). Exact Results for Shewhart Control Charts with Supplementary Runs Rules. *Technometrics*, 29(4):393–399.

Di Bucchianico A., and Van den Heuvel E. (2015). Shewhart's Idea of Predictability and Modern Statistics, in *Frontiers in Statistical Quality Control 11* (S. Knoth, W. Schmid, eds.). Springer International Publishing Switzerland, pp.237–248.

Frisén, M. (2003). Statistical surveillance. Optimality and methods. *International Statistical Review*, 71:403–434.

Frisén, M. (2011). Methods and evaluations for surveillance in industry, business, finance, and public health. *Quality and Reliability Engineering International*, 27:611–621.

Frisén, M. and de Maré, J. (1991). Optimal Surveillance, *Biometrika*, 78:271–280.

Frisén, M. and Sonesson, C. (2006). Optimal surveillance based on exponentially weighted moving averages. *Sequential Analysis*, 25:379–403.

Fuchs, C. and Kenett, R.S. (1987). Multivariate Tolerance Regions and F-tests. *Journal of Quality Technology*, 19:122–131.

Godfrey A.B. and Kenett R.S. (2007). Joseph M. Juran, a perspective on past contributions and future impact, Quality Reliability Engineering International, 23(6):653–663.

Goeb, R. (2006). Data Mining and Statistical Control—A Review and Some Links, in *Frontiers in Statistical Quality Control 8* (H-J. Lenz, P-Th. Wilrich – eds.). Springer-Verlag, Heidelberg, 285–308.

Gold, M. (1989). The Geometric Approximation to the Cusum Run Length Distribution, *Biometrika*, 76(4):725–733.

Hawkins, D., Qiu, P. and Kang, C. (2003). The Change-point Model for Statistical Process Control. *Journal of Quality Technology,* 35(4):355–65.

Hillier F.S. (1969). $\bar{X}$ and *R*- Chart Control Limits Based on a Small Number of Subgroups. *Journal of Quality Technology,* 1:17–26.

Kenett, R.S., Thyregod, P. (2006). Aspects of statistical consulting not taught by academia. *Statistica Neerlandica*, 60(3):396–412.

Kenett, R.S. (2016). On Generating High InfoQ with Bayesian Networks. *Quality Technology and Quantitative Management*, 13(3), in press.

Kenett, R.S. and Pollak, M. (1983). On Sequential Detection of a Shift in the Probability of a Rare Event. *Journal of the American Statistical Association*, 78:389–395.

Kenett, R.S. and Pollak, M. (1986). A Semi-Parametric Approach to Testing for Reliability Growth, With Application to Software Systems. *IEEE Transactions on Reliability*, R-35:304–311.

Kenett, R.S. and Pollak, M. (1996). Data-Analytic Aspects of the Shiryaev–Roberts Control Charts: Surveillance of a Non-Homogenous Poisson Process. *Journal of Applied Statistics,* 23:125–137.

Kenett, R.S. and Pollak, M. (2012). On Assessing the Performance of Sequential Procedures for Detecting a Change. *Quality and Reliability Engineering International*, 28:500–507.

Kenett, R.S. and Shmueli, M. (2014). On Information Quality. *Journal of the Royal Statistical Society, Series A* (with discussion), 177(1):3–38.

Kenett, R.S. and Shmueli, M. (2016). *Information Quality: The Potential of Data and Analytics to Generate Knowledge*, John Wiley and Sons.

Kenett, R.S. and Zacks, S. (1998). *Modern Industrial Statistics: Design and Control of Quality and Reliability*, Duxbury Press: Pacific Grove, CA, Spanish edition 2002, 2nd paperback edition 2002, Chinese edition 2004.

Kenett, R.S. and Zacks, S., with contributions by D. Amberti (2014). *Modern Industrial Statistics with Application in R, MINITAB and JMP*, John Wiley and Sons.

Knoth, S. (2006). The art of evaluating monitoring schemes—How to measure the performance of control charts? In H.-J. Lenz and P.-T. Wilrich, editors, *Frontiers in Statistical Quality Control* 8: 74–99. Physica-Verlag Heidelberg, http://cran.r-project.org/web/packages/spc.

Knoth, S. (2015). Run length quantiles of EWMA control charts monitoring normal mean or/and variance, *International Journal of Production Re-search*, 53;4629–4647.

Lai, T.L. (1995). Sequential Change-Point Detection in Quality Control and Dynamical Systems (with discussions). *Journal of Royal Statistical Society, Series B*, 57:613–658.

Lindsey, D. (2004). Some reflections on the current state of statistics, in *Applied Bayesian Statistics Studies in Biology and medicine*, di Bacco, M., d'Amore, G., Scalfari, F. (editors), Springer Verlag.

Lorden, G. (1971). Procedures for Reacting to a Change in Distribution. *Annals of Mathematical Statistics*, 42:1897–1908.

Lucas, J. (1976). The Design and Use of V-Mask Control Schemes. *Journal of Quality Technology*, 8: 1–12.

Lucas, J. and Crosier, R.B. (1982). Fast Initial Response for CUSUM Quality-Control Schemes: Give Your CUSUM a Head Start. *Technometrics*, 24:199–205.

Luceño, A. and Cofiño (2006). The Random Intrinsic Fast Initial Response of Two-Sided CUSUM Charts. *Sociedad de Estadıstica e Investigacion Operativa,* 15:505–524.

Mei, Y. (2008). Is Average Run Length to False Alarm Always an Informative Criterion? (with discussions). *Sequential Analysis*, 27:354–419.

Page, E.S. (1954). Continuous inspection schemes. *Biometrika*, 41:100–114.

Pollak, M. (1985). Optimal detection of a change in distribution, *Annals of Statistics* 13, 206–227.

Poor, V. (1988). Quickest detection with exponential penalty for delay. *Annals of Statistics*, 28: 2179–2205.

Shiryaev, A.N. (1961). The problem of the most rapid detection of a disturbance in a stationary process. *Soviet Math. Dokl.* 2:795–799.

Shiryaev, A.N. (2010). Quickest detection problems: Fifty years later. *Sequential Analysis*, 29: 345–385.

Shiryaev, A.N. (1963). On optimum methods in quickest detection problems. *Theory of Probability and Its Applications*, 8:22–46.

Weeze, M., Martinez, W., Megahed, F. and Jones-Farmer, L.A. (2015). Statistical Learning Methods Applied to Process Monitoring: An Overview and Perspective. *Journal of Quality Technology,* 48:4–27.

Woodall, W.H. and Montgomery, D.C. (1999). Research Issues and Ideas in Statistical Process Control. *Journal of Quality Technology,* 31:376–386.

Yashchin, E. (1985). On the analysis and design of CUSUM-Shewhart control schemes. *IBM Journal of Research and Development*, 29:377–391.

Zacks, S. (1981). The probability distribution and the expected value of a stopping variable associated with one-sided CUSUM procedures for non-negative integer valued random variables. *Communications Statistics A,* 10:2245–2258.

Zacks, S. (2004). Exact Determination of The Run Length Distribution of a One-Sided CUSUM Procedure Applied on An Ordinary Poisson Process. *Sequential Analysis*, 23:159–178.

Zacks, S. and Kenett, R.S. (1994). Process Tracking of Time Series with Change Points, in *Recent Advances in Statistics, Proceedings of the 4th international meeting of statistics in the Basque Country, San Sebastián*, Spain, 4–7 August, 1992. Utrecht: VSP, pp. 155–171.

This page intentionally left blank

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Two-factor hypothesis testing using the Gamma model

Nabendu Pal
Department of Mathematics, University of Louisiana at Lafayette, Louisiana, USA
Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Nuong Thi Thuy Tran & Minh-Phuong Tran
Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

ABSTRACT: We are familiar with the two-way Analysis Of Variance (ANOVA) using the normal model where we assume the additivity of the factor effects on the mean of the response variable assuming homoscedasticity (i.e., variances are all equal across the factor levels). But this type of normal set-up is not applicable in many problems, especially in engineering and biological studies where the observations are non-negative to begin with and likely to be positively skewed. In such situations one may use the Gamma model to fit the data, and proceed with further inferences. However, a normal type inference based on the decomposition of total Sum of Squares (SS) is not possible under the Gamma model, and further sampling distributions of the SS components are intractable. Therefore, we have looked into this problem from the scratch, and developed a methodology where one can test the effects of the factors. Our approach to tackle this interesting problem depends heavily on computations and simulation which bringa host of other challenges.

1 INTRODUCTION

Analysis Of Variance (ANOVA) with two factors is an important as well as powerful tool which is used to study the effects of two factors on the response variable. Suppose we have observations in the form of $\{X_{ijk}\}$ which indicates the *kth* observation under the influence of the *ith* level of Factor – 1 and *jth* level of Factor – 2, where $k = 1,2,\ldots,n_{ij}$. The standard statistical theory for two-factor ANOVA assumes the following linear additive model

$$X_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{1.1}$$

where ε_{ijk}'s are assumed to be independent and identically distributed as $N(0,\sigma^2)$. The remaining terms in (1.1) are the general effect $(=\mu)$, effect of the *ith* level of Factor – 1 $(=\tau_i)$, effect of the *jth* level of Factor – 2 $(=\beta_j)$, and the corresponding interaction term $(=\gamma_{ij})$. Due to the above normality and homoscedasticity (i.e., equality of variances for all *i* and *j*) assumptions the total Sum of Squares (SS), which measures the overall variability among the observations, can be decomposed into several independent components. These components can be used to construct the F-statistics to test relevant hypotheses regarding factor and/or interaction effects.

The classical two-way ANOVA theory does not work if the above assumptions fail to hold, although the practitioners may keep using it if the departures from assumptions are not significant. But in many engineering and biological studies the datasets tend to be positively skewed as well as nonnegative, and as a result the two standard assumptions may not hold. As a possible remedy one can transform the data using the well-known Box–Cox transformation. However, the standard interpretation of the variables get lost under such transformation, and as a result, the statistical inferences become difficult to relate to the original problem. To explain the situation further we use two real-life datasets given as follows.

Example 1.1: McDonald ((2014), *Handbook of Biological Statistics*, 3*rd* ed., pages 173–179) presented enzyme activity data for amphipod crustacean (Platorchestia platensis), classified in terms of gender and genotype, as given in Table 1.1.

Note that in Table 1.1 there are only 4 $(= n_{ij}$, for all *i* and *j*) observations for each combinations of Factor – 1 (genotype) and Factor – 2 (gender). Therefore, the standard tests for normality and/or homoscedasticity, which are asymptotic in nature, are not very effective. It would be prudent to consider the Gamma model, instead of the normal model, for a dataset with nonnegative observations. A two—parameter Gamma distribution is a more versatile model for nonnegative observations since the distribution can be extremely skewed (for small shape parameters) to almost symmetric (for large shape parameters).

Example 1.2: BOD (Biological Oxygen Demand) samples were taken from the Dongnai river basin in Vietnam. The BOD is being described by two factors: Location and Season. The Season factor has two levels – WET and DRY, and the Location factor has three levels – DN (Dongnai River), SG (Saigon River) and HL (Estuary). The dataset is given in Table 2.

In Table 2, the values of n_{ij}'s are different for every level of Factor – 1 (Location). Here we ignore the year (we combine the data for the years 2009 and 2010 into a single time period).

We are going to assume that the independent observations $X_{ijk}, (1 \le k \le n_{ij})$ follow Gamma $(\delta_{ij}, \beta_{ij})$ distribution, $1 \le i \le a$ and $1 \le j \le b$ with the pdf

$$f(x_{ijk} \mid \delta_{ij}, \beta_{ij}) = \frac{1}{\Gamma(\delta_{ij})\beta_{ij}^{\delta_{ij}}} \exp^{-x_{ijk}/\beta_{ij}} x_{ijk}^{\delta_{ij}-1}. \qquad (1.2)$$

The mean and variance of X_{ijk} are given as

$$\begin{aligned} &E(X_{ijk}) = \mu_{ij}(\text{say}) = \delta_{ij}\beta_{ij}, \\ &\text{and } V(X_{ijk}) = \sigma_{ij}^2(\text{say}) = \delta_{ij}\beta_{ij}^2. \end{aligned} \qquad (1.3)$$

To study whether the factor levels, individually or jointly, have any significant influence on the means, or whether the two factors have any interaction, we are going to consider the following four hypothesis testing problems.

Problem-1: Test $H_0^{(1)} : \mu_{ij} = \mu_{i\cdot} \forall j$ vs. $H_A^{(1)} : \mu_{ij} \ne \mu_{ij'}$, for some $j \ne j', 1 \le j, j' \le b$.

Problem-2: Test $H_0^{(2)} : \mu_{ij} = \mu_{\cdot j} \forall i$ vs. $H_A^{(2)} : \mu_{ij} \ne \mu_{i'j}$, for some $i \ne i', 1 \le i, i' \le a$.

Problem-3: Test $H_0^{(3)} : \mu_{ij} = \mu_{\cdot\cdot} \forall (i,j)$ vs. $H_A^{(3)} : \mu_{ij} \ne \mu_{i'j'}$, for some $i \ne i'$ and/or $j \ne j', 1 \le i, i' \le a, 1 \le j, j' \le b$.

Table 1. Enzyme activity data according to gender and genotype.

Gender Genotype	Male	Female
FF	1.884; 2.283; 4.939; 3.486;	2.838;4.216; 2.889; 4.198
FS	2.396; 2.956; 3.105; 2.649;	3.550; 4.556; 3.087; 1.943
SS	2.801; 3.421; 4.275; 3.110;	3.620; 3.079; 3.586; 2.669

Problem-4: Test $H_0^{(4)} : \mu_{ij} = \mu_{i\cdot} \times \mu_{\cdot j} \forall (i,j)$ vs. $H_A^{(4)} : \mu_{ij} \ne \mu_{i\cdot} \times \mu_{\cdot j}$ for some combinations of (i,j) .

If we fail to reject the null hypothesis (i) in Problem-1, then it implies that Factor – 2 has no influence on the mean response; (ii) in Problem – 2, then it implies that Factor – 1 has no influence on the mean response; (iii) in Problem – 3, then both the factors have no influence on the mean response when they act simultaneously; (iv) in Problem – 4, the joint interaction effect of the two factors is of multiplicative nature.

In this study, we are going to consider only Problem – 1 and Problem – 3, since by interchanging the roles of *i* and *j* (as well as *a* and *b*) we convert the Problem – 2 to Problem – 1. Problem – 4 will be taken up in a later study.

To keep the theory somewhat simpler we further assume that the scale parameters are all equal (but unknown), i.e.,

$$\beta_{ij} = \beta \forall (i,j). \qquad (1.4)$$

The most general case, i.e., where all scale parameters are unknown and possibly unequal will be considered in a later phase of our study. The above assumption (1.4) helps us in developing the ideas and concepts as well as computational tools which will be generalized rather easily for the next phase of our study.

Under the assumption of equal scale (i.e., (1.4)), the four hypothesis testing problems essentially boil down to studying the effects of the two factors on the shape parameters δ_{ij}'s only.

In Section – 2, we develop the test procedures for Problem – 1 (i.e., testing the significance of Factor – 2) followed by a comprehensive simulation. In Section – 3, we consider the test procedures for Problem 3 (i.e., testing the joint significance of Factor – 1 and Factor – 2). In each of the above two problems we first derive the Asymptotic Likelihood Ratio Test (ALRT). As we will see from our simulation, the ALRT performs poorly in maintaining the nominal level for small sample sizes, and hence an improvement is presented in terms of a Parametric Bootstrap (PB) version of the test based on the likelihood ratio statistic, henceforth called 'PBLRT'. In a series of recent papers (see Pal et al. (2007), Chang et al. (2008), Chang et al. (2010), Lin et al. (2015), it has been shown that the PBLRT works much better than the ALRT for many other problems where an exact optimal test either does not exist or hard to find due to a complicated sampling distribution. Although the PBLRT performs better than the ALRT in terms of maintaining the level (i.e., probability of type – I error) condition, especially for small to moderate sample sizes, it is heavily

Table 2. Data of BOD (Biological Oxygen Demand).

Season	DRY		WET	
Location	2009	2010	2009	2010
DN	7.0; 7.8; 13.0; 12.9; 14.7; 7.8; 8.0; 12.6	8.0; 7.7; 11.3; 13.2; 14.8; 7.9; 8.1; 12.7	5.9; 6.4; 9.0; 8.2; 9.6; 6.7; 6.8; 9.0	6.0; 6.5; 8.9; 8.1; 9.6; 5.8; 5.9; 8.1
SG	8.0; 9.5; 13.2; 13.0; 13.8; 26.8; 140.0; 17.2; 56.4; 23.8; 20.7; 162.0; 67.4; 17.5; 17.0; 11.3	8.0; 9.5; 13.2; 13.0; 13.8; 26.9; 141.0; 17.3; 56.5; 23.8; 18.9; 149.6; 59.9; 15.7; 18.8; 11.3	7.0; 8.1; 9.8; 12.4; 12.5; 21.9; 140.0; 15.5; 53.5; 20.8; 19.1; 132.0; 58.0; 15.6; 14.9; 9.2	7.3; 8.3; 10.1; 12.8; 12.9; 19.6; 122.0; 14.0; 47.3; 21.7; 18.2; 134.0; 59.5; 14.2; 15.5; 9.2
HL	10.4; 15.0; 8.0; 7.5; 7.1; 6.0; 7.6; 7.4; 6.3; 27.7; 22.9; 11.2	9.3; 15.2; 7.9; 7.6 7.4; 5.4; 7.7; 7.4; 6.3; 27.7; 29.9; 11.6	7.8; 8.7; 7.0; 6.2; 5.7; 4.4; 5.5; 5.7; 4.5; 19.1; 18.9; 9.4	6.8; 9.4; 6.5; 6.4; 5.8; 4.5; 5.6; 5.7; 4.5; 19.1; 18.9; 9.4

dependent on computations. But given the computational resources available today, implementation of PBLRT should not be any difficulty.

2 TESTING THE SIGNIFICANCE OF FACTOR – 2 (PROBLEM – 1)

Our goal in this section is to address the hypothesis testing problem $H_0^{(1)}: \mu_{ij} = \mu_{i.}, \forall j$ against the alternative which negates it. The null hypothesis is stating that Factor – 2 has no effect on the mean response, which under the assumption of equality of scales (i.e., (1.4)) can be written as

$$H_0^{(1)}: \delta_{ij} = \delta_{i.} \forall j, \text{for some suitable } \delta_{i.}; \tag{2.1}$$

where $\delta_{i.}$ can be thought as $(\mu_{i.}/\beta)$.

To test (2.1) we derive the classical Likelihood Ratio Test (LRT) statistic given as $\Lambda_* = (-2\ln\Lambda)$ where

$$\Lambda = \frac{\sup_{H_0^{(1)}} L}{\sup L}, \tag{2.2}$$

where L stands for the likelihood function of the combined data, the numerator in (2.2) stands for the restricted supremum of L under $H_0^{(1)}$, and the denominator in (2.2) represents the global supremum of L.

The standard asymptotic theory says that for all n_{ij} 'moderately large', the sampling distribution of Λ_* under $H_0^{(1)}$ can be approximated as

$$\Lambda_* = (-2\ln\Lambda) \sim \chi_\nu^2, \tag{2.3}$$

where the degrees of freedom ν is the difference between the number of free parameters in the global parameter space Θ and that under $H_0^{(1)}$. So, the LRT rejects $H_0^{(1)}$ if $\Lambda_* > \chi^2_{(\nu,(1-\alpha))} = (1-\alpha)100th$ percentile value of χ_ν^2 -distribution. But one must note that this test based on the Chi-square distribution is not very good (or accurate) when n_{ij}'s are 'small'. But first we are going to see the details of this LRT method.

Given the independent observations X_{ijk} ~ Gamma (δ_{ij}, β), the likelihood function L is given as

$$\begin{aligned} L &= L(\delta_{ij}, \beta, 1 \le i \le a, 1 \le j \le b \mid X_{ijk}, \forall(i,j,k)) \\ &= \prod_{i=1}^{a}\prod_{j=1}^{b}\prod_{k=1}^{n_{ij}}\left[\frac{1}{\Gamma(\delta_{ij})\beta^{\delta_{ij}}} \exp^{-X_{ijk}/\beta}(X_{ijk})^{\delta_{ij}-1}\right]. \end{aligned} \tag{2.4}$$

Thus, the log-likelihood function $L_* = \ln L$ can be written as

$$\begin{aligned} L_* = \sum_{i=1}^{a}\sum_{j=1}^{b}\{&-n_{ij}\ln\Gamma(\delta_{ij}) - n_{ij}\delta_{ij}\ln\beta - \\ &(1/\beta)\sum_{k=1}^{n_{ij}} X_{ijk} + (\delta_{ij}-1)\sum_{k=1}^{n_{ij}} \ln X_{ijk}\}. \end{aligned} \tag{2.5}$$

We use the notation $\bar{X}_{ij\cdot}$ and $\tilde{X}_{ij\cdot}$ to denote the Arithmetic Mean (AM) and Geometric Mean (GM) of the observations in the $(i,j)th$ cell, i.e.,

$$\bar{X}_{ij\cdot} = \sum_{k=1}^{n_{ij}} X_{ijk}/n_{ij}; \tilde{X}_{ij\cdot} = \left(\prod_{k=1}^{n_{ij}} X_{ijk}\right)^{1/n_{ij}}. \tag{2.6}$$

Then L_* can be simplified as

$$\begin{aligned} L_* = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\{&-\ln\Gamma(\delta_{ij}) - \delta_{ij}\ln\beta - (1/\beta)\bar{X}_{ij\cdot} \\ &+ (\delta_{ij}-1)\ln\tilde{X}_{ij\cdot}\}. \end{aligned} \tag{2.7}$$

By differentiating L_* in (2.7) w.r.t. δ_{ij} and β, and then setting them equal to zero yields the following system of equations

$$\psi(\delta_{ij}) + \ln\beta = \ln\widetilde{X}_{ij\cdot}, \forall(i,j);$$
$$\text{and}\quad \beta\Big(\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\delta_{ij}\Big) = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\bar{X}_{ij\cdot}; \tag{2.8}$$

where $\psi(c) = \{\partial \ln\Gamma(c)/\partial c\}$ is the di-gamma function defined at $c > 0$. Define the total sample size $n_{\cdot\cdot}$ and the grand mean $\bar{X}_{\cdot\cdot\cdot}$ as follows

$$n_{\cdot\cdot} = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij} \quad\text{and}\quad \bar{X}_{\cdot\cdot\cdot} = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\bar{X}_{ij\cdot}/n_{\cdot\cdot} \tag{2.9}$$

Then, solving the system of equations in (2.8), yields the MLEs of δ_{ij} and β, say $\hat{\delta}_{ij}$ and $\hat{\beta}$, as follows.

First obtain $\hat{\delta}_{ij}$ by solving the following system of $(a\times b)$ equations

$$\ln\left(\sum_{i_0=1}^{a}\sum_{j_0=1}^{b} n_{i_0 j_0}\hat{\delta}_{i_0 j_0}\right) - \psi(\hat{\delta}_{ij}) = \ln\left(\frac{n_{\cdot\cdot}\bar{X}_{\cdot\cdot\cdot}}{\widetilde{X}_{ij\cdot}}\right), \forall(i,j); \tag{2.10}$$

and then obtain $\hat{\beta}$ as

$$\hat{\beta} = (n_{\cdot\cdot}\bar{X}_{\cdot\cdot\cdot}) / \left(\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\hat{\delta}_{ij}\right). \tag{2.11}$$

Thus,

$$\sup L = L(\hat{\delta}_{ij}, \hat{\beta}, 1 \le i \le a, 1 \le j \le b \mid X_{ijk}, \forall(i,j,k)). \tag{2.12}$$

The log-likelihood function under $H_0^{(1)}$, henceforth denoted by $L_*^{0(1)}$, is

$$L_*^{0(1)} = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\{-\ln\Gamma(\delta_{i\cdot}) - \delta_{i\cdot}\ln\beta - (1/\beta)\bar{X}_{ij\cdot} + (\delta_{i\cdot} - 1)\ln\widetilde{X}_{ij\cdot}\}. \tag{2.13}$$

Differentiating $L_*^{0(1)}$ w.r.t. $\delta_{i\cdot}$ and β, and then setting them equal to zero yields

$$\sum_{j=1}^{b} n_{ij}\psi(\delta_{i\cdot}) = \sum_{j=1}^{b} n_{ij}\{\ln\widetilde{X}_{ij\cdot} - \ln\beta\}, \forall i;$$
$$\text{and}\quad \beta\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\delta_{i\cdot} = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\bar{X}_{ij\cdot} \tag{2.14}$$

Define the total sample size subject to *ith* level of Factor – 1, and the corresponding sampling proportion as

$$n_{i\cdot} = \sum_{j=1}^{b} n_{ij}, \text{ and } v_{ij} = n_{ij}/n_{i\cdot} \tag{2.15}$$

The MLEs of $\delta_{i\cdot}$ and β under $H_0^{(1)}$, denoted by $\hat{\delta}_{i\cdot}^{0}$ and $\hat{\beta}^{0}$, are obtained as follows. First obtain $\hat{\beta}^{0}$ by solving the following system of a equations

$$\ln\left(\sum_{q=1}^{a} n_{q\cdot}\hat{\delta}_{q\cdot}^{0}\right) - \sum_{j=1}^{b} v_{ij}\psi(\hat{\delta}_{i\cdot}^{0}) = \ln\left(\sum_{q=1}^{a}\sum_{l=1}^{b} n_{ql}\bar{X}_{ql\cdot} \Big/ \prod_{j=1}^{b}\widetilde{X}_{ij\cdot}^{v_{ij}}\right) \tag{2.16}$$

and then obtain $\hat{\beta}^{0}$ as

$$\hat{\beta}^{0} = (n_{\cdot\cdot}\bar{X}_{\cdot\cdot\cdot}) \Big/ \left(\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\hat{\delta}_{i\cdot}^{0}\right). \tag{2.17}$$

Thus,

$$\sup_{H_0^{(1)}} L = L(\hat{\delta}_{i\cdot}^{0}, \hat{\beta}^{0}, 1 \le i \le a \mid X_{ijk}\forall(i,j,k)). \tag{2.18}$$

As stated earlier, for 'moderately large' n_{ij} values, Λ_* follows χ_ν^2 under $H_0^{(1)}$, with $\nu = a(b-1)$. We will see later that for small n_{ij}'s, the size of the ALRT is higher than α whereas the proposed PBLRT keeps it within α. The beauty of the PBLRT is that it is a purely computational technique where one does not need to know the sampling distribution of the test statistic (which is the LRT statistic in this case), and the critical value is derived through a simulation. Before discussing further about the pros and cons of the PBLRT, we first describe how it is implemented through a series of steps as given below.

Steps of the proposed PBLRT

Step – 1: Given the original data $\{X_{ijk}, \forall(i,j,k)\}$, obtain the unrestricted MLEs $(\hat{\delta}_{ij}, \hat{\beta})$ as well as restricted MLEs $(\hat{\delta}_{i\cdot}^{0}, \beta^{0})$ (under $H_0^{(1)}$). Compute Λ_* using (2.12) and (2.18).

Step – 2:

i. Assuming that $H_0^{(1)}$ is true, generate artificial (bootstrap) observations in an internal loop of M replications. In the *mth* replications we generate $X_{ijk}^{(m)}$ from Gamma $(\hat{\delta}_{i\cdot}^{0}, \hat{\beta}^{0}), 1 \le k \le n_{ij}, 1 \le j \le b, 1 \le i \le a.$

ii. With the artificial observations $\{X_{ijk}^{(m)}, \forall(i,j,k)\}$ compute $(\hat{\delta}_{ij}, \hat{\beta})$ and $(\hat{\delta}_{i\cdot}^{0}, \hat{\beta}^{0})$ as done in Step – 1, and call them $(\hat{\delta}_{ij}^{(m)}, at\beta^{(m)})$ and $(\hat{\delta}_{i\cdot}^{0(m)}, \hat{\beta}^{0(m)})$, respectively. Then obtain Λ_* value as done in Step - 1, and call it Λ_*^{m}.

iii. By repeating above (i)–(ii) for $m = 1, 2, \cdots, M$, we have $\Lambda_*^{1}, \Lambda_*^{2}, \cdots, \Lambda_*^{M}$. Order these Λ_*^{m} values as $\Lambda_*^{(1)} \le \Lambda_*^{(2)} \le \cdots \le \Lambda_*^{(M)}$.

Step – 3: The critical value for the statistic Λ_* (in Step – 1) is obtained as $\Lambda_*^{((1-\alpha)M)}$, where α is the level of the test. If $\Lambda_* > \Lambda_*^{((1-\alpha)M)}$, then reject $H_0^{(1)}$; retain $H_0^{(1)}$ if otherwise. Alternatively, the p-value of PBLRT is approximated by $\sum_{m=1}^{M} I(\Lambda_*^{(m)} > \Lambda_*)/M$.

Remark 2.1: Why the PBLRT might work better than the ALRT is not counter intuitive. The ALRT approximates the true distribution of Λ_* under $H_0^{(1)}$ by the Chi-square distribution which may be far from reality when n_{ij}*'s* are not large. On the other hand, the proposed PBLRT tries to replicate the true distribution of Λ_* under $H_0^{(1)}$ by drawing samples from Gamma $(\hat{\delta}_{i\cdot}^{0}, \hat{\beta}^{0})$ which is an approximation to the distribution Gamma (δ_{ij}, β) under $H_0^{(1)}$. Thus, the relative frequency histogram of $\Lambda_*^{(m)}, 1 \le m \le M$, comes very close to that of Λ_* under $H_0^{(1)}$, for large *M*, and it appears to be a better fit, as the simulation results indicate, than the χ_ν^2 distribution.

Remark 2.2: In order to compare the proposed PBLRT with ALRT in terms of size and power, we have undertaken a comprehensive simulation study. In our simulation, we generate the dataset $\{X_{ijk}, \forall(i,j,k)\}$ a large number (say, *Q*) times. In each replication we observe whether the test under consideration rejects the null hypothesis or not. Then the size or power of the test is approximated by the proportion of times (out of *Q*) it rejects $H_0^{(1)}$. When our input parameters (δ_{ij}, β), i.e., the parameters used to generate the observations $\{X_{ijk}, \forall(i,j,k)\}$, obey $H_0^{(1)}$, then the proportion of times a test rejects the null hypothesis becomes the estimated size of that test. When the input parameters do not obey the null hypothesis, then we obtain the estimated power of the test. To be specific, in every replication of the data $\{X_{ijk}, \forall(i,j,k)\}$, say in the *qth* replication, we define $I_{ALRT}^{(q)}$ and $I_{PBLRT}^{(q)}$ as $I_{ALRT}^{(q)} = 1$ if ALRT rejects $H_0^{(1)}$, $I_{ALRT}^{(q)} = 0$, otherwise; and $I_{PBLRT}^{(q)} = 1$ if PBLRT rejects $H_0^{(1)}$, $I_{PBLRT}^{(q)} = 0$, otherwise. Then, depending on the input parameters, (size or power of ALRT) = $\sum_{q=1}^{Q} I_{ALRT}^{(q)}/Q$, and (size or power of PBLRT) = $\sum_{q=1}^{Q} I_{PBLRT}^{(q)}/Q$.

Remark 2.3: The utility of the proposed PBLRT lies in its simplicity. One does not need to know the true sampling distribution of the test statistic Λ_*. However, it is very computation intensive. In our simulation study, while the size (or power) of ALRT is computed through a single loop (of *Q* replications), that of PBLRT is done through a double loop (of *Q* replications in the outer loop and *M* replications in the inner loop). As a result, running the simulation study becomes a challenge in terms of computational time. But in real-life applications where a decision has to be made, based on the PBLRT, whether to reject the null hypothesis or not, then that decision-making process is not that time consuming, since it is done through a single loop (the inner loop of *M* replications only to find the critical value for the test statistic). In Section 5, where four datasets have been analyzed, we have used M = 10,000.

The next section is devoted to size comparison of the two tests mentioned above.

3 COMPARISON OF ALRT AND PBLRT IN TERMS OF SIZE

For size comparison, the datasets are generated under the null hypothesis, i.e., $\delta_{ij} = \delta_{i\cdot}, \forall j$, for some $\delta_{i\cdot} > 0$. Not only we are going to vary $\delta_{i\cdot}$ but also n_{ij}*'s* as well as the nominal level α. Three widely used α values will be used, which are 0.01, 0.05, 0.10. We have noted that $M = Q = 5000$ gives quite

Table 3. Simulated size of two tests with $a = b = 2$, $\delta_{1\cdot} = \beta = 1.0$ under $H_0^{(1)}$.

		$\delta_{2\cdot} = 0.5$		$\delta_{2\cdot} = 1.0$		$\delta_{2\cdot} = 2.0$		$\delta_{2\cdot} = 5.0$		$\delta_{2\cdot} = 10.0$	
α	n_{ij}	P_{ALRT}	P_{PBLRT}	P_{ALRT}	P_{PBLRT}	P_{ALRT}	P_{PBLRT}	P_{ALRT}	P_{PBLRT}	P_{ALRT}	P_{PBLRT}
0.01	5	0.019	0.011	0.018	0.011	0.021	0.013	0.022	0.014	0.024	0.016
	10	0.015	0.009	0.015	0.008	0.014	0.007	0.017	0.009	0.015	0.006
	25	0.011	0.006	0.012	0.007	0.013	0.005	0.017	0.006	0.011	0.006
	50	0.012	0.005	0.012	0.006	0.010	0.005	0.011	0.006	0.010	0.005
0.05	5	0.076	0.047	0.075	0.047	0.079	0.053	0.091	0.058	0.088	0.054
	10	0.076	0.045	0.071	0.046	0.069	0.039	0.070	0.042	0.070	0.040
	25	0.051	0.029	0.055	0.031	0.053	0.030	0.064	0.035	0.054	0.029
	50	0.059	0.035	0.055	0.032	0.050	0.025	0.054	0.028	0.049	0.026
0.10	5	0.142	0.097	0.142	0.093	0.144	0.100	0.154	0.110	0.156	0.113
	10	0.133	0.093	0.129	0.086	0.127	0.081	0.129	0.082	0.126	0.082
	25	0.106	0.063	0.110	0.071	0.104	0.063	0.118	0.072	0.111	0.061
	50	0.109	0.071	0.110	0.068	0.106	0.061	0.108	0.060	0.098	0.054

stable results with standard error bounded above by 0.003. For convenience we use $n_{ij} = n \quad \forall(i, j)$, and n is taken as 5, 10, 25, 50. The overall two-factor problem size, identified with (a, b), will be varied as (2,2), (5,5), (10,10); but in Table 3 we report the results for a = b = 2 only. We use the notation P_{ALRT} and P_{PBLRT} to denote the simulated size of ALRT and PBLRT respectively.

Remark 3.1: It is noted that the size of the ALRT is much higher than the nominal level α when $n < 25$, making the test very liberal. For $n \geq 25$, the ALRT's size stays very close to α, albeit a bit higher. On the other hand, the proposed PBLRT's size is always within α. In fact for $n \geq 25$, the PBLRT behaves like a more conservative test. Therefore, as a rule of thumb, we suggest that the PBLRT be used for "small" sample sizes, and the ALRT be used for "large" sample sizes.

4 TESTING THE JOINT SIGNIFICANCE OF FACTOR – 1 AND FACTOR – 2

Our goal in this section is to consider the Problem - 3, with the hypothesis testing of $H_0^{(3)}: \mu_{ij} = \mu_{..} \forall(i,j)$ against the alternative which negates it. The null hypothesis is stating that both the factors have no influence on the mean response when they act simultaneously, which under the assumption of equality of scales $(\beta_{ij} = \beta, \forall(i,j))$ can be written as

$$H_0^{(3)}: \delta_{ij} = \delta_{..} \forall(i,j) \text{ for some suitable } \delta_{..}; \tag{4.1}$$

where $\delta_{..}$ can be thought as $(\mu_{..} / \beta)$. Similar to Section 2, the likelihood ratio test statistic is given as

$$\Lambda_* = -2 \ln \frac{\sup_{H_0^{(3)}} L}{\sup L} \tag{4.2}$$

With all n_{ij} 'moderately large', we can approximate the sampling distribution of Λ_* under $H_0^{(3)}$ as

$$\Lambda_* \sim \chi^2_\nu,$$

where $\nu(= ab - 1)$ is the degrees of freedom. Next we are going to see the details of the LRT method for Problem – 3. We also consider the log-likelihood function under $H_0^{(3)}$, denoted by $L_*^{0(3)}$, as

$$L_*^{0(3)} = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij} \{-\ln \Gamma(\delta_{..}) - \delta_{..} \ln \beta - (1/\beta) \bar{X}_{ij\cdot} + (\delta_{..} - 1) \ln \widetilde{X}_{ij\cdot}\}. \tag{4.3}$$

Table 4. Simulated size of two tests with $a = b = 2$; $\delta_{1.} = \delta_{2.} = 1.0 = \beta$ under $H_0^{(3)}$.

α	0.01		0.05		0.10	
n_{ij}	P_{ALRT}	P_{PBLRT}	P_{ALRT}	P_{PBLRT}	P_{ALRT}	P_{PBLRT}
5	0.020	0.004	0.082	0.034	0.140	0.056
10	0.012	0.005	0.061	0.029	0.115	0.063
25	0.010	0.007	0.053	0.030	0.106	0.067
50	0.011	0.008	0.052	0.033	0.096	0.065

By differentiating $L_*^{0(3)}$ w.r.t. $\delta_{..}$ and β, and then setting them equal to zero yields

$$n_{..} \psi(\delta_{..}) = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij} (\ln \widetilde{X}_{ij\cdot} - \ln \beta); \quad \text{and} \quad \beta \delta_{..} = \bar{X}_{...} \tag{4.4}$$

The MLEs of $\delta_{..}$ and β under $H_0^{(3)}$, denoted by $\hat{\delta}_{..}^0$ and $\hat{\beta}^0$, are obtained as follows. First obtain $\hat{\delta}_{..}^0$ by solving the following equation

$$\psi(\hat{\delta}_{..}^0 - \ln\left(\hat{\delta}_{..}^0 n_{..}\right) = \frac{\sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij} \ln \widetilde{X}_{ij\cdot}}{n_{..}} - \ln(n_{..} \bar{X}_{...}) \tag{4.5}$$

and then obtain $\hat{\beta}^0$ as

$$\hat{\beta}^0 = \frac{\bar{X}_{...}}{\hat{\delta}_{..}^{(0)}}. \tag{4.6}$$

Thus,

$$\sup_{H_0^{(3)}} L = L(\hat{\delta}_{..}^0, \hat{\beta}^0 \mid X_{ijk} \forall(i,j,k)). \tag{4.7}$$

Similar to Section 2, a PBLRT version test can be constructed based on the LRT statistic. A comprehensive simulation has been carried out, and it has been noted that the PBLRT adheres to the level much closer than the ALRT. Table 4 shows the simulated size values of the two tests (ALRT and PBLRT).

5 ANALYSIS OF DATASETS

In this section we revisit the two datasets presented in Section 1 and see two other datasets. These datasets have been used as demonstration purposes for our proposed gamma distribution based analysis of two factors.

Example 5.1: Recall the dataset of McDonald (2014) presented in Example 1.1. The following Table 5 provides the results of the usual normal distribution based ANOVA with interaction.

The following Table 6 presents the analysis of the dataset in Example 1.1 under the gamma model.

Remark 5.1: Note that under the gamma model both the ALRT as well as PBLRT retain the null hypothesis negating the effect of the two factors on enzyme activity. The results are consistent with the findings under the normal model, though the corresponding p-values are slightly different.

Example 5.2: Recall the dataset presented in Example 1.2. Similar to the previous example the following two tables (Table 7 and Table 8) show the results under the normal as well as the gamma models. Here also the final inferences are consistent with each other.

In the following sequel we present two new datasets which show the divergence in outcomes of the two approaches (one based on the normal model, and the other based on the gamma model).

Example 5.3: Montgomery ((2005), *Design and Analysis of Experiments*, 6*th* ed., page 201) cites an experiment, similar to a study reported in an article in the *IEEE Transactions on Electronic Devices* (Nov. 1986, page - 1754), where the response variable, that is Base Current (BC), was observed subject to various levels of two factors: Polysilicon Doping (ions) and Anneal Temperature (degree Centigrade) as shown below.

The following Table 10 and Table 11 show the normal based ANOVA results as well as those based on the gamma model.

Remark 5.2: Take a look at the Table 9. While the BC values clearly differ greatly for the levels of AT, they do not vary much for the levels of PD. If one uses the level $\alpha = 0.01$, then the factor PD is not significant under the normal model. However, under the gamma model, PD is clearly significant (along with AT).

Example 5.4: Hogg and Ledolter ((1987), *Engineering Statistics*, page - 238) reported a study done by a textile engineer regarding the effect of temperature (degree Fahrenheit) and time (in cycles) on the brightness of a synthetic fabric which uses a particular dye. Brightness was measured on a 50-point scale, and three observations were taken at each combination of temperature and time as shown in the following Table 12.

Table 5. Results for the Enzyme data (normal model).

Source	F-value	p-value
Genotype	0.332	0.722
Gender	0.489	0.493
Interaction	0.351	0.709

Table 6. Results for the Enzyme data (gamma model).

Factor	P_{ALRT}	P_{PBLRT}
Genotype	0.613	0.847
Gender	0.647	0.781

Table 7. Results for the BOD data (normal model).

Source	F-value	p-value
Location	14.460	0.000
Season	0.558	0.456
Interaction	0.011	0.992

Table 8. Results for the BOD data (gamma model).

Factor	P_{ALRT}	P_{PBLRT}
Location	0.000	0.000
Season	0.660	0.693

Table 9. BC dataset according to factors PD and AT.

Polysilicon Doping (PD)	Anneal Temperature (AT)		
	900	950	1000
1×10^{20}	4.60, 4.40	10.15, 10.20	11.01, 10.58
2×10^{20}	3.20, 3.50	9.38, 10.02	10.81, 10.60

Table 10. Results for the BC data (normal model).

Source	F-value	p-value
PD	10.216	0.019
AT	648.906	0.000
Interaction	3.474	0.0995

Table 11. Results for the BC data (gamma model).

Factor	P_{ALRT}	P_{PBLRT}
PD	0.000	0.000
AT	0.000	0.000

Table 12. Brightness dataset according to Time and Temp.

Time (in cycles)	Temperature (Temp) (in Fahrenheit)		
	350	375	400
40	38, 32, 30	37, 35, 40	36, 39, 43
50	40, 45, 36	39, 42, 46	39, 48, 47

Table 13. Results for the Brightness data (normal model).

Source	F-value	p-value
Time	9.692	0.009
Temperature	2.606	0.115
Interaction	0.111	0.896

Table 14. Results for the Brightness data (gamma model).

Factor	P_{ALRT}	P_{PBLRT}
Time	0.027	0.030
Temperature	0.131	0.300

As before, the Table 13 and Table 14 summarize the findings under the normal as well as the gamma models.

Remark 5.3: Interestingly, while the normal model indicates 'Time' as a significant factor for brightness for any α, the gamma model infers it as insignificant using $\alpha = 0.01$. Also, under the gamma model, PBLRT differs greatly from the ALRT in terms of the p-value for the factor 'Temperature'.

Concluding Remark: This work sheds some light on an approach alternative to the normal model which is widely used to analyze a dataset subject to two factors. Our proposed gamma model, and the corresponding PBRLT can be used effectively for nonnegative datasets, especially when the sample sizes are small and/or the normality assumption fails to hold.

ACKNOWLEDGMENT

We sincerely thank Mr. Huynh Van Kha of the Faculty of Mathematics and Statistics, Ton Duc Thang University, for his generous help with the numerical computations of this work.

We are indebted to our colleague Dr. Pham Anh Duc of the Faculty of Environment and Labour Safety, Ton Duc Thang University, for providing the BOD dataset.

REFERENCES

Chang, C.-H., J.-J. Lin, & N. Pal (2011). Testing the equality of several gamma means: a parametric bootstrap method with applications. *Computational Statistics 26*(1), 55–76.

Chang, C.-H. & N. Pal (2008). Testing on the common mean of several normal distributions. *Computational Statistics & Data Analysis 53*(2), 321–333.

Chang, C.-H., N. Pal, & J.-J. Lin (2010). A note on comparing several poisson means. *Communications in Statistics-Simulation and Computation 39*(8), 1605–1627.

Lin, J.-J., C.-H. Chang, & N. Pal (2015). A Revisit to Contingency Table and Tests of Independence: Bootstrap is Preferred to Chi-Square Approximations as well as Fishers Exact Test. *Journal of biopharmaceutical statistics 25*(3), 438–458.

Pal, N., W.K. Lim, & C.-H. Ling (2007). A computational approach to statistical inferences. *Journal of Applied Probability & Statistics 2*(1), 13–35.

Applied Mathematics in Engineering and Reliability – Briš, Snášel, Khanh & Dao (Eds)
© 2016 Taylor & Francis Group, London, ISBN 978-1-138-02928-6

Author index

This page intentionally left blank

Applied Mathematics in Engineering and Reliability contains papers presented at the International Conference on Applied Mathematics in Engineering and Reliability (ICAMER 2016, Ho Chi Minh City, Viet Nam, 4-6 May 2016). The book covers a wide range of topics within mathematics applied in reliability, risk and engineering, including:

- Risk and Reliability Analysis Methods
- Maintenance Optimization
- Bayesian Methods
- Monte Carlo Methods for Parallel Computing of Reliability and Risk
- Advanced Mathematical Methods in Engineering
- Methods for Solutions of Nonlinear Partial Differential Equations
- Statistics and Applied Statistics, etc.

The application areas range from Nuclear, Mechanical and Electrical Engineering to Information Technology and Communication, Safety Engineering, Environmental Engineering, Finance to Health and Medicine. The papers cover both theory and applications, and are focused on a wide range of sectors and problem areas. Integral demonstrations of the use of reliability and engineering mathematics are provided in many practical applications concerning major technological systems and structures.

Applied Mathematics in Engineering and Reliability will be of interest to academics and professionals working in a wide range of industrial, governmental and academic sectors, including Electrical and Electronic Engineering, Safety Engineering, Information Technology and Telecommunications, Civil Engineering, Energy Production, Infrastructures, Insurance and Finance, Manufacturing, Mechanical Engineering, Natural Hazards, Nuclear Engineering, Transportation, and Policy Making.

ISBN 978-1-138-02928-6

9 781138 029286

an **informa** business

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
Schipholweg 107C
2316 XC Leiden, NL
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK