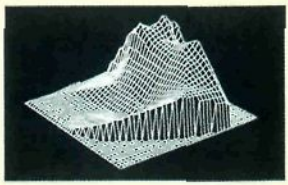
Proceedings of Statistics 2001 Canada
The 4th Conference in Applied Statistics



RECENT ADVANCES IN STATISTICAL METHODS

edited by **Yogendra P. Chaubey**

Imperial College Press

RECENT ADVANCES IN STATISTICAL METHODS

edited by **Yogendra P. Chaubey**
Concordia University, Canada

RECENT ADVANCES IN STATISTICAL METHODS

Proceedings of Statistics 2001 Canada
The 4th Conference in Applied Statistics

Montreal, Canada 6 – 8 July 2001

Imperial College Press

Published by

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

Distributed by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: Suite 202, 1060 Main Street, River Edge, NJ 07661
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

RECENT ADVANCES IN STATISTICAL METHODS
Proceedings of Statistics 2001 Canada: The 4th Conference in Applied Statistics

ISBN 1-86094-333-0

This book is printed on acid-free paper.

Printed in Singapore by Mainland Press

PREFACE

The present volume consists of a selection of 31 papers which were presented at the conference *Statistics 2001 Canada: Fourth Canadian Conference in Applied Statistics* held at Concordia University, Montreal, Quebec, Canada during July 6-8, 2001. The Conference attracted approximately 250 participants from all over the globe and featured 144 speakers in seven plenary sessions and 42 invited and contributed papers sessions.

This conference was held as a sequel to the ones previously held at Concordia University in 1971, 1981 and 1991 as the first, second and third Canadian conferences in applied Statistics, all under the chairship of Professor T.D. Dwivedi. The 1971 conference was initiated with local Statistics groups in Montreal and Ottawa, and led to the creation of the Statistical Society of Canada. The next two conferences were held under the joint sponsorship of the two departments: Mathematics & Statistics and DS & MIS. The Conference brought eminent academics and practitioners in the discipline of Statistics from all over the world, a majority mainly from Canada and USA. The tradition of these conferences has become a permanent feature in the minds of the Statistics community on the Canadian scene - to hold such conferences every 10 years in order to assess the existing techniques and new directions in Applied Statistics.

The success of the Conference was summarized by one of the Plenary Speakers, in an email to me: "This is just a note to express my pleasure in attending the highly successful conference you have arranged. You indeed did a splendid job." Here, "you" must be understood in a plural sense. The task of this magnitude required help from many people, and the organizers are thankful to all those who helped. The success of the Conference is a pride for the team on the Organizing Committee, Scientific Committee and Student Volunteers.

The papers included in this volume have gone through serious refereeing process. The editorial contribution provided by the members of the Editorial Board and many referees is highly appreciated. Furthermore, I am very thankful to all the authors for submitting their papers and continued cooperation. I would also like to thank Mr. Anthony Doyle of World Scientific Publishing (UK) Ltd. for enthusiastically supporting this project.

I sincerely hope that this volume will prove to be an important research resource for the scientific community.

Yogendra P. Chaubey

Concordia University
Montreal, May 2002

This page is intentionally left blank

CONTENTS

Preface v

On the Estimation of Size and Mean Value of a Stigmatized Characteristic of a Hidden Gang in a Finite Population
R. Arnab and S. Singh 1

Unemployment, Search and the Gender Wage Gap: A Structural Model
C. Belzil and X. Zhang 12

Kullback-Leibler Optimization of Density Estimates
A. Berlinet and E. Brunel 31

The Asymptotic Distribution of Spacings of Order Statistics
M. G. Bickis 42

Theoretical and Computational Issues in Bayesian Analysis of Multivariate Ordinal Categorical Data with Reference to an Ophthalmologic Study
A. Biswas 50

Second-Order Moments and Mutual Information in the Analysis of Time Series
D. R. Brillinger 64

Spatial Association Between Community Air Pollution and Heart Disease: Analysis of Correlated Data
S. Cakmak, R. Burnett, M. Jerrett, M. S. Goldberg, Arden Pope III and R. Ma 77

On the Robustness of Relative Surprise Inferences to the Choice of Prior
M. Evans and T. Zou 91

Using Survival Analysis in Pretern Birth Study
C. Y. Fu and S. H. Liu 107

Asymptotic Forms and Bounds for Tails of Convolutions of Compound Geometric Distributions, with Applications
J. Cai and J. Garrido 114

Improved Finite-Sample Inference in Overidentified Models with Weak Instruments
N. Gospodinov 132

Probability of Correct Selection of Gamma versus GE or Weibull versus GE based on Likelihood Ratio Test
R. D. Gupta, D. Kundu and A. Manglick 147

Survey Methodology for Studying Substance Use Prevention Programs in Schools
S. M. Jones, B. C. Sutton and K. E. Boyle 157

One-Step Estimation for the Partially Linear Proportional Hazards Model
X. Lu and R. S. Singh 169

A Nested Frailty Survival Model for Recurrent Events
R. Ma, J. D. Willms and R. T. Burnett 186

Multiple Comparison Procedures for Linear Models under Order Restrictions
H. Mansouri and R. Paige 198

Testing Goodness-of-Fit of the Gamma Models
C. E. Marchetti, G. S. Mudholkar and G. E. Wilding 207

On Frailty Models and Copulas
D. Oakes 218

Surrogate Data and Fractional Brownian Motion
P. Rabinovitch 225

Analysis of Occult Tumour Trial Data with Varying Lethality
S. N. Rai, J. Sun and D. Hunt 233

Nonlinear Mixed Effects Models: Recent Developments
P. S. R. S. Rao and N. Zaino 252

The Sampling Bias of Heckman's Sample Bias Estimator
P. Rilstone and A. Ullah 263

Computational Sequence Analysis: Genomics and Statistical Controversies
P. K. Sen 274

Universal Optimality of Completely Randomized Designs
K. R. Shah and B. K. Sinha 290

A Nonparametric Comparison of Tumor Incidences with Intermediate Lethality and Different Death Rates
J. Sun, Q. Zhao and S. N. Rai 296

Generalized Smoothed Estimating Functions with Censored Observations
A. Thavaneswaran and J. Singh 304

Large Sample Asymptotic Properties of Ordinary Least Squares, Two Stage Least Squares and Limited Information Maximum Likelihood Estimators in Simultaneous Equations Models
R. Tiwari and V. K. Srivastava 312

Relative Stability of Weighted Maxima of Bounded i.i.d. Random Variables
R. J. Tomkins 324

Transient Analysis of Some Infinite Server Queues
G. E. Willmot and S. Drekic 329

Empirical Likelihood Method for Finite Populations
C. Wu 339

Second Order Estimating Equations for Clustered Longitudinal Binary Data with Missing Observations
G. Y. Yi and R. J. Cook 352

ON THE ESTIMATION OF SIZE AND MEAN VALUE OF A STIGMATIZED CHARACTERISTIC OF A HIDDEN GANG IN A FINITE POPULATION

RAGHUNATH ARNAB
Department of Statistics, University of Durban-Westvile, Durban-4000, South Africa

SARJINDER SINGH
School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada
E-mail: sarjinder@yahoo.com

Arnab and Singh [1] developed theoretical formulae for estimation of size and the mean value of a sensitive character of a sub group (such as hidden gang) of a finite population using a randomized response survey in a unified set-up. In this article expressions for the estimators of the population characteristics and their variances are obtained for various sampling strategies used in practice. Performance of the proposed estimators are compared using numerical studies.

1 Introduction

In the collection of data of a sensitive nature, like induced abortion, drug addiction, suffering from AIDS *etc.* directly from the respondents, the respondents very often report untrue value and even refuse to answer. Warner [8] introduced an ingenious method known as randomized response (RR) technique for collection of data of a stigmatized nature by protecting confidentiality of respondents and produced an unbiased estimator for the proportion of persons belonging to a certain sensitive group under simple random sampling with replacement (SRSWR). This technique was extended for quantitative characteristics and or other sampling designs by various authors: see Chaudhuri and Mukherjee [3] and Arnab and Singh [1]. Singh, Horn and Chowdhury [7] were the first who used RR technique to estimate population parameters for both the qualitative and quantitative characters.

Arnab and Singh [1] extended the method of Singh, Horn and Chowdhury [7] for SRSWR for an arbitrary sampling design. In the present investigation we have studied the performance of a few well-known sampling strategies which may be used in estimating parameters of two population characteristics in practice. Efficiencies of the proposed strategies are compared on the basis of simulation studies and a real survey data.

2 Formulation of the Problem

Suppose that a finite population P consists of N (known) identifiable units. Let N_G (unknown) denote the total number of persons belonging to some sensitive group (Gang) $G(\subset P)$ and x_i be the value of the stigmatized quantitative character X under study for the $i-$th person in the Gang G. Arnab and Singh [1] considered the following sampling strategies for estimation of proportion $\pi = \frac{N_G}{N}$ and the mean $\mu_x = \sum_{i=1}^{N} x_i/N$ of the sensitive character$X$ for elements of the population P belonging to the subgroup G in a unified setup.

From the population P, two independent samples s_1 and s_2 of sizes n_1 and n_2 are selected by using some suitable sampling design p_1 and p_2 respectively. If the respondent labeled i, selected in the sample $s_k(k = 1, 2)$ belongs to the sensitive group G then he or she has to disclose the true value x_i. On the other hand if the respondent does not belong to the sensitive group G the respondent has to perform certain randomized response trial using a suitable randomized device. It is expected that the chosen randomized device D_k should produce a value similar in range to the confidential character X for generating more co-operation from the respondents.

Thus each of respondents selected in the sample will supply either the true value of X or a number obtained by randomized device D_k. The confidentiality of the respondent is maintained since the interviewer will not know whether the respondent is supplying true value or a value generated by the randomized device. The mean and variance of the randomized device D_k are assumed to be known to the interviewer and will be denoted by θ_k and σ_k^2 respectively. Thus for the i-th respondent, included in the sample $s_k, k = 1, 2$, we obtain a randomized response:

$$z_{ki} = \begin{cases} x_i, & \text{if } i \in G \\ R_{ki}, & \text{if } i \notin G \end{cases} \tag{1}$$

where x_i and R_{ki} denote respectively the true value of the stigmatized character X and the response obtained by the randomized device. The above response can be written as

$$z_{ki} = x_i I_i + (1 - I_i)R_{ki} = x_i I_i + I_i' R_{ki} \tag{2}$$

where

$$I_i = \begin{cases} 1, & \text{if } i \in G \\ 0, & \text{if } i \notin G \end{cases} \tag{3}$$

with $I_i' = 1 - I_i$. Denoting for $E_R(V_R)$ as expectation (variance) with respect to the randomize device, we get

$$E_R(z_{ki}) = x_i I_i + I_i^{'}\theta_k = \gamma_i \text{ (say)} \tag{4}$$

and

$$V_R(z_{ki}) = I_i^{'} V_R(R_{ki}) = I_i^{'}\sigma_k^2. \tag{5}$$

Conditional on the randomized device D_k, let the population mean of $z-$values be given $\bar{Z}_k = \sum_{i=1}^{N} z_{ki}/N$. By using the standard results in finite population survey sampling, it may be estimated by a linear homogeneous unbiased estimator given by

$$T_k = \sum_{i \in s_k} b_{s_k i} z_{ki}/N \tag{6}$$

where $b_{s_k i}$ are known constants satisfying the following design unbiasedness condition $\sum_{s_k \ni i} b_{s_k i} p_k(s_k) = 1$ with $p_k(s_k)$ being the probability of selection of the sample s_k for the design p_k. It is easy to see from (4) as in Arnab and Singh [1]:

$$E(T_k) = \pi\mu_x + (1-\pi)\theta_k, k = 1, 2. \tag{7}$$

This gives rise to unbiased estimation of π and μ_x as given in the following theorems.

Theorem 2.1. An unbiased estimator of the proportion π is given by

$$\hat{\pi} = 1 - \frac{T_1 - T_2}{\theta_1 - \theta_2} \tag{8}$$

with the variance

$$V(\hat{\pi}) = \frac{V(T_1) + V(T_2)}{(\theta_1 - \theta_2)^2} \tag{9}$$

Theorem 2.2. An approximately unbiased estimator of μ_x is given by

$$\hat{\mu}_x = \frac{d_1}{d_2} \tag{10}$$

where $d_1 = T_2\theta_1 - T_1\theta_2$ and $d_2 = (T_2 - \theta_2) - (T_1 - \theta_1)$. An approximate expression of the variance of the above estimator is given as

$$V(\hat{\mu}_x) = (\theta_1 - \mu_x)^2 V(T_2) + (\theta_2 - \mu_x)^2 V(T_1) \tag{11}$$

Remark: Note that the variances of the unbiased estimators in the above theorems may be obtained by replaing $V(T_1)$ and $V(T_2)$ by their respective unbiased estimators. The following theorem is useful in this regard.

Theorem 2.3. The variance of T_k and its unbiased estimator are respectively given as

$$V(T_k) = (\sigma_k^2/N^2)\left[\sum_i \alpha_i(k) I_i' + \sum_i \gamma_i^2(k)\{\alpha_i(k) - 1\} + \sum\sum_{i \neq j} \gamma_i(k)\gamma_j(k)\{\alpha_{ij}(k) - 1\}\right] \tag{12}$$

An unbiased estimator of the variance $V(T_k)$ is given by

$$\hat{V}(T_k) = \hat{V}_k + \frac{\sigma_k^2}{N}(1 - \hat{\pi}(k)) \tag{13}$$

where

$$\alpha_i(k) = \sum_{s_k \ni i} b_{s_k i}^2 p_k(s_k), \;\; \alpha_{ij}(k) = \sum_{s_k \ni i,j} b_{s_k i} b_{s_k j} p_k(s_k), \tag{14}$$

$$\hat{V}_k = \frac{1}{N^2}[\sum_{i \in s_k} \frac{Z_{ki}^2}{\pi_i(k)}(\alpha_i(k) - 1) + \sum_{i \neq j} \sum_{j \in s_k} \frac{z_{ki} Z_{kj}}{\pi_{ij}(k)}(\alpha_{ij}(k) - 1)] \tag{15}$$

and

$$\hat{\pi}(k) = \frac{1}{N} \sum_{i \in s_k} \frac{I_i}{\pi_i(k)}. \tag{16}$$

Now any of the standard sampling strategies, such as Horvitz-Thompson [5], SRSWR, SRSWOR, Hansen-Hurwitz [4], and Rao-Hartley-Cochran [6] can be used to obtain T_k in obtaining the estimator of π and $\hat{\mu}_x$. The results are summarized in the following section.

3 Estimators under Standard Sampling Strategies

The results given below under different strategies can be obtained by using different choices of T_1 and T_2 in the general estimator of π and μ_x described in section 2.

3.1 *Horvitz-Thompson Estimator Based on an Arbitrary Sampling Scheme*

The Horvitz-Thompson [5] $\hat{\pi}_{HT}$ of π, based on arbitrary probability sampling scheme can be obtained by choosing $b_{s_k i} = \frac{1}{\pi_i(k)}$ and is given by

$$T_k = \frac{1}{N} \sum_{i \in s_k} \frac{Z_{ki}}{\pi_i(k)} = T_{kHT} \tag{17}$$

The variance of T_{kHT} is obtained by using Theorem 2.1 and is given in the following theorem.

Theorem 3.1. The variance of T_{kHT} is given by

$$V(T_{kHT}) = \frac{1}{N^2}[\sigma_k^2 \sum_{i \in \bar{G}} \frac{1}{\pi_i(k)} + \sum_{i \in G} X_i^2 (\frac{1}{\pi_i(k)} - 1) + \sum_{i \neq j} \sum_{j \in G} X_i X_j \beta_{ij}(k) \tag{18}$$

$$+\theta_k^2 (\sum_{i \in \bar{G}} (\frac{1}{\pi_i(k)} - 1) + \sum_{i \neq j} \sum_{j \in G} \beta_{ij}(k)) + 2\theta_k (\sum_{i \in G} X_i \sum_{j \neq i \in \bar{G}} \beta_{ij}(k))] \tag{19}$$

where $\bar{G}$ is the complement of the set G and $\beta_{ij}(k) = \frac{\pi_{ij}(k)}{\pi_i(k)\pi_j(k)} - 1$ for $k = 1, 2$.

Using standard results in sampling with unequal probabilities, we can set two unbiased estimators for V_{kHT} gibem below:

$$\hat{V}_{kHT}^{(1)} = \frac{1}{N^2}[\sum_{i \in s_k} \frac{z_{ki}^2}{\pi_i(k)} (\frac{1}{\pi_i(k)} - 1) + \sum_{i \neq} \sum_{j \in s_k} \frac{_{ki} z_{kj}}{\pi_{ij}(k)} \beta_{ij}(k)] + \frac{\sigma_k^2}{N}(1 - \hat{\pi}) \tag{20}$$

and

$$\hat{V}_{kHT}^{(2)} = \frac{1}{N^2} \sum_{i <} \sum_{j \in s_k} \frac{\pi_i(k)\pi_j(k) - \pi_{ij}(k)}{\pi_{ij}(k)} (\frac{z_{ki}}{\pi_i(k)} - \frac{z_{kj}}{\pi_j(k)})^2 + \frac{\sigma_k^2}{N}(1 - \hat{\pi}) \tag{21}$$

The first estimator is obtained from (13) and the second one is an alternative estimator.

3.2 *SRSWOR Sampling Scheme*

The results given above may be specialized to the SRSWOR sampling by substituting, $\pi_i(k) = \frac{n_k}{N}$ and $\pi_{ij}(k) = \frac{n_k(n_k-1)}{N(N-1)}$. The expression for T_k,

$V(T_k)$ and $\hat{V}_{kHT}$ for SRSWOR are obtained as folows:

$$T_k = \frac{1}{n_k} \sum_{i \in s_k} z_{ki} = \bar{z}(k) = T_{kWOR} \tag{22}$$

$$V(T_{kWOR}) = \sigma_k^2 \frac{(1-\pi)}{n_k} + (\frac{1}{n_k} - \frac{1}{N}) \frac{N}{N-1} [(\pi - \frac{1}{N}) S_x^2 + \pi(1-\pi)(\mu_x - \theta_k)^2] \tag{23}$$

and

$$\hat{\hat{V}}_k = (\frac{1}{n_k} - \frac{1}{N}) s_z^2(k) + \frac{\sigma_k^2}{N}(1-\hat{\pi}) \tag{24}$$

where $(N_G - 1)s_x^2 = \sum_{i \in G} x_i^2 - (\sum_{i \in G} x_i)^2 / N_G$ and $(n_k - 1)s_z^2(k) = \sum_{i \in s_k} (z_{ki} - \bar{z}(k))^2$

3.3 Hansen-Hurwitz Estimator Based on PPSWR

Here we assume that the sample s_k of size n_k is selected by PPSWR method of sampling with p_i as a probability of selection for the $i-$th unit of the population at every draw. Using the Hansen-Hurwitz [4] estimator of mean for PPSWR, an unbiased estimator of $\pi\mu_x + (1-\pi)\theta_k$ is obtained by choosing $b_{s_k i} = \frac{n_i(s_k)}{p_i}$ and it is given by

$$T_k = \frac{1}{N n_k} \sum_{i=1}^{N} \frac{n_i(s_k)}{p_i} \bar{z}_{ki} = T_{kHH} \tag{25}$$

where $n_i(s_k)$ denotes the number of times the ith unit appears in the kth sample s_k and $\bar{Z}_{ki}$ denotes the average of the randomized responses from the sample s_k. For the PPSWR sampling scheme we have:

Theorem 3.2. The variance of T_{kHH} is given by

$$V(T_{kHH}) = \frac{1}{N^2 n_k} \Big[\pi (\frac{1}{N_G} \sum_{i \in G} \frac{X_i^2}{N p_i} - \mu_x^2) + \theta_k^2 (\frac{1}{N} \sum_{i \in \bar{G}} \frac{1}{N p_i} - (1-\pi)) + \pi(1-\pi)(\mu_x - \theta_k)^2 \Big] + \frac{\sigma_k^2}{N^2 n_k} \sum_{i \in \bar{G}} \frac{1}{p_i} \tag{26}$$

3.4 SRSWR Sampling Scheme:

Specializing the above result for the SRSWR sampling, an unbiased estimator of $\pi\mu_x + (1-\pi)\theta_k$ is obtained by putting $n_i(s_k) = \frac{1}{n_k}$ in T_k^* and is given by

$$T_{kHH} = \frac{1}{n_k}\sum_{i\in s_k} z_{ki} = \bar{z}(k) = T_{kWR}. \tag{27}$$

The variance of T_{WR} is obtained by putting $p_i = \frac{1}{N}$ in the expression for $V(T_{HH})$ and is given by

$$V(T_{kWR}) = \frac{1}{n_k}[\pi\sigma_x^2 + \pi(1-\pi)(\mu_x - \theta_x)^2 + (1-\pi)\sigma_k^2] \tag{28}$$

where

$$\sigma_x^2 = \frac{1}{N_G}\sum_{i\in G} X_i^2 - \mu_x^2.$$

3.5 Rao-Hartley-Cochran Strategy:

Following the standard terminology of Rao, Hartley and Cochran [6] (RHC) strategy, an unbiased estimator of $\pi\mu_x + (1-\pi)\theta_k$ is given by

$$T_k = \frac{1}{N}[\sum_{i=1}^{n_k} \frac{z_{ki}}{p_i} P_i(k)] = T_{kRHC} \tag{29}$$

for $k = 1, 2$ where $P_i(k)$ denotes the sum of p_i's belonging to the ith group $Q_i(k)(i = 1, 2, ..., n_k)$ that was formed in selection of sample s_k by RHC sampling scheme. For the RHC sampling scheme we have:

Theorem 3.3.

$$V(T_{kRHC}) = \frac{\sigma_k^2}{N^2}\left[\frac{N(n_k-1)}{n_k(N-1)}\sum_i I_i' + \frac{N-n_k}{n_k(N-1)}\sum_i \frac{I_i'}{p_i}\right] + \frac{N-n_k}{N^2(N-1)}\left[\sum_i \frac{\gamma_i^2}{p_i} - \left(\sum_i \gamma_i\right)^2\right] \tag{30}$$

4 Relative Efficiency of Estimators

We next numerically compute the relative efficiency of various strategies. The measure for this is called Percent Relative Efficiency (PRE) of e_1 relative to

Table 1. Percent Relative Efficiency of SRSWOR *vs.* SRSWR for estimating proportion

			π						
N	n_1	n_2	0.1	0.3	0.4	0.6	0.7	0.8	0.9
100	20	20	109.4	115.9	117.5	119.6	120.5	121.4	122.6
500	100	200	114.8	129.7	133.8	138.7	139.9	140.3	139.5
1000	100	200	106.9	112.9	114.4	115.5	116.6	116.7	116.5
1000	200	200	109.3	116.2	117.9	119.1	121.0	121.9	122.9
5000	500	500	104.4	107.4	108.2	108.8	109.5	109.9	110.3
10000	2000	2000	109.3	116.2	117.9	119.2	121.1	121.9	123.0
10000	2000	3000	112.3	123.2	126.1	128.1	130.7	131.5	131.9

e_2 defined as

$$PRE(e_1, e_2) = 100 \times \frac{Var(e_2)}{Vare_1}. \tag{31}$$

4.1 SRSWR vs. SRSWOR

In order to study the performance of the proposed estimator under SRSWOR sampling with respect to SRSWR sampling design, we considered here a few fixed values of the parameters of the sensitive character and that of the randomization devices. We considered a hypothetical situation with $\theta_1 = 20.5$, $\sigma_1^2 = 2.50$,$\theta_2 = 25.6$,$\sigma_2^2 = 2.54$,$\mu_x = 20$ and $S_x^2 = 2.50$. PRE of SRSWOR sampling with respect to SRSWR for estimating proportion for a selection of values of N, n_1, n_2 and π is summarized in Table 1. It has been observed that for moderate sample sizes and population sizes the relative efficiency of SRSWOR sampling over SRSWR sampling remains appreciable. Similar values are obtained for estimating $\hat{\mu}_x$, which are not displayed here.

4.2 PPSWR vs. SRSWR and SRSWOR

We consider here the parameters from the following small artificial population, which was used for checking the behaviour of the estimators under SRSWOR:

We observed that it is possible to make PPSWR sampling more efficient than SRSWR and SRSWOR sampling designs based on the choice of selection probabilities. For illustration we consider $\theta_1 = 63.5$,$\sigma_1^2 = 2.50$,$\theta_2 = 65.6$ and $\sigma_2^2 = 2.54$. From the artificial population we have $\mu_x = 65$,$S_x^2 = 9$,$N_G = 3$

Table 2. Values in an artificial population

65	68	62	10	14	25	36	23	18	20
10	25	39	25	36	47	45	24	25	36

Table 3. Percent Relative Efficiency of PPSWR *vs.* SRSWR and SRSWOR

	$n_1 = n_2 = 3$		$n_1 = 5, n_2 = 7$	
π	WR	WOR	WR	WOR
0.1	419.4	269.1	422.7	249.1
0.2	207.1	152.4	207.9	133.8
0.3	168.6	131.2	169.2	112.9
0.4	152.3	122.1	152.9	104.1
0.5	143.2	117.0	143.7	99.1
0.6	137.3	113.6	137.8	95.8
0.7	133.1	111.2	133.6	93.5
0.8	129.9	109.3	130.4	91.7
0.9	127.3	107.8	127.9	90.3

and $N = 20$. We consider $P_1 = 0.07$,$P_2 = 0.08$,$P_3 = 0.06$ and $P_4 = P_5 = ... = P_{20} = 0.79/17 = 0.046471$. Here $\pi = 3/20 = 0.15$ and the percent relative efficiency of PPSWR sampling over SRSWOR sampling is 256.7 and 179.7 percent,respectively, for $n_1 = n_2 = 3$. Similar experiments have been done by considering different values of π (except the value of N_G) and changing the above artificial population accordingly, then the relative efficiency of PPSWR sampling over SRSWR sampling and SRSWOR sampling, respectively, for different choices of sample sizes and values of true proportion is given in Table 3.

In case, the sample sizes are large relative to the population size, the loss in efficiency of PPSWR *vs.* WOR is not surprising. This may happen because in case of large sample sizes, WOR samples become more representative of the population and hence it may produce gains in efficiency with respect to any WR sampling design. It is more interesting to note that for small values of the true proportion of the gang, the PPSWR sampling shows efficient results than both designs for moderate sample sizes in comparison to population size.

Similar results hold for efficiency of estimators of μ_x.

Table 4. Percent Relative Efficiency of RHC *vs.* PPSWR for estimation of π

π	$n_1 = n_2 = 3$	$n_1 = 5, n_2 = 7$
0.1	102.5	101.3
0.2	103.6	102.6
0.3	104.2	103.6
0.4	105.6	104.5
0.5	107.4	106.6
0.6	105.4	104.6
0.7	103.9	102.4
0.8	103.1	102.2
0.9	102.8	101.9

Table 5. Percent Relative Efficiency of RHC *vs.* PPSWR for estimation of μ_x

π	$n_1 = n_2 = 3$	$n_1 = 5, n_2 = 7$
0.1	110.2	109.7
0.2	111.4	110.4
0.3	113.6	111.5
0.4	115.3	112.5
0.5	116.3	114.6
0.6	114.6	112.5
0.7	113.2	110.2
0.8	110.2	109.8
0.9	108.2	106.2

4.3 RHC Scheme vs. PPSWR

Percent Relative Efficiencies of RHC scheme *vs.* PPSWR have been computed for the same values of parameters as given in the earlier section and reported in Table 4, for estimating proportion. It is interesting to note that under RHC scheme there is slight but non-ignorable gain in efficiency over the PPSWR sampling for estimating proportion. However, for estimating μ_x, there may be considerable gain in efficiency under RHC scheme, as seen from Table 5.

Acknowledgments

The authors are thankful to the referees for constructive comments and the editor, Yogendra P. Chaubey for bringing the original manuscript in the present form.

References

1. R. Arnab and S. Singh, *Ann. Inst. Math. Stat.* (To appear, 2002).
2. P. Bratley, B. L. Fox and L.E. Scharge, *A Guide to Simulation* (Springer-Verlag, New York, 1983).
3. A. Chaudhuri and R. Mukherjee, *Randomized Response: Theory and Techniques* (Marcel Dekker, New York, 1988).
4. M.H. Hansen and W.N. Hurwitz, *Ann. Math. Stat.* **14**, 333 (1943).
5. D.G. Horvitz and D.J. Thompson, *J. Amer. Statist. Assoc.* **47**, 663 (1952).
6. J.N.K. Rao, H.O. Hartley and W.G. Cochran, *J. Roy. Statist. Soc. Ser. B* **24**, 482 (1962).
7. S. Singh, S. Horn and S. Chowdhury, *Austral. & New Zealand J. Statist.* **40**, 291 (1998).
8. S.L. Warner, *J. Amer. Statist. Assoc.* **60**, 63 (1965).

UNEMPLOYMENT, SEARCH AND THE GENDER WAGE GAP: A STRUCTURAL MODEL

CHRISTIAN BELZIL AND XUELIN ZHANG
Department of Economics, Concordia University, H3G 1M8
and Statistics Canada
E-mail: belzilc@vax2.concordia.ca

Using a structural model in which the decision to search is indogenous, we analyze how various parameters such as the mean wage offer, the offer probability and the value of non-market time can explain the gender wage gap. The model, implemented on a sample of young Canadian men and women who suffered a permanent job displacement, is able to explain both the higher incidence of right-censored unemployment spells and longer completed unemployment duration for females. The structural parameters imply that around 30% of the gender re-employment wage gap is explained by the presence of young children.

1 Introduction

For several decades, labor economists have tried to explain the existence of the gender wage gap. As women have constantly increased their share of the labor force, interest in the persistence of a significant difference in wages paid to female versus male workers (given identical observable characteristics) has grown steadily. To date, economists have used two fundamental economic theories to explain the gender wage gap: human capital theory and statistical discrimination.

In the human capital approach, which dates back to Mincer and Polachek [5], the gender wage gap is explained by the fact that females are relatively more productive in household activities than males. For this reason, women tend to invest less in labor market human capital or tend to work in occupations which do not require heavy human capital investments. The gender wage gap is therefore a result of discontinuous work pattern expectations.

The literature on discrimination has, on the other hand, focused on the differential treatment of male and females workers who are otherwise identical. The notion of discrimination, dating back to Becker [1], is based on the fact that employers, facing uncertainty about individual productivity or individual labor force attachment, must sometimes focus on observed differences between males and females when hiring new workers. As a result, women may systematically receive lower wages or may be excluded from various occupations.

The approach to the gender wage gap suggested in this paper is quite

different from most previous work. We use a partial equilibrium job search framework in order to investigate the gender differences in job search outcomes which take place following a permanent job displacement and analyze how much of the gender re-employment wage gap can be explained by the presence of young children.

The model has several distinct features. First, the decision to search is endogenous. Using data on the willingness to search for a new job upon job displacement (participation data), we specify a model where both males and females may decide to drop out of the labour force. By allowing the decision to search to be endogenous, we avoid self-selection bias introduced if we were to sample only those women searching and work with duration and wage data (such as is typically done in the literature). Second, we examine the possibility that the information on job search activities provided by individuals over-estimates the fraction of displaced workers who decided to search. Third, we do not rely on homogeneity assumptions. We use observable characteristics such as age, education, marital status and child status (number of young children) to parameterize three important aspects of the search process; the value of non-market time, the mean wage offer, the offer probability. Reservation wages are treated as a function of unknown parameters and exogenous regressors; this must be solved using dynamic programming principles. Fourth, we also introduce measurement errors in observed re-employment wages.

We believe that investigation of the gender wage gap using a structural model is particularly promising. First, the imposition of all the restrictions imposed by dynamic programming allows us to obtain separate estimates for all parameters of the mean wage offer and the reservation wage function. This means that we can actually compute how various regressors such as the number of young children, marital status, and education, effect on males and females differently. This can be achieved without having to impose exclusion restrictions such as those needed in reduced-form analysis of female wage functions. In other words, our model allows us to distinguish between supply side versus demand side factors affecting the gender wage gap. As a consequence, the structural estimates can be used to investigate gender differences in offered wages, reservation wages and re-employment wages, unemployment duration and on the incidence of non-participation following job displacement.

The model is estimated from a sample extracted from the Canadian Labour Market Activity Survey. We use a sample of men and women who have experienced a permanent job displacement. The likelihood function is based on information on the decision to search or not to upon displacement (we refer to this as participation data), duration data and re-employment wages.

The paper is arranged as follows. In Section 2, the theoretical model is presented in detail. Section 3 is devoted to a discussion of econometric issues and a presentation of the likelihood function. The data set is presented in Section 4 while Section 5 is devoted to a discussion of the main results while an analysis of the gender wage gap resulting from the structural parameter estimates is performed in Section 6. The conclusions are summarized in Section 7.

2 A Model with Endogenous Job Search

To investigate gender differences in job search outcome, we specify a stationary search model similar to the one estimated by Belzil and Zhang [2] to investigate child care and search costs. The model is applied to full-time male and female workers who are affected by a permanent job displacement. We disregard temporary layoffs. The model is constructed around the following assumptions.

1. Expected lifetime earnings are maximized over an infinite horizon and discounted at rate $\beta = \frac{1}{1+r}$
2. Individuals receive at most one offer per period and the probability of receiving an offer is given by ξ. Search is costless.
3. The unemployed receive unemployment benefit b for each period of unemployment.
4. For those who decide not to search, the value of non-market time per period is given by

$$\vartheta(K) = \frac{1}{\tau}\exp(\tau K)$$

   in which $\vartheta(K)$ may be interpreted as the monetary value of the output produced at home, K denoting the number of young children at the time of the displacement.
5. We assume that job offers are indexed by an hourly wage rate and that, upon acceptance, a job is held forever. Wage offers are normally distributed with mean μ and variance σ_u^2.

Using the previous assumptions, it is straightforward to derive the value functions associated with each state. These are

$$V_e(w) = \frac{w}{1-\beta} \tag{1}$$

$$V_u = b + \beta E[V] \tag{2}$$

$$V_n = \frac{1}{1-\beta}\left(\frac{1}{\tau}\exp(\tau K)\right) \tag{3}$$

in which $V_e(w)$ is the value of accepting employment at wage w, V_u is the value of unemployment (search) and V_n is the value of leaving the labour force to involve in household activities. The value of following the optimal policy in the future, $E[V]$, is given by

$$E[V] = \frac{1}{1-\beta}\left[(1-\xi)w^* + \xi P(w < w^*)w^* + \xi P(w \geq w^*)E(w|w \geq w^*)\right] \tag{4}$$

where w^* denotes the re-employment reservation wage (for those who decide to search and remain in the labour force) . In the case where wage follow a normal distribution with density $\phi(\frac{w-\mu}{\sigma_u})$ and cdf $\Phi(\frac{w-\mu}{\sigma_u})$, $E[V]$ can be re-expressed as

$$E[V] = \frac{\xi}{1-\beta}\left[\mu + (w^*-\mu)\Phi(\frac{w^*-\mu}{\sigma_u}) + \sigma_u\phi(\frac{w^*-\mu}{\sigma_u})\right] + (1-\xi)\left(\frac{w^*}{1-\beta}\right) \tag{5}$$

The necessary and sufficient condition to remain in the labour force and search (following displacement) is given by

$$\frac{1}{1-\beta}\left(\frac{1}{\tau}\exp(\tau K)\right) \leq b + \beta E[V] = \frac{1}{1-\beta}(w^*) \tag{6}$$

3 Econometric Specification

In this section, we present the estimation strategy used to investigate the gender wage gap and other related issues.

3.1 Parameterization

To take into account individual unobserved heterogeneity in the value of non market time, we allow the value of non-market time to incorporate a stochastic element:

$$\vartheta(K) = \alpha + \frac{1}{\tau}\exp(\tau K)$$

in which

$$\alpha \sim N(0, \sigma_n^2)$$

In order to introduce observed heterogeneity, we parametrize the offer probability and the mean wage offer. Initially, the job offer probability is allowed

to depend on child status only. This might be relevant if for instance, those women with young children searching for a new job are less flexible or have less time to devote to search activities.[a] To take this into account, we specify the offer probability as

$$\xi = \exp(\xi_0 + \xi_1 K)$$

We parametrize the mean wage offer distribution as a function of education binary variables (primary, secondary and university) indicating the highest degree attained by a given individual, and age dummies: age_1 (16-24, inclusive), age_2 (25-34, inclusive) and age_3 (35-44, inclusive). Note that the reference groups are those who have a secondary education and are aged between 25-34.

$$\mu = \mu_{0j} + \mu_1 \cdot \text{primary} + \mu_2 \cdot \text{university} + \mu_3 \cdot \text{age}_1 + \mu_4 \cdot \text{age}_3 \qquad (7)$$

where $\mu_{0j}, (j = 1, 2)$ is an individual effect intended to capture unobserved heterogeneity in the mean wage offer.

In order to estimate the discount factor and the sample proportion for the unobserved heterogeneity term, we use the following transformations:

$$\beta = \frac{\exp(b_\beta)}{1 + \exp(b_\beta)}$$

$$p_1 = \frac{\exp(\mu_p)}{1 + \exp(\mu_p)}.$$

3.2 Likelihood Functions

Let $s_i = 1$ for those women who decided to search and 0 for those who dropped out. Using condition (6), the probability that a woman i will search is given by

$$\Pr(s_i = 1) = \Pr\left[\alpha \leq h(w_i^*)\right] = \Phi\left[\frac{h(w_i^*)}{\sigma_n}\right] \qquad (8)$$

where

$$h(w_i^*) = w_i^* - \frac{1}{\tau}\exp(\tau K) \qquad (9)$$

The probability that a woman decides not to search, $Pr(s_i = 0)$, follows trivially from equations 6 and 7. As it is the case in most economic surveys of

[a] Of course, it is impossible to distinguish this hypothesis from the hypothesis that employers are less likely to offer employment opportunities to females with children.

the unemployed, some of the individual reporting that they are searching are still unemployed by the end of the survey time. For women, we only observe a censored unemployment duration. For these women who completed their unemployment spell, we observe a completed duration t_i and a re-employment hourly wage w_i. Completed observations are indexed by the binary variable $c_i = 1$ while censored observations are indexed by $c_i = 0$.

It is well known that wage data is often subject to measurement errors. We assume that observed wages, $\tilde{w}_t$ are given by

$$\tilde{w}_t = w_t + \tilde{\varepsilon}_t, \quad \text{with} \tilde{\varepsilon}_t \overset{i.i.d.}{sim} N(0, \sigma_e^2) \tag{10}$$

in which

$$w_t = \mu + \varepsilon_t, \quad \text{with} \varepsilon_t \overset{i.i.d.}{sim} N(0, \sigma_u^2) \tag{11}$$

Assuming that $\tilde{\varepsilon}_t$ and ε_t are independent, then

$$\theta_t = \tilde{\varepsilon}_t + \varepsilon_t \quad \text{i.i.d. } N(0,\sigma_\theta^2) \tag{12}$$

where $\sigma_\theta^2 = \sigma_u^2 + \sigma_e^2$. It is easy to see that the joint probability of receiving an acceptable offer and observing a re-employment wage $\tilde{w}_t$, is given by

$$Pr(w_t \geq w^*, \tilde{w}_t) = \left[1 - \Phi\left(\frac{w^* - \mu - \frac{\sigma_u^2}{\sigma_\theta^2}(\tilde{w}_t - \mu)}{(\sigma_w^2(1-\rho^2))^{\frac{1}{2}}}\right)\right] \cdot \frac{1}{\sigma_\theta}\phi\left(\frac{\tilde{w} - \mu}{\sigma_\theta}\right) \tag{13}$$

where

$$\rho = \frac{\sigma_u}{\sqrt{\sigma_u^2 + \sigma_e^2}} \tag{14}$$

In the paper, we consider two distinct likelihood functions which differ according to whether or not participation data is judged reliable or not.

1. **Model with Participation Data (Model 1).** When we rely on the answer provided by each individual to the question(s) pertaining to their job search status, the log likelihood function for the entire sample is given by

$$L^1 = \sum_{i:s_i=0} \log\left(1 - \Phi\left[\frac{h(w_i^*)}{\sigma}\right]\right) + \sum_{i:s_i=1} \log\Phi\left[\frac{h(w_i^*)}{\sigma}\right] + \log\left[(1 - \xi\pi(\omega_i^*))^{t_i - 1}\xi . \Pr(w \geq w^*, \tilde{w})\right] \tag{15}$$

2. **Model without Participation Data (Model 2).** As reported in Section 4, we have found that quite a large number of individuals mistakenly reported a willingness to search, For example, some individuals who reported they were not willing to work and were not searching for job after being laid-off were reemployed before the end of the survey, while a large number of individuals, reporting that they were searching, remained unemployed for a long period. We consider that the probability that an individual did not find re-employment by the end of the survey is the sum of the probability that he(she) was not searching and the probability that he(she) was indeed searching weighted by the probability that no acceptable offer had been found. The log likelihood is then given by

$$L^2 = \sum_{i:,c_i=1} \log \left[\Phi \left(\frac{h(w^*)}{\sigma_n} \right) (1 - \xi\pi(\omega_i^*))^{t_i - 1} \xi . \Pr(w \geq w^*, \tilde{w}) \right]$$

$$+ \sum_{i:c_i=0} \log \left[\left(1 - \Phi(\frac{h(w^*)}{\sigma_n}) \right) + \Phi \left(\frac{h(w^*)}{\sigma_n} \right) (1 - \xi\pi(w_i^*))^{t_i} \right] \qquad (16)$$

It is easy to see that the following parameters are identifiable (and hence estimable); the standard deviation of α, (σ_n), the parameter of the function representing productivity at home (τ), the wage offer distribution parameter (μ) and σ_u, the standard deviation of the measurement error term σ_ϵ, the discount rate β,[b] the offer probability (ξ). Using a Newton-Raphson procedure, values for reservation wages can be obtained relatively easily although these calculations must be updated at each of the iterations needed to maximize the log likelihood function.[c] With unobserved heterogeneity in the mean wage offer, it is easy to see that the likelihood function is simply a weighted sum of two contributions to the likelihood (for each value of the support points μ_{01} and μ_{02}) where the weights are the probabilities of belonging to a particular type.

4 The Canadian Labour Market Activity Survey

The sample analyzed in this paper is drawn from the 1986-1987 Canadian labour Market Activity Survey (LMAS). The LMAS was designed as a replacement for the Annual Work Patterns Survey in order to provide measures

[b] Many authors set the discount rate β to a constant.

[c] To do so, we make use of the optimality condition and apply a Newton-Raphson algorithm to obtain estimates of the reservation wage.

of labor market dynamics. The data have been collected by Statistics Canada and Employment Immigration Canada.

4.1 Description

The LMAS is based on a stratified sample of Canadians aged between 16 and 69. The sample contains workers who held a full time job and experienced a job displacement. Workers are defined as full-time workers if they worked at least 120 hours per month. Job displacement information is based on question 34 of the LMAS: "What was the main reason for stopping work?". The reasons corresponding to a permanent separation are non-seasonal economic/business conditions, company moving or going out of business, end of a temporary (non-seasonal) job and dismissal by the employer. The question is therefore consistent with the notion of displacement used in the empirical literature. Given these restrictions, we started with 1910 observations: 1091 females and 819 males. Since we were interested in young and prime-age male and female workers, we eliminated all those older than 45 years old as well as those working in primary sector occupations such as farming, fishing, hunting and other occupations of this type. The resulting sample was 794 females and 494 males (1288 observations in total).

In order to estimate the model, a measure of unemployment duration (perhaps censored) is typically needed. Jones and Riddell [4] point out that the LMAS is unique in that the design of the survey attempts to distinguish between individuals who search throughout the entire non-employment spell and those who did not (those out of the labor force and those who have a "marginal attachment" to the labor force). The distinction is made according to the number of consecutive weeks of search reported by every individual and the number of weeks of non-employment in which the individual reported whether he/she was willing to accept any job. A "marginal attachment" applies when an individual reports not searching while also reporting that available work would have been accepted. The post-displacement status can actually be a complex sequence of states and, as documented by Jones and Riddell[4], the exact nature of the non-employment spell is not clear.

To get around these problems, we estimate our model with fixed non-market time value using the information provided in Question 21 (Did you look for work at any time during this period ?) and Question 24 (Did you want to work at any time during this period ?). We define a displaced worker as a non-participant if the individual did not look or did not want to work during the period. The duration of unemployment is computed from the difference between the starting week of the new job (when applicable) and the week of

termination of the previous job.

The LMAS has information on hourly wages as well as hours per week for all jobs sampled. The hourly wage rate used in this study is derived from information on hours worked per day, hours worked per week, total weeks worked and annual earnings in a given job. Using information on whether one has received some unemployment benefit or not, we can construct a measure of the amount of unemployment benefit received for those who report having received UI benefit. To do so, we use the fact that, in 1986, the maximum insurable earnings were $495 per week and the replacement ratio was 60%.

Apart from the information on the decision to search, non-employment duration and accepted wages, the LMAS reports information on the number of young children; this can identify those women who had children when displacement took place. We also observe the marital status (married, divorced, single or cohabitating) so that lone mothers and married mothers can be identified. Education level is reported by class variables (there are 5 classes). For the purpose of this study, we constructed three (3) education dummies. The first group contains those with secondary schooling, the second one contains individuals with post-secondary schooling while those with university training are in the third group. Age is also reported as a class variable. For our sample, we construct three (3) age dummies corresponding to 16-24, 25-34, and 35-44. Variable definitions and sample statistics can be found in the Appendix.

4.2 Some Features of the Data

One of the most striking features of the data is the difference between male and female unemployment duration. Although the overall averages are quite close (20 weeks for males and 23 weeks for females), censoring appears more important for females. Among 794 women aged below 45, only 264 (33%) had found a new job by the end of the survey. The remaining 558 women are split between those who were still searching at the end of the survey and those who report not searching after the loss of their previous job. The difference in sample averages between male and female unemployment duration is, however, not explained by censoring alone. Belzil and Zhang [2] also report that simple non-parametric rank tests indicate clearly that females experience longer unemployment spells than males. It is interesting to note that despite the large number of censored non-employment spells, only 6% (48/794) report not being available to search. Among 494 males, more than 50% had found a new job before the end of the survey. Interestingly, although pre-unemployment wages and re-employment wages for both males and females are relatively similar (10.09$ vs 9.61$ for males and 6.28$ vs 6.95$ for fe-

males), we found female average pre-unemployment wages for those who have no children and those with children to be almost the same (6.30$ vs 6.35$). However, the average re-employment wage of females with no children (7.00$) exceeds the average re-employment wage of those with children (6.26$) by more than 11%.

5 Empirical Results

As discussed earlier, we worked with two versions of the model. The first version is estimated under the assumption that the reported participation decision information is reliable (Section 5.1) while the second version uses the likelihood function which does not use participation data (Section 5.2). Section 5.3 is devoted to the introduction of marital status while, in Section 5.4, we investigate the initial condition problem that could arise if permanent unobserved heterogeneity in the labor market is correlated with child status.

5.1 Model with Participation Data

The results obtained under the assumption that the reported participation decision is reliable are in Table 1-A (Model I). A summary of the value of non-market time and the offer probability implied by the structural parameters are in Table 1-B (Model I). Estimates for the variance of the true wage offer (1.36 and 0.09) and the variance of the measurement error (2.69 and 2.13) illustrate the importance of measurement error. Unobserved heterogeneity in mean wage offer is also found to be important; about 10% of male and female workers are at the high end of mean wage offer ($16 for males and $15 for females).[d] Although the gender gap in mean wage offer appears to be very small, the differences observed in the value of non-market time, offer probabilities and the discount factors reveal differences in job search behavior between males and females. The parameter estimates of the effect of young children on the offer probability (ξ_1) indicates that men with children receive more offers than those without children while it is the reverse for females. For those without children, these estimates imply that males face a probability of 0.05 per week of receiving an offer while females would have a probability of 0.31.With one child, female workers still receive offers at higher rate than male workers (0.15 vs. 0.13). These estimates do not seem to fit the data very well because, in the sample, females experience longer spells of unemployment.

[d]We started to allow for three mass points in the mean wage offer. However, we found that two of them are fairly close to each other. This lead us to specify only two mass points for μ.

The estimates for b_β imply a very large a difference in discounting behavior between males and females as the annual discount rates are 29% for males and 0.1% for females. Given that our sample consists of relatively homogeneous individuals (in age and education, for example), it is quite unlikely that such a difference is the case. The estimates for the value of non-market time (Table 1-B, Model I) indicate that males and females tend to have similar home productivities (a questionable result) and that home productivity is increasing with child status (as expected).

5.2 Model without Participation Data

Table 1-A (Model II) contains the result for the model estimated under the assumption that the reported participation decision may not be reliable. The estimates of σ_e for males and females (2.44 and 1.21 respectively) indicate again the importance of measurement error in wage data while the two constant terms μ_{01} and μ_{02} (15.33 vs. 7.03 for males) and (11.38 vs. 3.77 for females) indicate the importance of unobserved heterogeneity. Other things equal, university education raises the wage rate by \$1.32 for male and \$1.80 for female workers when compared to the reference group (high school education). Although the sample contains only those who were below 45, we still find a positive effect of age on expected wages although it is insignificant for females. This is perhaps a reflection of the fact that females have flatter age-earnings profiles than males.

From Table 1-B (Model II), we see that offer probabilities increase with child status for males but decrease with child status for females. Female displaced workers with no children receive offers at a higher rate than males with no children (0.20 vs. 0.07). With one child, male and female workers receive job offers at a similar rate (about 0.1); with two children, the probability a male worker receives job offer per period of time (one week in the current study) becomes 0.15, but that for a female worker becomes 0.05. The discount factors estimated are 0.965 for male workers and 0.94 for female workers. These estimates are close to those obtained by Christensen and Kiefer [3]. Although the results show that male job searchers put a relatively higher value for future offers than their female counterparts, we no longer observe a huge gender differences in discount rates as obtained with Model 1.[e]

The most striking difference between Models I and II lies in the value of home productivity. When we let the data determine the probability that a

[e]In a companion paper (Belzil and Zhang [2], 1996), we estimate of structural search model with child care costs in which the discount factor in treated as an individual effect. We find a particularly large dispersion in individual discount factors amongst women.

Table 1-A: Estimation Results for Models I and II

	Male		Female	
	Model I	Model II	Model I	Model II
σ_u	1.3571(2.38)	1.7748(4.23)	0.0882(3.72)	2.6149(4.15)
σ_e	2.6860(12.60)	2.4400(11.09)	2.1324(18.99)	1.2058(8.44)
μ_p	-2.3626(-7.94)	-1.8875(-6.57)	-2.5443(-7.75)	-2.8972(-5.30)
μ_{01}	16.0007(15.12)	15.3292(17.42)	15.0150(20.42)	11.3806(4.26)
μ_{02}	7.2211(7.50)	7.0281(10.75)	7.0468(68.98)	3.7692(1.86)
μ_1	-0.8517(-1.36)	-1.1636(-1.36)	-0.3807(-2.28)	-1.2674(-1.54)
μ_2	0.8228(1.65)	1.3208(2.14)	0.3431(3.54)	1.8027(4.69)
μ_3	-0.8546(-1.97)	-1.2983(-2.29)	-1.2255(-10.9)	-0.9423(-1.64)
μ_4	0.0953(0.22)	1.2450(1.64)	-0.3014(-3.62)	0.4177(0.75)
ξ_0	-2.9462(-6.47)	-2.6618(-14.7)	-1.1577(-10.9)	-1.6048(-2.86)
ξ_1	0.8731(2.67)	0.3829(2.63)	-0.7581(-3.34)	-0.7103(-3.16)
τ	0.3972(5.51)	0.5303(2.76)	0.4418(8.94)	0.1628(5.24)
σ_n	2.8689(7.27)	10.3291(2.99)	2.8087(12.95)	9.3314(1.85)
b_β	5.1807(8.85)	3.3061(7.80)	10.7921(17.08)	2.7810(10.48)
logl	-1851.46	-1832.16	-1897.55	-1892.86

Note: Asymptotic t-ratios are in parantheses

Table 1-B: Home Time Value and Offer Probability

		Model I		Model II	
		Hm. Value	Offer Prob.	Hm. Value	Offer Prob.
Female	$K=0$	2.26	0.31	6.14	0.20
	$K=1$	3.52	0.15	7.23	0.10
	$K=2$	5.48	0.07	14.46	0.05
Male	$K=0$	2.52	0.05	1.89	0.07
	$K=1$	3.75	0.13	3.21	0.10
	$K=2$	5.57	0.30	5.45	0.15

given individual is searching, the household productivity parameter τ estimates are 0.53 for males and 0.16 for females. The implied home time values, shown in Table 1-B (Model II) indicate that females are more productive at home than males. This seems to indicate that answers provided by individuals on their job search status tend to over-estimate search intensity and under-estimate home productivity.

Overall, we believe that the estimates obtained without relying on participation data yield more convincing results. However, on the matter of which

model fits the data better, the log likelihood values do not yield a clear indication.

5.3 The Effects of Marital Status

The estimates presented in Table 1-A were obtained from model specifications in which marital status played no part. As most static and dynamic models of female labor market behavior are based on the fact that married and single females behave differently, it is reasonable to investigate whether marital status has an impact on job search outcomes for female displaced workers. To do so, we allow the offer probability and home time value to be a function of marital status. The parameterization for ξ and τ are $\xi = exp(\xi_0 + \xi_1 K + \xi_2 M)$ and $\tau = exp(\tau_0 + \tau_1 M)$ respectively where M is equal to 1 for those who are married or live in a permanent union and 0 otherwise. Marital status is introduced in the offer probability function in order to reflect the possibility that married women have less (or more) time to devote to search activities. We retained a specification without participation data (such as in Model 2).

The results (in Table 2-A, under Model III) show that marital status has basically no effect on male offer probability and male home time value, though it does raise female home productivity. The level of significance is however relatively low (t-ratio of 1.48). Although being married seems to raise the offer probability for males (0.35) and females (0.15), standard errors are also large. Introducing marital status does not seem to lead to any major improvement.

5.4 The Initial Condition Problem: Child Status and Unobserved Heterogeneity

The estimates obtained in Section 5.1, 5.2 and 5.3 were generated under the assumption that unobserved ability in the labor market production is independent from any individual characteristic, including child status. It is possible that the distribution of unobserved labor market ability among women who have children is different than for women with no children. For instance, one could imagine that females who were working full-time in the presence of young children have relatively higher labor market ability than those women with no children.

In what follows,we extend the model to take into account correlation between unobserved heterogeneity in the mean wage offer and female child status. The probability of belonging to the high level, p_1, is given by

$$p_1 = \frac{\exp(\mu_{p0} + \mu_{p1} K)}{1 + \exp(\mu_{p0} + \mu_{p1} K)} \tag{17}$$

The results (Table 2-A, Model IV and Table 2-B) do not support the hypothesis that permanent unobserved heterogeneity in labor market ability is correlated with child status. Although the probability of belonging to the high ability class ($\mu_{01} = 15.3$) increases with K, the very low t-ratio (0.04) implies that sorting is insignificant.

6 Investigating the Gender and the Fertility Wage Gap

In this section, we use the structural parameters obtained for Model II (where participation data are not necessarily reliable) to calculate the mean wage offer, reservation wage and observed re-employment wage rate (all measured in \$ terms on an hourly basis).[f] The gender gap (1 minus the ratio of female re-employment wage to male re-employment wage) can be calculated at different levels of education, age and number of children. Controlling for age and education, we can define a **gross gender wage gap** that reflects gender as well as child status. This gross gender wage gap is then decomposed into a portion that is due to child status and a residual part.

The mean wage offer μ is calculated according to equation 8, while the reservation wage is obtained through the Newton-Raphson procedure using the estimates of the structural parameters. The expected re-employment wage can be calculated using,

$$E[w|w \geq w^*] = \mu + \sigma_u \left(\frac{\phi(\frac{w^*-u}{\sigma_u})}{1 - \Phi(\frac{w^*-\mu}{\sigma_u})} \right).$$

Table 3 presents the expected mean wage offer, mean reservation wage and average expected re-employment wage for both male and female workers at different levels of age, education, and child status (mean wage offer is independent of child status). Table 4 contains the corresponding gender wage gap. The gender wage gap in mean wage offers ranges between 60% (for low education workers) and 40% for high education workers. We note that the reservation wage gap is smaller than the wage offer gap. This implies that female hazard rates will be smaller than male hazard rates. Note also that the reservation wage gap rises significantly as the number of children increases; it goes from around 20% (for those without children) to around 50% for those with 2 children. Both the reservation wage gap and the re-employment wage gap are widening with age and shrinking with respect to education level. More importantly, the gender wage gap becomes larger as the number of children increases. For younger females, the gender wage gap goes from 15% to 20%

[f] All values are averaged over the two support points in the mean wage offer.

Table 2-A: Estimation Results for Models III and IV

	Male		Female	
	Model III	Model IV	Model III	Model IV
σ_u	1.8470(4.32)	1.7688(4.25)	2.6141(4.68)	2.6265(4.13)
σ_e	2.4122(11.35)	2.4370(11.15)	1.2085(8.59)	1.2041(8.46)
μ_p	-1.9204(-6.82)	——	-2.9910(-5.39)	——
μ_{p0}	——	-1.9904(-0.72)	——	-2.9026(-4.87)
μ_{p1}	——	0.2244(0.04)	——	0.0296(0.04)
μ_{01}	15.2750(17.03)	15.3283(17.53)	11.7045(4.58)	11.3294(4.23)
μ_{02}	6.8664(9.43)	7.0213(10.85)	3.9000(2.09)	3.7276(1.83)
μ_1	-1.1565(-1.38)	-1.0958(-1.29)	-1.2362(-1.49)	-1.2580(-1.53)
μ_2	1.3364(2.18)	1.3650(2.21)	1.7563(4.66)	1.8044(4.70)
μ_3	-1.1204(-1.90)	-1.2926(-2.29)	-1.0650(-1.86)	-0.9287(-1.63)
μ_4	1.0686(1.36)	1.2206(1.62)	0.1950(0.34)	0.4181(0.75)
ξ_0	-2.8432(-12.6)	-2.6669(-14.9)	-1.6698(-3.62)	-1.5964(-2.81)
ξ_1	0.3281(2.00)	0.3843(2.65)	-0.6615(-3.06)	-0.7064(-3.17)
ξ_2	0.3458(1.24)	——	0.1506(0.56)	——
τ	——	0.5274(2.75)	——	0.1637(5.42)
τ_0	-0.0461(-0.73)	——	-1.6430(-8.01)	——
τ_1	-0.9884(-1.31)	——	-0.2838(-1.48)	——
σ_n	8.6643(2.38)	10.2342(3.01)	8.6094(2.12)	9.1749(1.92)
b_β	3.3617(7.50)	3.3077(7.82)	2.7884(10.06)	2.7833(10.58)
Logl	-1831.14	-1832.16	-1891.70	-1892.86

Note: Asymptotic t-ratios are in parantheses

Table 2-B. Marital, Offer Prob. and Hm. Value

		Married		Single	
		Home Value	Offer Prob.	Home Value	Offer Prob.
Female	$K=0$	6.87	0.22	5.17	0.19
	$K=1$	7.95	0.11	6.27	0.10
	$K=2$	9.19	0.06	7.61	0.05
Male	$K=0$	2.81	0.08	1.05	0.06
	$K=1$	4.02	0.11	2.72	0.08
	$K=2$	5.73	0.16	7.07	0.11

Table 3: Mean Wage, Reservation Wage and Re-employment Wage

		Male			Female		
		K=0	K=1	K=2	K=0	K=1	K=2
Mean Wage Offer (μ)							
	Primary	5.66			1.96		
Age_1=1	Secondary	6.82			3.23		
	University	8.14			5.03		
	Primary	6.96			2.90		
Age_2=1	Secondary	8.12			4.17		
	University	9.44			5.97		
	Primary	8.20			3.32		
Age_3=1	Secondary	9.37			4.59		
	University	10.69			6.39		
Reservation Wage							
	Primary	4.92	5.30	5.68	3.86	3.42	3.10
Age_1=1	Secondary	5.57	6.03	6.47	4.45	3.87	3.39
	University	6.35	6.89	7.40	5.43	4.64	3.94
	Primary	5.65	6.11	6.56	4.29	3.74	3.31
Age_2=1	Secondary	6.34	6.88	7.39	4.94	4.25	3.66
	University	7.15	7.78	8.36	6.00	5.09	4.27
	Primary	6.39	6.93	7.45	4.49	3.90	3.42
Age_3=1	Secondary	7.11	7.73	8.30	5.18	4.43	3.79
	University	7.94	8.65	9.29	6.25	5.30	4.42
Re-employment Wage							
	Primary	6.73	6.92	7.13	5.40	5.10	4.88
Age_1=1	Secondary	7.63	7.83	8.06	6.17	5.77	5.48
	University	8.72	8.92	9.16	7.40	6.91	6.53
	Primary	7.74	7.94	8.17	5.96	5.57	5.31
Age_2=1	Secondary	8.70	8.91	9.14	6.79	6.34	6.00
	University	9.85	10.04	10.28	8.09	7.56	7.16
	Primary	8.77	8.98	9.21	6.23	5.82	5.53
Age_3=1	Secondary	9.78	9.98	10.21	7.08	6.61	6.25
	University	10.97	11.15	11.38	8.41	7.87	7.46

(with no children) up to 30% (with 2 children). For the older age group in the sample, the wage gap goes from 25% (with no children) to 35% to 40% (with 2 children).

Table 4: Predicted Gender Wage Gap (%)

		Mean*	Reserv. Wage			Re-empl. Wage		
		(μ)	K=0	K=1	K=2	K=0	K=1	K=2
	Primary	65	22	36	45	20	26	32
Age_1=1	Secondary	53	20	36	48	19	26	32
	University	38	15	33	47	15	22	29
	Primary	58	24	39	45	23	30	45
Age_2=1	Secondary	49	22	38	51	22	29	34
	University	37	16	35	49	18	25	30
	Primary	59	30	44	54	29	35	40
Age_3=1	Secondary	51	27	43	54	28	37	39
	University	40	21	39	52	22	29	34

Our measurement of the contribution of child status to the gender wage gap is based on three expected wage values; one for a male worker with no children, one for a female worker who has no children and the third value is from another female worker who has one child. The difference in expected wage between a male worker who has no children and a female worker who has one child becomes the **gross gender wage gap** (reflecting the effects of gender and child status) while the gap between the two female workers is referred to as the **fertility wage gap**. We normalize the gross gender wage gap to be unity, and calculate the percentage contribution by fertility. Table 5 presents the calculated fertility wage gaps for different age and education groups. In the first column, we compare a male worker who has no children and two female workers, one has no children, the other has 1 child. In the second column, the fertility wage gap is computed using a male and a female worker with no children, and a female worker with two children. On average, more than 20% of the gender wage gap observed for females with one child is due to the presence of this child. For female workers who have two children, the presence of young children contributes about 30% to the gross gender gap in expected wage.

7 Conclusion

The fact that males tend to earn more than females is a stylized fact present in all industrialized countries. Several explanations, reviewed in the Introduction, have been proposed by labor economists. In this paper, we have

proposed a new approach to the issue; namely, we have investigated how the job search process may differ across males and females. Using a structural job search model with endogenous search, we have estimated how differences in structural parameters can affect the discounted expected lifetime earnings of unemployed females and found that the presence of young children plays an important role in the setting of the optimal reservation wage. More precisely, we found that female workers with young children have a substantially lower probability of receiving an offer while the opposite is true for males. In other words, the presence of young children translates into a significant efficiency loss in their of job search. We also found that the value of non-market time for females with young children is actually higher than for males. Overall,the self-reported search status in the Labour Market Activity Survey seems to over-estimate the number of female displaced workers who are actually searching. This is consistent with the higher incidence of censoring among women than among men.

Table 5: Fertility Wage Gap (%)

		K=1	K=2
$Age_1=1$	Primary	18	28
	Secondary	22	32
	University	27	40
$Age_2=1$	Primary	18	27
	Secondary	19	29
	University	23	35
$Age_3=1$	Primary	13	22
	Secondary	14	24
	University	17	27

One of the advantages of obtaining structural estimates is that we can distinguish supply side effects (non-market time) from demand side effects (offer probability) and impute a fraction of the gender wage gap to the efficiency loss in job search explained by the presence of young children. Our estimates imply that females with no young children face parameters much the same as male workers. When females with no children are compared with males with no children, the gender wage gap is between 15% to 25%. We found the gender wage gap to be increasing with the number of young children (up to 35% for females with no children). However, the results indicate that around 30% of the gender wage gap is actually accounted for by the effect of young children on the valuation of job search activities (through the offer probability). As well, our results indicated that the gender gap in the mean wage offer

received tends to exceed the gender reservation wage gap and is therefore consistent with the fact that females experience longer spells of unemployment than males. Finally, the results presented are consistent with the claim that the gender wage gap is typically small when males and females enter the labor market but tends to increase with age or experience.

Appendix A: Variable Definition and Sample Mean

Variable	Male	Female	Definition
Marital	0.5992	0.5781	Marital Status, 1 if married
Kid	0.3765	0.3325	Number of children under 6
Prim	0.1174	0.0781	=1 for elementary or lower education
Second	0.6336	0.5982	=1 for high school or post-secondary
Univ	0.2490	0.3237	=1 for university education
Age_1	0.2794	0.4181	=1 if age falls between 16-24
Age_2	0.4413	0.3967	=1 if age falls between 25-34
Age_3	0.2794	0.1851	=1 if age falls between 35-44
Wage1	10.0923	6.2843	Hourly wage rate before separation ($)
Wage2	9.6082	6.9574	Hourly re-employment wage rate ($)
U-Durat	20.5243	22.6373	Unemployment duration in weeks
UIB	3.9212	1.7465	Hourly UI benefit ($)
Censor	0.5041	0.3325	=1 for censored unemployment spell
Parti	0.9575	0.9396	=1 for labour market participant
N	494	794	Sample size

References

1. Becker, Gary (1971) The Economics of Discrimination , University of Chicago Press, Chicago.
2. C. Belzil and X. Zhang, *Young Children and the Search Costs of Unemployed Females* (Working paper, European University Institute, San Domenico di Fiesole, Italy, 1996).
3. B. J. Christensen and N. Kiefer, *Econometric Theory* **7**, 464 (1991).
4. S. Jones and W. C. Riddell, *Journal of Labor Economics* **13**, 351 (1995).
5. J. Mincer and S. Polachek, *Journal of Political Economy* **82**, ((1)974).
6. N. Stokey and R. E. Lucas, *Recursive Methods in Economic Dynamics* (Harvard University Press, Cambridge, MA, 1989).
7. P. Swaim and M. Podgursky, *Journal of Labor Economics* **12**, 640 (1994).
8. K. Wolpin, *Econometrica* **55**, 801 (1987).

KULLBACK-LEIBLER OPTIMIZATION OF DENSITY ESTIMATES

A. BERLINET AND E. BRUNEL

Laboratoire de Probabilités et Statistique, Université Montpellier II, CC 051, Place Eugène Bataillon, F-34095 Montpellier Cedex 5, France
E-mail: berlinet@stat.math.univ-montp2.fr

We analyze the asymptotic behavior of the expected value of the Kullback-Leibler information divergence between the true density and modified histograms. This provides an asymptotically optimal value for the smoothing parameter which can be used in plug-in methods.

1 Introduction

The Kullback-Leibler divergence is a basic tool in decision and information theory. It has a lot of attractive features such as its additivity property used in projection pursuit density estimation and it defines a strong notion of convergence (stronger than L_1-convergence) in the set of probability densities. Also in some parametric models the Kullback-Leibler divergence can be expressed as simple function of the parameters. However, with this error criterion, difficulties appear when standard kernel estimates or histograms are considered to estimate an unknown probability density from a sample of observations. On the one hand tail properties of the kernel dramatically influence the behavior of estimates (Hall [9]). On the other hand empty cells, occuring with high probability, make the Kullback-Leibler divergence between the true density and the histogram equal to infinity. Recently, modified histograms circumventing the problem of empty cells have been shown to have nice properties with respect to information divergences (Barron, Györfi and van der Meulen [2], Berlinet, Györfi and van der Meulen [5], Berlinet, Vajda and van der Meulen [6], Györfi, Liese, Vajda and van der Meulen [7]). As usual in such a non parametric density estimation framework the crucial issue is about the number of cells to take into account. From upper bounds given by Barron and Sheu [3] and Barron *et al.* [2] for densities with bounded support and finite Fisher information it follows that the number of cells should be of order $n^{1/3}$, where n is the number of observations. Here we carefully analyze the behavior of the expected value of the Kullback-Leibler divergence between the true density and the modified histogram and give the exact constants in its asymptotic expansion. This provides an asymptotically optimal value for the number of cells which can be used in plug-in methods.

2 Kullback-Leibler divergence and modified histograms

First recall that if f and g are two probability densities on $\mathbf{R}$, the Kullback-Leibler information divergence of f with respect to g is defined by

$$\mathcal{I}(f,g) = \begin{cases} \int f(x)\ \log(f(x)/g(x))\ d\lambda(x) & \text{if } f << g \\ \infty & \text{otherwise,} \end{cases} \tag{1}$$

where λ is the Lebesgue measure. Integrals without limits are taken over the whole real line.

Let us now turn to the estimation problem and the definition of estimates. Let $(X_i)_{i\geq 1}$ be a sequence of independent real random variables with the same unknown distribution μ on $\mathbf{R}$. The probability measure μ is supposed to have a density f with respect to λ. Define a sequence of integers $(m_n)_{n\in\mathbf{N}^*}$ such that $0 < m_n < n$ and put $h_n = 1/m_n$. The integer m_n will be the number of cells for n observations $X_1, \ldots, X_n$ and h_n is called the smoothing parameter. We suppose that we know g, a density which can be seen as a *reference density*, satisfying $\mathcal{I}(f,g) < \infty$ and denote by ν the probability measure on $\big(\mathbf{R}, \mathcal{B}(\mathbf{R})\big)$ with density g. In practice g will correspond to some a priori idea on the unknown density. For instance one can suspect that the true density f is not far from some known parametric density g. Contamination and more generally mixture models are examples of such situations. When the entropy of f and $E(\log|X|)^+$ are finite a density constant on $(-1,1)$ and behaving like $constant/x^2$ if $|x| \geq 1$ is such that $\mathcal{I}(f,g) < \infty$ (Györfi and van der Meulen [8]). Define finite partitions $\mathcal{P}_n = \{A_{n,1}, \cdots, A_{n,m_n}\}$ of $\mathbf{R}$ such that the $A_{n,k}$'s are m_n consecutive real intervals with $\nu(A_{n,k}) = h_n$, $k \in \{1, \cdots, m_n\}$. Let $(a_n)_{n\in\mathbf{N}^*}$ denote the sequence of strictly positive numbers given by $a_n = 1/(1+nh_n)$, $n \geq 1$. The density estimate introduced by Barron [1] is then defined as

$$\hat{p}_{h_n}(x) = a_n\left[1 + n\mu_n\big(A_n(x)\big)\right] g(x) = \left[(1-a_n)\frac{\mu_n\big(A_n(x)\big)}{h_n} + a_n\right] g(x) \tag{2}$$

where $A_n(x) = A_{n,k}$ if $x \in A_{n,k}$ and μ_n stands for the empirical measure associated with the sample $X_1, \ldots, X_n$. As shown by Barron *et al.* [2] (who study the case with general $(a_n)_{n\in\mathbf{N}^*}$), under the assumption $\mathcal{I}(f,g) < \infty$, if $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$, then

$$\mathcal{I}(f,\hat{p}_{h_n}) \to 0 \ a.s. \quad \text{and} \quad \mathbf{E}\left(\mathcal{I}(f,\hat{p}_{h_n})\right) \to 0. \tag{3}$$

Moreover, Berlinet et al (1997) showed, with the additional assumption $\nu(\bar{S}_\mu - S_\mu) = 0$ where $S_\mu = \{x : f(x) > 0\}$, that

$$n\sqrt{2\,h_n}\left[\mathcal{I}(f,\hat{p}_{h_n}) - \mathbf{E}\left(I(f,\hat{p}_{h_n})\right)\right] \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0,\nu(S_\mu)\Big). \tag{4}$$

The condition $\mathcal{I}(f,g) < \infty$ is minimal to ensure the finiteness of $\mathcal{I}(f,\hat{p}_{h_n})$ and of its expectation. Now, the choice of g can easily meet the condition on the boundary of the support of μ. It is worth noting that (4) is independent of the dimensionality of the problem (see Berlinet et al (1997)). As $\nu(S_\mu) \leq 1$, one gets from (4), for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}\left(n\sqrt{2h_n}|\mathcal{I}(f,\hat{p}_{h_n}) - \mathbf{E}[\mathcal{I}(f,\hat{p}_{h_n})]| > \varepsilon\right) \leq 2\,\Phi(-\varepsilon), \tag{5}$$

where $\Phi(\cdot)$ stands for the cumulative distribution function of a standard gaussian random variable. The upper bound in (5) is independent of both f and g. For other properties of modified histograms see the above mentioned references and the recent paper by Berlinet and Biau [4].

3 Asymptotics

Let us write $\mathcal{I}(f,\hat{p}_{h_n}) = B(h_n) + V(h_n)$, where

$$B(h_n) = \int f(x) \log\left(\frac{f(x)}{\mathbf{E}\,[\hat{p}_{h_n}(x)]}\right) d\lambda(x)$$

and

$$V(h_n) = \int f(x) \log\left(\frac{\mathbf{E}\,[\hat{p}_{h_n}(x)]}{\hat{p}_{h_n}(x)}\right) d\lambda(x).$$

The terms $B(h_n)$ and $V(h_n)$ are respectively named bias and variance components by analogy with the theory of quadratic loss. But balance between these two components cannot be achieved so easily than between analogues in L^2-theory. As Theorem 1 shows, the order of convergence of the bias component depends crucially on tail properties of f and g whereas the variance component (in which g simplifies and disappears) is typically of order $(nh_n)^{-1}$ without almost no restriction.

From the very definition of $\hat{p}_{h_n}(x)$ it appears that a strange fact can happen with that estimate. We can be very lucky and have $g = f$, the unknown density. Then the choice $h_n = m_n = 1$ leads for any x to

$$\hat{p}_{h_n}(x) = f(x).$$

This is unbeatable whatever the error criterion one can consider! In the following we will remove that case by considering densities such that the derivative of (f/g) cannot be identically zero.

Assumptions

Let $f_0 = f/g$ be the density of μ with respect to the probability measure ν, denote by G the distribution function of ν and by $\varphi = f_0 \circ G^{-1}$ the density of $G(X)$. Let $S_\mu = \{x : f(x) > 0\}$, $\bar{S}_\mu$ be the closure of S_μ, and for $\varepsilon \in (0,1)$ let $\mathcal{N}_n = \{x \in S_\mu : \mathbf{E}[\hat{p}_{h_n}(x)] < (1-\varepsilon) f(x)\}$.
We consider the following assumptions

$(\mathcal{A}_0)$ $\mathcal{I}(f,g) < \infty$; $n \to \infty$; $nh_n \to \infty$;

$(\mathcal{A}_1)$ $\nu(\bar{S}_\mu - S_\mu) = 0$;

$(\mathcal{A}_2)$ φ admits first and second order bounded derivatives;

$(\mathcal{A}_3)$ there exists $\gamma > 0$ so that for all $x \in S_\mu$, $f_0(x) \geq \gamma$.

Comments on assumptions

The only assumptions required for the analysis of the variance component are $(\mathcal{A}_0)$ which is the minimal condition for consistency of Barron estimates and condition $(\mathcal{A}_1)$ already given by Berlinet et al (1997). This last condition is not too much restrictive since it is satisfied for all measures ν for which the boundary of the support of μ is negligible. Although f is unknown, it is often the case that its support S_μ is known. Then one can choose ν such that $\nu(S_\mu) = 1$. Otherwise $\nu(S_\mu)$ has to be estimated, for instance by $\nu(\min_i X_i, \max_i X_i)$. Here we only consider the case where $\nu(S_\mu)$ is known. The two following weaker conditions could be given in place of $(\mathcal{A}_3)$.

$$(\mathcal{A}'_3) \quad \int_{\mathcal{N}_n} f(x) \log\left(\frac{\mathbf{E}[\hat{p}_{h_n}(x)]}{f(x)}\right) d\lambda(x) = o\left(h_n^2\right);$$

$$(\mathcal{A}'_4) \quad \sup_{\mathbf{R}/\mathcal{N}_n} \frac{g(x)}{f(x)}\, d\lambda(x) = O\left(n^2 h_n^4\right) \text{ and } \int_{\mathbf{R}/\mathcal{N}_n} \frac{g^2(x)}{f(x)}\, d\lambda(x) = o\left(n^2 h_n^4\right);$$

Under $(\mathcal{A}_3)$ the ratio $\mathbf{E}[\hat{p}_{h_n}(x)]/f(x)$ converges uniformly to 1, thus for n large enough, the set $\mathcal{N}_n = \{x \in S_\mu : \mathbf{E}[\hat{p}_{h_n}(x)]/f(x) < 1-\varepsilon\}$ is empty and assumption $(\mathcal{A}'_3)$ follows directly. Under $(\mathcal{A}'_3)$ and $(\mathcal{A}'_4)$, the set $\mathcal{N}_n$ (which can be intuitively understood as a kind of "non-uniform convergence set of the ratio") is asymptotically negligible in the sense that $\lambda(\mathcal{N}_n)$ tends to 0. Conditions $(\mathcal{A}'_3)$ and $(\mathcal{A}'_4)$ are less easy to interpret than the condition $(\mathcal{A}_3)$

but they allow more flexibility in the choice of g. Indeed, under $(\mathcal{A}_3')$ and $(\mathcal{A}_4')$, f/g is allowed to tend to 0 with slow rate of convergence. Also conditions $(\mathcal{A}_3')$ and $(\mathcal{A}_4')$ involve the expectation of the estimate $\mathbf{E}\left[\hat{p}_{h_n}(x)\right]$, so $(\mathcal{A}_3)$ can be preferred. Although it is stronger, it only involves f and g. So it appears as a more clear and understandable condition on the model.
As well known conditions on tail behavior *must* appear when the accuracy of estimates is evaluated by means of Kullback-Leibler divergence.
The following theorem gives the asymptotic behavior of the bias component and of the expectation of the variance component of the Kullback-Leibler divergence of modified histograms.

Theorem 1 *Under the assumptions* $(\mathcal{A}_0)$ *and* $(\mathcal{A}_1)$,

$$\mathbf{E}\left[V(h_n)\right] = \frac{\nu(S_\mu)}{2nh_n} + o\left(\frac{1}{nh_n}\right). \tag{6}$$

If moreover the assumptions $(\mathcal{A}_2)$ *and* $(\mathcal{A}_3)$ *are satisfied, if*

$$0 < \int \frac{f_0'^2(x)}{f(x)} d\lambda(x) < \infty$$

and $nh_n^2 \to \infty$, *as* $n \to \infty$, *then*

$$B(h_n) = \frac{h_n^2}{24} \int \frac{f_0'^2(x)}{f(x)} d\lambda(x) + o\left(h_n^2 + \frac{1}{nh_n}\right). \tag{7}$$

Corollary 1 *Under the assumptions required in the second part of Theorem 1 the asymptotically optimal smoothing parameter is given by*

$$h_n^* = (6\,\nu(S_\mu))^{1/3} \left(n \int \frac{f_0'^2(x)}{f(x)} d\lambda(x)\right)^{-1/3}.$$

For this value h_n^*, *one has*

$$\mathbf{E}\left[\mathcal{I}(f, \hat{p}_{h_n^*})\right] = (2^{1/3} + 2^{-2/3}) \left(\frac{1}{24} \int \frac{f_0'^2(x)}{f(x)} d\lambda(x)\right)^{1/3} \left(\frac{2\,n}{\nu(S_\mu)}\right)^{-2/3} + o\left(n^{-2/3}\right).$$

As already mentioned the choice of g can often be made such that $\nu(S_\mu) = 1$. Then a pilot estimator can be used to compute an approximated value of h_n^*. Our results are in accordance with the upper bounds given by Barron and Sheu [3] and Barron *et al.* [2]. Up to our knowledge no minimax results are available in the present setting.
In the proof of Theorem 1 the following lemma is used several times. This

lemma extends a technical result given by Berlinet et al (1997). Its proof, which is omitted, relies on the application of the Lebesgue dominated convergence theorem to suitable functions.

Lemma 1 *Let μ and ν be two probability measures with densities f and g with respect to the Lebesgue measure on $\left(\mathbf{R}, \mathcal{B}(\mathbf{R})\right)$. Let $(A_{n,1}, \ldots, A_{n,m_n})$ be a partition of $\mathbf{R}$ satisfying $\nu(A_{n,k}) = h_n = 1/m_n$. The measure μ is supposed to be absolutely continous with respect to ν and to satisfy $\nu\left(\overline{S}_\mu - S_\mu\right) = 0$, where $S_\mu = \{x : f(x) > 0\}$ and $\overline{S}_\mu$ is the closure of S_μ. If $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$ then, for any couple (p, q) of positive integers, we have*

$$\lim_{n\to\infty} h_n \sum_{k=1}^{m_n} \frac{\left(n\mu(A_{n,k})\right)^p}{\left(1 + n\mu(A_{n,k})\right)^q} = \begin{cases} 0 & \text{if} \quad 0 < p < q \\ \nu(S_\mu) & \text{if} \quad p = q \end{cases}$$

and

$$\lim_{n\to\infty} h_n \sum_{k=1}^{m_n} \left(n\mu(A_{n,k})\right)^p \log\left(1 + n\mu(A_{n,k})\right)^q \exp\left[-C\,(n-\kappa)\mu(A_{n,k})\right] = 0,$$

where $C > 0$ and $\kappa \in \mathbf{R}$ are any fixed constants.

Proof of Theorem 1. Let $R_n(x) = \left(\hat{p}_{h_n}(x) - \mathbf{E}\left[\hat{p}_{h_n}(x)\right]\right)/\mathbf{E}\left[\hat{p}_{h_n}(x)\right]$.

$$\mathbf{E}\left[V(h_n)\right] = -\int f(x)\mathbf{E}\left[\log\left(1 + R_n(x)\right)\right] d\lambda(x).$$

Elementary analysis shows that

$$\forall r > -1\,, \quad \left|\log(1+r) - r + \frac{r^2}{2}\right| \le |r|^3 - \log(1+r)\,\mathbf{I}_{(r<-1/2)}\,, \tag{8}$$

so that $\left|\mathbf{E}\left[V(h_n)\right] - \frac{1}{2}\int f(x)\,\mathbf{V}\left(R_n(x)\right) d\lambda(x)\right|$

$$\le \int f(x)\left\{\mathbf{E}\left|R_n(x)\right|^3 + \mathbf{E}\left(\left|\log\left(1 + R_n(x)\right)\right|\mathbf{I}_{(R_n(x)<-1/2)}\right)\right\} d\lambda(x)\,. \tag{9}$$

By Hölder inequality, we get $\mathbf{E}\left|R_n(x)\right|^3 \le \mathbf{E}\left(R_n(x)^4\right)^{3/4}$ and since

$$R_n(x) = \frac{n}{1 + n\mu(A_n(x))}\left[\mu_n(A_n(x)) - \mu(A_n(x))\right],$$

simple calculations give

$$\mathbf{E}\left(R_n(x)^4\right)^{3/4} = \Big(\frac{3\,n^2\mu(A_n(x))^2 + n\mu(A_n(x))}{\big(1+n\mu(A_n(x))\big)^4}\Big)^{3/4}.$$

Consequently,

$$nh_n \int f(x)\,\mathbf{E}\left|R_n(x)\right|^3 d\lambda(x) \le 2^{3/2}h_n \sum_{k=1}^{m_n} \frac{\big(n\mu(A_{n,k})\big)^{5/2}}{\big(1+n\mu(A_{n,k})\big)^3} \quad .$$

Hence by Lemma 1, it follows that

$$\lim_{n\to\infty} nh_n \int f(x)\,\mathbf{E}\left|R_n(x)\right|^3 d\lambda(x) = 0 \tag{10}$$

Now for all $x \in A_{n,k}$, $R_n(x) \in \left[-\dfrac{n\mu(A_{n,k})}{1+n\mu(A_{n,k})}\,;\,\dfrac{n}{1+n\mu(A_{n,k})}\right]$ so that

$$\Big|\log\Big(1+R_n(x)\Big)\,\mathbf{I}_{(R_n(x)<-1/2)}\Big| \le \log\Big(1+n\mu(A_{n,k})\Big).$$

Applying Bernstein inequality, we obtain the following bound for the second term in the right hand side of (9),

$$\begin{aligned} &nh_n \int f(x)\,\mathbf{E}\Big(\big|\log\big(1+R_n(x)\big)\big|\,\mathbf{I}_{(R_n(x)<-1/2)}\Big)\,d\lambda(x) \\ &\le nh_n \sum_{k=1}^{m_n} \mu(A_{n,k}) \log\big(1+n\mu(A_{n,k})\big)\,\exp\big(-\frac{3}{32}\,n\mu(A_{n,k})\big). \end{aligned}$$

Lemma 1 applies and we can conclude that

$$\lim_{n\to\infty} nh_n \int f(x)\,\Big\{\mathbf{E}\Big(\big|\log\big(1+R_n(x)\big)\big|\,\mathbf{I}_{(R_n(x)<-1/2)}\Big)\Big\}\,d\lambda(x) = 0.$$

From this and (10), we get

$$\Big|\mathbf{E}\left[V(h_n)\right] - \frac{1}{2}\int f(x)\,\mathbf{V}\big(R_n(x)\big)\,d\lambda(x)\Big| = o\big(\frac{1}{nh_n}\big)\,. \tag{11}$$

Now by Lemma (1), it is easily seen that

$$\lim_{n\to\infty} \frac{nh_n}{2} \int f(x)\, \mathbf{V}\big(R_n(x)\big)\, d\lambda(x) = \lim_{n\to\infty} \frac{nh_n}{2} \sum_{k=1}^{m_n} \frac{n\mu(A_{n,k})^2}{\big(1+n\mu(A_{n,k})\big)^2}$$
$$= \frac{\nu(S_\mu)}{2}.$$

To prove (7), we decompose the pointwise bias as follows

$$\mathbf{E}\left[\hat{p}_{h_n}(x)\right] - f(x) = (1-a_n)\left[\frac{\mu(A_n(x))}{h_n} - f_0(x)\right] g(x) + a_n[g(x) - f(x)].$$

Now, changing variable in the integrated bias, we obtain

$$B(h_n) = -\int_0^1 \varphi(y)\, \log\Big(1+U_n(y)\Big)\, d\lambda(y),$$

where, for $y \in [0,1]$,

$$U_n(y) = (1-a_n)\left[\frac{\mu\Big(A_n(G^{-1}(y))\Big)}{h_n} - \varphi(y)\right] \frac{1}{\varphi(y)} + a_n\left[\frac{1}{\varphi(y)} - 1\right].$$

Thanks to assumption $(\mathcal{A}_3)$, putting

$$\mathcal{M}_n = \{y \in G^{-1}(S_\mu) \cap [0,1] : U_n(y) < -\varepsilon\},$$

where $0 < \varepsilon < 1$, we have

$$\int_{\mathcal{M}_n} \varphi(y)\, \log\Big(1+U_n(y)\Big)\, d\lambda(y)$$

$$= \int_{\mathcal{N}_n} f(x)\, \log\Big(\frac{\mathbf{E}[\hat{p}_{h_n}(x)]}{f(x)}\Big)\, d\lambda(x) = o(h_n^2)$$

so

$$\int_0^1 \varphi(y)\, \log\big(1+U_n(y)\big)\, d\lambda(y)$$

$$= \int_0^1 \varphi(y)\, \log\big(1+U_n(y)\big)\, \mathbf{I}_{(U_n(y)\geq -\varepsilon)}\, d\lambda(y) + o\left(h_n^2\right).$$

Applying Taylor expansion with integral remainder, we can see that

$$B(h_n) = -\int_{[0,1]/\mathcal{M}_n} \varphi(y) \left[U_n(y) - \frac{1}{2} U_n(y)^2 + R_n(y) \right] d\lambda(y) + o(h_n^2)$$

with

$$R_n(y) = \int_0^{U_n(y)} \frac{(t - U_n(y))^2}{(1+t)^3} \, d\lambda(t) \,,$$

for $U_n(y) \geq -\varepsilon$. Simple arguments lead to

$$\lim_{n\to\infty} \frac{1}{h_n^2} \int_{[0,1]/\mathcal{M}_n} \varphi(y) \, |R_n(y)| \, d\lambda(y) = 0$$

and by checking that

$$0 \leq \int_{[0,1]/\mathcal{M}_n} \varphi(y) U_n(y) \, d\lambda(y) = -\int_{\mathcal{M}_n} \varphi(y) U_n(y) \, d\lambda(y)$$

which is upper bounded by

$$-\int_{\mathcal{M}_n} \varphi(y) \log\Big(1 + U_n(y)\Big) \, d\lambda(y) = o(h_n^2),$$

we get finally

$$B(h_n) = \frac{1}{2} \int_{[0,1]/\mathcal{M}_n} \varphi(y) \, U_n(y)^2 \, d\lambda(y) + o(h_n^2) \,.$$

So, it remains to compute the integral in the above expression. For this, we write by Taylor expansion, for $y \in [(k-1)h_n \,;\, kh_n]$

$$\mu\Big(A_n(G^{-1}(y))\Big) = h_n \varphi(y) + \varphi'(y) h_n \left(kh_n - \frac{h_n}{2} - y\right) + o(h_n^2) \qquad (12)$$

and that gives

$$\begin{aligned} \varphi(y) \, U_n(y)^2 = {} & (1-a_n)^2 \frac{\varphi'(y)^2}{\varphi(y)} \left(kh_n - \frac{h_n}{2} - y\right)^2 \\ & + 2\, a_n (1-a_n) \frac{\varphi'(y)}{\varphi(y)} \left(1 - \varphi(y)\right) \left(kh_n - \frac{h_n}{2} - y\right) \\ & + a_n^2 \left(\frac{(1-\varphi(y))^2}{\varphi(y)} \right) + o(h_n^2). \end{aligned}$$

By using the assumption $nh_n^2 \to \infty$, the second and the third terms in the right-hand side are proved to be negligible in front of h_n^2. Therefore,

$$\int_{[0,1]/\mathcal{M}_n} \varphi(y)\, U_n(y)^2\, d\lambda(y)$$

is equal to

$$(1-a_n)^2 \int_{[0,1]/\mathcal{M}_n} \frac{\varphi'(y)^2}{\varphi(y)} \left(kh_n - \frac{h_n}{2} - y\right)^2 d\lambda(y) + o\left(h_n^2\right).$$

Now, apply the mean formula. Since φ'^2/φ is continous on $[0,1]$, for all $k \in \{1, \ldots, m_n\}$ there exists $\xi_k \in](k-1)h_n; kh_n[$ such that

$$\int_0^1 \frac{\varphi'(y)^2}{\varphi(y)} \left(kh_n - \frac{h_n}{2} - y\right)^2 d\lambda(y)$$

$$= \sum_{k=1}^{m_n} \frac{\varphi'(\xi_k)^2}{\varphi(\xi_k)} \int_{(k-1)h_n}^{kh_n} \left(kh_n - \frac{h_n}{2} - y\right)^2 d\lambda(y)$$

$$= \frac{h_n^2}{12} \sum_{k=1}^{m_n} h_n \frac{\varphi'(\xi_k)^2}{\varphi(\xi_k)}.$$

From this result and by checking that $\int_{\mathcal{M}_n} \varphi'^2(y)/\varphi(y)\, d\lambda(y)$ tends to zero as $\lambda(\mathcal{M}_n)$ does, we obtain

$$\lim_{n\to\infty} \frac{1}{h_n^2} B(h_n) = \lim_{n\to\infty} \frac{1}{24} \sum_{k=1}^{m_n} h_n \frac{\varphi'(\xi_k)^2}{\varphi(\xi_k)}$$

$$= \frac{1}{24} \int_0^1 \frac{\varphi'(y)^2}{\varphi(y)}\, d\lambda(y) = \frac{1}{24} \int \frac{f_0'(x)^2}{f(x)}\, d\lambda(x).$$

□

Acknowledgments

We thank the referees for interesting questions and remarks.

References

1. A. R. Barron, in *Proc. IEEE Int. Symp Inform. Theory,* Kobe, Japan, (1988).
2. A. R. Barron, L. Györfi, and E. C. van der Meulen, *IEEE Trans. on Information Theory* **38**, 1437 (1992).
3. A. R. Barron and C. Sheu, *Ann. Statist.* **19**, 1347 (1991).
4. A. Berlinet G. and Biau, *Iterated modified histograms* (Submitted for publication, 2001).
5. A. Berlinet, L. Györfi snd E. C. van der Meulen, *Publications de l'ISUP.* **41**, 3 (1997).
6. A. Berlinet, I. Vajda and E. C. van der Meulen, *IEEE Trans. on Inform. Theory* **44**, 999 (1998).
7. L. Györfi, F. Liese, I. Vajda and E. C. van der Meulen, *Statistics* **32**, 31 (1998).
8. L. Györfi, and E. C. van der Meulen, in *Transactions of the Twelfth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, 88 (1994).
9. P. Hall, *Ann. Statist.* **15**, 1491 (1987).

THE ASYMPTOTIC DISTRIBUTION OF SPACINGS OF ORDER STATISTICS

MIKELIS G. BICKIS

Mathematical Sciences Group, Department of Computer Science, University of Saskatchewan, 106 Wiggins Road, Saskatoon, SK S7N 5E6, Canada
E-Mail: bickis@snoopy.usask.ca

It is well known that for an i.i.d. sample from a uniform distribution, the spacings between the order statistics are asymptotically distributed like i.i.d. exponential random variables. Although the spacings between order statistics from an arbitrary distribution are not in general identically distributed, one can still consider the distribution function to which the empirical cdf of such spacings from an i.i.d. sample would converge as the sample size goes to infinity. The limit of this empirical cdf is called the asymptotic spacing distribution. If X is a random variable with density f, then the asymptotic spacing distribution is $1 - L$, where L is the Laplace transform of the distribution of $f(X)$. Although the explicit form for L is not tractable in general, one can use Tauberian theorems to relate the tail behaviour of L to the order of contact of f to the x-axis: The more gradual the approach of f to the axis, the heavier the tail of the asymptotic spacing distribution.

1 Introduction

Imagine a large number of marathon runners crossing a finish line, and consider the set of time intervals between their arrivals. What would be the distribution of the lengths of these intervals? The question is ill-posed, since in general these lengths would not have the same distribution. Nonetheless, one can calculate the empirical distribution of such a collection and ask the question "Is there a probability distribution to which this empirical distribution converges as the sample size goes to infinity?"

It was noted by Katz [1] that the empirical distribution appears to follow a "power law". More precisely, one could say that the tail of the distribution function behaves like a negative power of its argument (although, of course, integrability requirements prohibit such behaviour extending to the origin). This observation led Katz to speculate the spacings in an ordered Normal sample would have such a "power law" distribution.

In this paper, we study the relationship between a parent distribution and the asymptotic distribution of spacings between its order statistics. We characterize the distributions for which the spacings distribution has a negative power tail, and discover that these do not, in fact, include the Normal distribution.

2 Empirical distribution of spacings

Let $X_1, \ldots, X_n$ be i.i.d. from a distribution F, and define

$$Y_i = X_{(i+1)} - X_{(i)}, \quad i = 1, \ldots, n-1$$

to be the spacings between consecutive order statistics. We will assume in what follows that F is absolutely continuous.

It is well known that if F is the uniform distribution on the unit interval then nY_i, $i = 1, \ldots, n-1$ are asymptotically i.i.d. exponential, and thus it makes sense to talk about the asymptotic distribution of the spacings. In general, however, the Y_i's will not be identically distributed, even asymptotically, so care is needed in defining what is meant by "the" asymptotic distribution.

Let

$$\hat{S}_n(t) = \frac{1}{n-1} \sum_{i=1}^{n-1} I_{(t,\infty)}(nY_i) \tag{1}$$

be the empirical survivor function of the scaled spacings, where $I_{(t,\infty)}(\cdot)$ represents the indicator function of the interval (t, ∞). (Formulas look simpler in this context if we deal with survivor functions rather than cumulative distribution functions. It should cause no confusion if we refer to a distribution by its survivor function rather than its cdf.) Suppose that there is a distribution S such that with probability 1, $\hat{S}_n$ converges weakly to S. Then we will call S the *asymptotic spacings distribution* (ASD) of the distribution F.

Pyke [2] considered the limit of (1), and presented an elegant proof that

$$\hat{S}_n(t) \to \int_{-\infty}^{\infty} e^{-tf(x)} f(x)\, dx \quad \text{in probability,}$$

where $f = F'$ is the probability density function of the original sample. He added, without proof, the stronger statement that the convergence is, in fact, uniform with probability 1, referring to a paper by Blum and Weiss [3].

Monte-Carlo investigation led the present author to speculate that the empirical distribution of spacings is, moreover, asymptotically equivalent to the empirical distribution of an appropriate i.i.d. sample, as suggested by the following heuristic argument, under the additional assumption that the density f is continuous.

Conjecture: Let $X_1, \ldots, X_n, \ldots$ be i.i.d. from a distribution F with a continuous density f, and let $Z_1, \ldots, Z_n, \ldots$ be i.i.d. from an exponential distribution with mean 1, independent of the X_i's. Then the empirical spacings distribution $\hat{S}_n(t)$ converges to the same limiting distribution as the empirical distribution from the i.i.d. sample $Z_1/f(X_1), \ldots, Z_n/f(X_n)$.

Heuristic Argument: By the mean value theorem

$$F(X_{(i+1)}) = F(X_{(i)}) + f(V_i)(X_{(i+1)} - X_{(i)})$$

for some $X_{(i)} < V_i < X_{(i+1)}$. Thus

$$F(X_{(i+1)}) - F(X_{(i)}) = f(V_i)Y_i. \tag{2}$$

However, since $F(X_i)$ has a uniform distribution, the left side of (2) gives the spacings of an i.i.d. sample from the uniform distribution. Thus, the joint distribution of

$$nf(V_1)Y_1, \ldots, nf(V_n)Y_n$$

will approach that of $Z_1, \ldots, Z_n$. For large samples, the order statistics will approximate the appropriate quantiles, as will the V_i's sandwiched between them. Thus, the joint distribution of $nY_1, \ldots, nY_n$ will approximate that of $Z_1/f(X_{(1)}), \ldots, Z_n/f(X_{(n)})$.

The random function

$$S_n^*(t) = \frac{1}{n-1} \sum_{i=1}^{n-1} I_{(t,\infty)} \left(Z_i/f(X_{(i)}) \right),$$

on the other hand, will have the same distribution as

$$\tilde{S}_n(t) = \frac{1}{n-1} \sum_{i=1}^{n-1} I_{(t,\infty)} \left(Z_i/f(X_i) \right), \tag{3}$$

which is the empirical survivor function from the i.i.d. sample $Z_1/f(X_1), \ldots, Z_n/f(X_n)$.

A rigorous proof of the conjecture will depend first on a careful specification of the nature of the putative asymptotic equivalence between (1) and (3). A strong result would obtain if it could be shown that the stochastic process $\sqrt{n}\left(\hat{S}_n(t) - S(t)\right)$ converges to the same limit as $\sqrt{n}\left(\tilde{S}_n(t) - S(t)\right)$.

The asymptotic spacings distribution

$$S(t) = \int_{-\infty}^{\infty} e^{-tf(x)} f(x)\, dx \tag{4}$$

is just the distribution of the random variable $Z/f(X)$, which, by conditioning on X, can be viewed as a mixture of exponential distributions.

Example 1. If F is exponential, then the spacings themselves are exponential with $E(Y_i) = 1/(n-i)$ and the distribution of $\hat{S}_n$ can be calculated explicitly for finite n.

We have

$$\begin{aligned} E\left(\hat{S}_n(t)\right) &= (n-1)^{-1} \sum_{i=1}^{n-1} \Pr\{Y_i > t/n\} \\ &= (n-1)^{-1} \sum_{i=1}^{n-1} e^{-(n-i)t/n} \\ &= \frac{1 - \exp(-(1-1/n)t)}{(n-1)(e^{t/n}-1)} \end{aligned}$$

from which we can see that $E\left(\hat{S}_n(t)\right) \to (1-e^{-t})/t$ as required by (4).

Example 2. Consider a random variable with density function

$$f(x) = \left(\frac{x}{k+1}\right)^k \quad 0 \le x \le k+1$$

where $k > 0$ is a parameter. (These distributions are members of the Beta family, rescaled for ease of computation.) Making the change of variable

$$v = t\left(\frac{x}{k+1}\right)^k$$

we can calculate

$$\begin{aligned} S(t) &= \int_0^{k+1} \exp\left[-t\left(\frac{x}{k+1}\right)^k\right] \left(\frac{x}{k+1}\right) dx \\ &= \frac{1+1/k}{t^{1+1/k}} \int_0^t e^{-v} v^{1/k}\, dx \\ &= \frac{\Gamma(2+1/k) G_{1+1/k}(t)}{t^{1+1/k}}, \end{aligned}$$

where $G_{1+1/k}(t)$ is the Gamma cdf with shape parameter $1+1/k$.

These two examples illustrate ASD's with tails behaving like negative powers. Such distributions have at most a finite number of moments. In

particular, the expected asymptotic spacing length will be

$$\begin{aligned}\int_0^\infty S(t)dt &= \int_0^\infty \int_{f(x)>0} e^{-tf(x)} f(x)\, dx\, dt \\ &= \int_{f(x)>0} f(x)/f(x)\, dx \\ &= \infty\end{aligned}$$

unless the support of f is bounded. Thus the distribution of an unbounded random variable, such as the Normal, cannot have an ASD with a tail $O(t^{-\alpha})$ with $\alpha > 1$ for then its expectation would be finite.

3 Characterization of asymptotic spacings distributions

Note that the ASD is in fact a Laplace transform. If X has density f, consider the random variable $W = f(X)$. Then $S(t) = E(e^{-tW})$. If f is monotone, then the density of W is given by

$$\frac{f \circ f^{-1}(w)}{|f' \circ f^{-1}(w)|} = \frac{w}{|f' \circ f^{-1}(w)|}$$

provided the density is differentiable. (If f is not monotone, then the density of W must include contributions from the components of $f^{-1}(w)$).

Thus, for example, if X is Exponential, then W is uniform with Laplace transform $(1 - e^{-t})/t$ as seen previously. If X is Uniform, W is degenerate at 1, with Laplace transform e^{-t}.

The representation of the ASD as a Laplace transform allows one to use Tauberian theorems [4] which relate the tail behaviour of the Laplace transform with the behaviour of the distribution near the origin.

Specifically, if L is the Laplace transform of a distribution F, then as $t \to 0$

$$F(t) \sim L(1/t)/\Gamma(\rho + 1)$$

where

$$\rho = \lim_{t \to 0} \frac{\log(F(tx)/F(t))}{\log x}.$$

We see that what controls the tail behaviour of the ASD is the behaviour near the origin of W, which is determined by the way that the density f approaches the axis. It can readily be seen that if $f(x)$ behaves like $O(x^k)$ as $x \to 0$, then W will have a density like $O(w^{1/k})$ as $w \to 0$. The cdf of W will then behave like $O(w^{1/k+1})$ and the ASD will have a tail like $O(t^{-(1+1/k)})$.

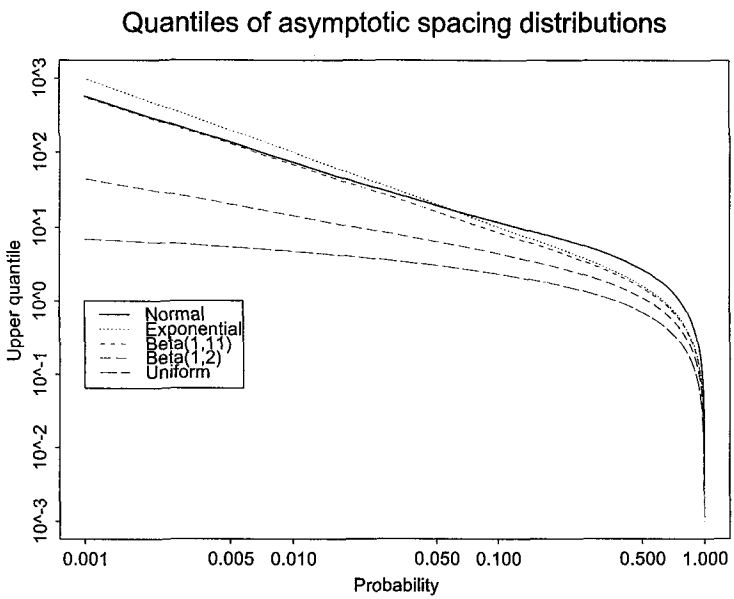


Figure 1. Upper quantiles of asymptotic spacings distributions from selected parent distributions

Obviously, the distribution of W will be unaffected by location shifts or reflections, and will be reciprocally scaled by rescaling of X. Thus the tail behaviour of the ASD is the same for all distributions in a location-scale family.

If X is Normal, then the density of W is messy, but after some calculation can be seen to be approximately proportional to $(-\log w)^{-1/2}$ near the origin. This does not behave like a power of w, indicating that the ASD will not have a negative-power tail. By making the change of variable $s = -\log w$, we can see that the cdf of W will be approximately proportional to

$$\int_{-\log w}^{\infty} s^{-1/2} e^{-s}\, ds.$$

It follows that the tail of the ASD will be like

$$\int_{\log t}^{\infty} s^{-1/2} e^{-s}\, ds,$$

i.e., like the tail of $\exp(C)$ where C has a χ^2_1 distribution.

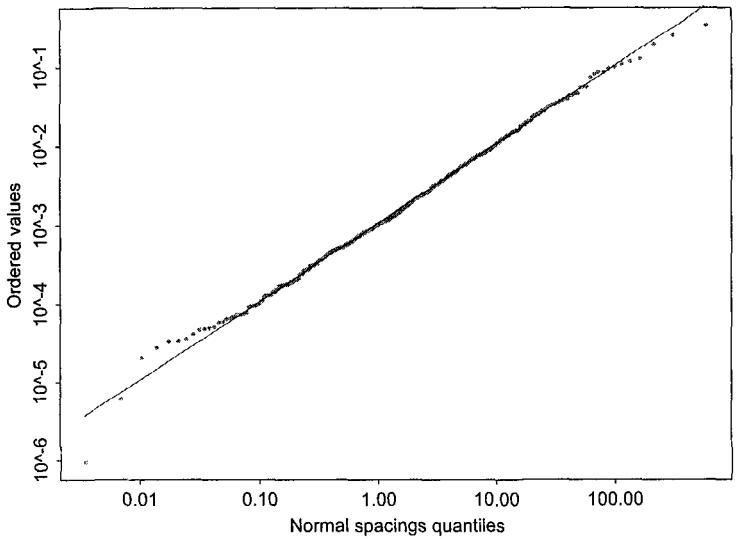


Figure 2. Q-Q-plots of Normal spacings

4 Numerical results and simulations

The quantiles for the ASD corresponding to given parent distributions can be computed numerically. These are plotted in Figure 1 for distributions with different tail behaviours.

One can examine the suitability of the ASD for describing the empirical distribution of spacings for large but finite sample sizes. Figures 2 and 3 present Q-Q-plots for spacings of samples 1000 *i.i.d.* observations from the Normal and Exponential distributions, respectively, plotted against the respective ASD. These plots demonstrate a reasonably good fit. The plot is presented on a log-log scale to accommodate the heavy-tailed distribution. Note that on such a scale, a well-fitting plot should lie on a straight line *with unit slope.* Such a line is superimposed on the points, constrained to pass through the empirical upper 10 percentile.

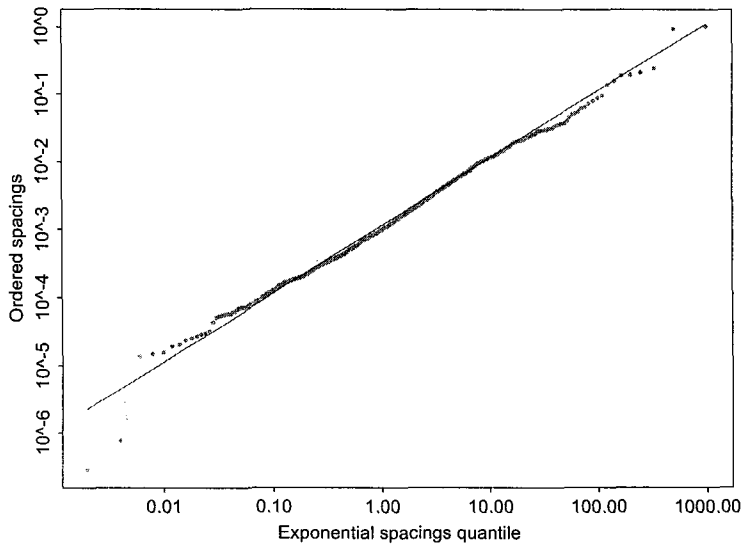


Figure 3. Q-Q-plots of Exponential spacings

Acknowledgments

I want to thank Leon Katz for raising the question that spawned this paper. I am indebted to Richard Lockhart and Michael Stephens for useful discussions on this topic. This research was supported by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

1. L. Katz, *Personal communication*, 2000.
2. R. Pyke, *J. Roy. Stat. Soc.* **27B**, 395 (1965).
3. J. R. Blum and L. Weiss, *Ann. Math. Stat.* **28**, 242 (1957).
4. W. Feller, *An Introduction to Probability Theory and its Applications v. 2, 2nd ed.* (Wiley, New York, 1971).

THEORETICAL AND COMPUTATIONAL ISSUES IN BAYESIAN ANALYSIS OF MULTIVARIATE ORDINAL CATEGORICAL DATA WITH REFERENCE TO AN OPHTHALMOLOGIC STUDY

ATANU BISWAS

Applied Statistics Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700 108, India

E-mail: atanu@isical.ac.in

Bivariate or multivariate ordinal categorical data is quite common in different real life situations. Several frequentist's approaches are available for the analysis of such data. In almost all those approaches, the computational burden is tremendous. As an alternative, some Bayesian latent variable based approach were suggested by some authors. In such cases also, the computation is a key issue. The computational burden of such Bayesian analyzes can be remarkably reduced by using the software WinBUGS. The present paper discusses one real life dataset on an ophthalmologic study, called the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR), which provides bivariate ordinal categorical data with several eye-specific and subject-specific covariates. The techniques of the available Bayesian analyzes with reference to the WESDR data are discussed in the present paper. Some results on Bayesian model selection for this data is also discussed. By the help of some exploratory data analysis an appropriate model is selected, and it is then analyzed in the Bayesian semiparametric way taking a Dirichlet process prior for the random effect (in the frequentist's sense). The final result on the analysis of WESDR data is presented. Several computational issues in this context are also discussed. Finally, the applicability of the present approach to some more general situation are also pointed.

1 Introduction

In several social, psychological and biomedical studies the response variable is ordinal categorical in nature, sometimes they are ordinal due to the absence of well-defined non-invasive direct measurements (*e.g.*, mild, moderate, severe, etc.). If the response is of multivariate nature and each component of the response is ordinal categorical, we have multivariate ordinal categorical data to deal with. This kind of data are available in many real life situations.

Ever since Dale [7] proposed the analysis of bivariate ordinal categorical data, a lot of subsequent studies were carried out in this interesting and important research area. Much of the early works were done on the frequentist's view point. Molenberghs and Lesaffre [27] used a multivariate Placket distribution as an extension of Dale's [7] model. These likelihood methods are, of course, computationally extensive. As an alternative, Williamson *et al.* [37] considered

the generalized estimating equations (GEE) approach. They fitted cumulative probit margins and a global odds ratio association model. Williamson *et al.* [38] discussed the applicability and usefulness of their computer programs GEECAT (generalized estimating equations for categorical data) and GEEGOR (generalized estimating equations using global odds ratio). Kim [19] carried out a latent variable based estimation technique for a bivariate ordinal data of an ophthalmologic study called the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR). Kim [19] used the Newton-Raphson iteration technique to find the estimates of the underlying parameters including the typically unknown cut-off points. But he admitted the computational difficulty of the technique in a more general set up. Williamson and Kim [36], also with reference to the WESDR dataset, considered bivariate latent variable regression model using global odds ratio, which required no specific choice of underlying latent distribution except its continuity, and no specific structure of the correlation. A quasi-Newton method in a full maximum likelihood procedure was employed for the estimation of the model parameters. Kim *et al.* [20] discussed regression models for bivariate ordered categorical data from ophthalmologic studies.

Some statisticians wanted to look at the multivariate ordinal categorical data from a Bayesian philosophy and tried to apply the Bayesian theory for analyzing such a multivariate ordinal categorical data with covariates. It is observed that, besides the difference in philosophical perspective, the Bayesian computation becomes more tractable and one may arrive at good solutions after some easy computational effort. In the present paper we discuss some of the available approaches and discuss some possible new directions, mainly with reference to a real life bivariate ordinal dataset obtained from the WESDR study. We discuss the WESDR study and the nature of the resulting bivariate ordinal data briefly in the next section. In Section 3, two available analyzes on WESDR data are briefly indicated. A discussion on appropriate model selection is provided in Section 4. In Section 5, a Bayesian analysis of the WESDR data using the selected model from Section 4 is carried out. The results are briefly discussed. Some important computational issues are discussed in Section 6. Finally, Section 7 ends with some concluding remarks.

2 WESDR

2.1 Description of the WESDR

It was a population-based study conducted in Southern Wisconsin between 1980 and 1982 by Dr. Ronald Klein and his medical colleagues at the University of Wisconsin and supported by the National Eye Institute, NIH. In the baseline study, a total of 996 insulin-taking, younger onset diabetic patients were examined using standard protocols to determine the prevalence and severity of diabetic retinopathy and associated risk variables. The population of the study consisted of a probability sample selected from 10135 diabetic persons who received primary care in an 11-county area in southern Wisconsin from 1979 to 1980. A detailed description of the population can be available in Klein *et al.* [21]. Of the younger-onset persons (less than 30 years of age), 996 participated in the baseline examination (1980 to 1982). Subsequently there were two follow-up examinations on the same population after 4 and 10 years (cf. Klein *et al.* [24]; Klein *et al.* [22]), with some missing data in this course of study. The baseline and the two follow-up examinations were performed in a mobile examination van in or near the city where the participants lived. The ocular and physical examinations included taking stereoscopic color fundus photographs of seven standard fields. The basic goals of the study (cf. Klein *et al.* [25]) were

- To find the associated risk factors which are important in planning a well-coordinated approach to the public health problem posed by the complications of diabetes (cf. Hamman [17]; Rand [30]).
- Identifying the possible patients at high risk level of severe retinopathy was quite important for advising ophthalmologic care.
- Planning future studies such as controlled clinical trials of treatment of diabetes and diabetic retinopathy (cf. Rand [30]; Palmberg *et al.* [28]).
- Progression of diabetic retinopathy over time and associated factors was also an important aspect of study (cf. Wahba *et al.* [33]).

2.2 Data description

WESDR data is a bivariate ordinal data. The retinopathy scale (RS) provided in the dataset is a more current one than the one used in some earlier works. Both the right and left eye retinopathy severity levels are recorded as two components of the bivariate response. The possible values of the retinopathy severity levels are 10, 21, 31, 37, 43, 47, 53, 60, 61, 65, 71, 75 and 85, corresponding to increasing levels of severity of retinopathy within an eye. A commonly used grouping is 10, 21-37, 43-53 and 60-85, which correspond to

no retinopathy (category 0), mild nonproliferative retinopathy (category 1), moderate to severe nonproliferative retinopathy (category 2), and proliferative retinopathy (category 3), respectively. Such a grouping is usually done for easy interpretation of the data, and it reduces the computational burden to a great extent. Most of the available works are done using this grouping.

Three eye-specific covariates are recorded separately for each eye. The first one is the presence or absence of macular edema (ME), which is the effusion of serious fluid into the intersects of cells in tissues. The right and left eye refractive error (RE) in diopters, which is the ability of the eye to refract light which enters into it so as to form an image on the retina, is also recorded. The values can be negative or positive, negative values represent myopia (nearsightedness), and positive values represent hyperopia (farsightedness). In addition, the right and left eye intraocular pressure (IOP) in mmHg. is also measured.

In addition there are 11 person-specific covariates. The first two are age at diagnosis (AgD) of diabetes in years and duration of diabetes (DuD) in years. To get current age at the time of examination, one has to add AgD and DuD. The other person-specific covariates are glycosylated hemoglobin (GH) in percent (a measure of control of blood sugar where lower values are considered better), systolic and diastolic blood pressures (SBP & DBP) in mmHg., body mass index (BMI) in kilograms per meter squared (using weight and height), pulse rate (PR) in beats per 30 seconds, sex, urine protein (UP) (present/absent), doses of insulin (DI) per day and area of residence (AR) (urban/rural).

3 Bayesian Analyses

3.1 Bayesian Analysis of Baseline Data

The first major Bayesian analysis of the WESDR (baseline) data is due to Biswas and Das [2]. They considered 691 observations with full response and covariate information for their analysis. Suppose y_{Li} and y_{Ri} be the responses from the left and right eye, respectively, for the ith individual. Both y_{Li} and y_{Ri} can take values 0, 1, 2 and 3. It is assumed that there are some underlying latent variables which are unobservable, but in effect we observe these y_{Li} and y_{Ri}. Underlying latent variables y^*_{Li} and y^*_{Ri} are postulated, which are responsible for observing y_{Li} and y_{Ri} in the following way:

$$y_{Li} = j \text{ if } y^*_{Li} \in (\gamma_{1j}, \gamma_{1j+1}], \quad j = 0, 1, 2, 3,$$
$$y_{Ri} = l \text{ if } y^*_{Ri} \in (\gamma_{2l}, \gamma_{2l+1}], \quad l = 0, 1, 2, 3,$$

where γ_{1j}'s and γ_{2l}'s are typically unknown cut-off points such that $\gamma_{s0} < \gamma_{s1} < \cdots < \gamma_{s4}$ for $s = 1, 2$. Some mild conditions on these γ_{1j}'s and γ_{2l}'s, such as

$$(1)\ \gamma_{10} = \gamma_{20} = -\infty,\quad \gamma_{14} = \gamma_{24} = \infty,$$
$$(2)\ \gamma_{11} = \gamma_{21} = \text{ a known constant,}$$

are needed for identifiability (see Chen and Shao [3], for details). The latent vector $y_i^* = (y_{Li}^*, y_{Ri}^*)^T$ is assumed to follow a bivariate normal distribution $N_2(X_i\beta, \Sigma)$. The joint distribution of $y_i = (y_{Li}, y_{Ri})^T$, y_i^*'s are written as

$$\begin{aligned}
&\pi(y_i^*, y_i, i = 1, \cdots, N | \beta, \Sigma, \gamma) \\
&= \prod_{i=1}^{N} \left[\sum_{j,l \in S} 1_{jl}^i \times I(y_{Li}^* \in (\gamma_{1j}, \gamma_{1j+1}], y_{Ri}^* \in (\gamma_{2l}, \gamma_{2l+1}]) \right] \\
&\times N_2(X_i\beta, \Sigma), \qquad (1)
\end{aligned}$$

where $N = 691$, the number of individuals under study; γ is the collection of all unknown γ_{1j}'s and γ_{2l}'s; $S = \{(j, l) : j = 0, 1, 2, 3; l = 0, 1, 2, 3\}$; $I(X \in A) = 1$ or 0 according as $X \in A$ or not; and $1_{jl}^i = 1$ or 0 according as $y_i^* = (j, l)^T$ or not. Without sufficient prior knowledge, noninformative priors for β, Σ and γ are taken. The conditional posteriors of the parameters are of known form. The computations can be carried out using Markov Chain Monte Carlo (MCMC) technique. In particular, posterior summary statistics of all the parameters are obtained using the software WinBUGS.

The analysis shows that the severity of retinopathy among the younger onset diabetic persons is directly affected by DuD and DBP. Also GH was one vital covariate for retinopathy.

3.2 Analysis of the Baseline and 4-year Follow-up Data

Das and Biswas [9] considered the analysis of bivariate ordinal data repeated over time. They took the WESDR data for two time points only (baseline and 4-year follow-up data). Some notational adjustments are to be needed from the subsection 3.1. In place of y_{Li}, y_{Ri}, y_{Li}^*, y_{Ri}^*, 1_{jl}^i, X_i we just write y_{Lit}, y_{Rit}, y_{Lit}^*, y_{Rit}^*, 1_{jl}^{it}, X_{it} to indicate that they are for time point t, $t = 0, 1$. For the baseline data we write $t = 0$ and for the 4-year follow-up data we write $t = 1$. The joint distribution of $y_{it} = (y_{Lit}, y_{Rit})^T$, $y_{it}^* = (y_{Lit}^*, y_{Rit}^*)^T$'s are now written as

$$\pi(y_{it}^*, y_{it}, i = 1, \cdots, n, t = 0, 1 | \beta, \Sigma, \gamma)$$

$$= \prod_{i=1}^{n} \prod_{t=0}^{1} \left[\sum_{j,l \in S} 1_{jl}^{it} \times I(y_{Lit}^* \in (\gamma_{1j}, \gamma_{1j+1}], y_{Rit}^* \in (\gamma_{2l}, \gamma_{2l+1}]) \right]$$
$$\times N_2(X_{it}\beta + Z_{it}b_i, \Sigma), \tag{2}$$

where the sample size n is now 548, as some of the individuals of the baseline study were missing in the follow-up study. A random effect model (in the frequentist's sense) is assumed for y_{it}^* as

$$y_{it}^* = X_{it}\beta + Z_{it}b_i + \epsilon_{it},$$

where b_i, a q-dimensional vector, is the ith cluster (subject) effect, which is responsible for the longitudinal dependence. Here ϵ_{it} is assumed to be bivariate normal with mean vector 0 and dispersion matrix Σ. For likelihood identifiability, Σ needs to be a correlation matrix (see Chib and Greenberg [5]; Chen and Shao [3]). The prior for β is taken as $N_p(\beta_0, \Sigma_0)$. Without sufficient prior information a relatively vague prior is chosen by setting $\beta_0 = 0$ and $\Sigma_0 = 10^4 \times$ identity matrix. This ensures the posterior to be driven by data. A noninformative prior for γ is taken. A prior proportional to the normal density (with known mean ρ_0 and known variance τ^{-1}) in the domain $(-1, 1)$ is taken for ρ, the only correlation parameter in the 2×2 correlation matrix Σ. A Dirichlet Process (DP) prior for the unknown distribution G of b_i's are taken (see Ferguson [12]; Kleinman and Ibrahim [26]). The software WinBUGS is used for the Markov Chain Monte Carlo (MCMC) computations. Metropolis-Hastings algorithm (Hastings [18]) was employed for sampling from the posterior of ρ, as it becomes non-standard; for other parameters the popular Gibbs sampler (see Geman and Geman [16]; Gelfand *et al.* [13]) is used. Note that in case Σ is a proper dispersion matrix, a suitable inverted Wishart prior could be taken for Σ. In that case the Bayesian MCMC computations could be same, possibly requiring some smaller time as the conditional posterior of Σ would be inverted Wishart also.

3.3 Present Analysis Using all the Data

The WESDR data provides information of only baseline values (and not of the 4-year follow-up) on $N - n = 143$ individuals. Das and Biswas [9] ignored this information. Now this data on 143 individuals can be successfully used to get more information. The likelihood thus becomes

$$\pi(y_{it}^*, y_{it}, i = 1, \cdots, n, t = 0, 1; y_{i0}^*, y_{i0}, i = n+1, \cdots, N | \beta, \Sigma, \gamma)$$

$$= \prod_{i=1}^{n} \prod_{t=0}^{1} \left\{ \left[\sum_{j,l \in S} 1_{jl}^{it} \times I(y_{Lit}^* \in (\gamma_{1j}, \gamma_{1j+1}], y_{Rit}^* \in (\gamma_{2l}, \gamma_{2l+1}]) \right] \right.$$
$$\left. \times N_2(X_{it}\beta + Z_{it}b_i, \Sigma) \right\}$$
$$\times \prod_{i=n+1}^{N} \left\{ \left[\sum_{j,l \in S} 1_{jl}^{i0} \times I(y_{Li0}^* \in (\gamma_{1j}, \gamma_{1j+1}], y_{Ri0}^* \in (\gamma_{2l}, \gamma_{2l+1}]) \right] \right.$$
$$\left. \times N_2(X_{i0}\beta + Z_{i0}b_i, \Sigma) \right\}, \qquad (3)$$

after modifying (2). Our ultimate goal is to analyze the baseline and 4-year follow-up data by making full use of the data after suitable prior elicitation and model selection in terms of covariate selection. These issues are discussed in the next section.

4 Prior Elicitation and Model Selection

Most of the priors considered so far are vague priors. An empirical Bayes procedure could be carried out if some past data were available. For example, Angers and Biswas [1] considered the analysis of the 4-year follow-up data only. They used the baseline data to construct priors for different parameters. Some sensitivity analyses were also carried out by Angers and Biswas [1]. But, in the absence of such past information, the vague priors are appropriate. In such a case, the posterior will be driven by the data. In fact, we have a quite large dataset, and the data should dominate the posterior in any case.

In the context of the 4-year data of the WESDR, model selection in terms of covariate inclusion was carried out by Angers and Biswas [1]. They computed the standardized Bayes estimators and ordered them. Then the model with highest marginal probability is selected. It was observed in their study that the baseline values, DBP and GH are important covariate in different situations. They included these covariates in their model to analyze the data. But if we carry out our analysis simultaneously for the baseline and 4-year follow-up data, there is no question of taking the baseline values as covariates. The covariates which are not important can be excluded from considerations to ease the computational burden.

Wahba *et al.* [33] carried out their analysis on a subgroup of the younger onset population, consisting of 669 subjects with no or nonproliferative retinopathy. They considered retinopathy scale (RS) as the response variable. Thus they reduced the bivariate ordinal data problem to a simpler univariate ordinal data problem. Some exploratory GLIM modeling using the SAS procedure LOGISTIC (SAS Institute [31]) were carried out and after some exploratory

considerations they took the model. Wahba *et al.* [33] observed that the effect of GH was very strong and fairly linear in the logit and that the effects of AgD, DuD and BMI were strong and nonlinear. We have done a similar detailed exploratory study using SAS with the baseline and 4-year data. We observe that the effect of DuD in a log-scale explains the data in a better way. Note that Wahba *et al.* [33] also observed that DuD has a nonlinear effect. We also find that the effects of GH and DBP are significant in the context of retinopathy levels. In fact, Klein *et al.* [23] reported that GH is a strong predictor of progression of diabetic retinopathy in the younger onset group. Angers and Biswas [1] also observed the same scenario. DBP came out as an important covariate in the study of Kim [19] and Biswas and Das [2]. In the present study we observe possible interaction between GH and DBP. Thus the model has to be taken with care. This is discussed in Section 5.

5 Present Bayesian Analysis With the Selected Model

We now carry out a Bayesian semiparametric analysis of the baseline and 4-year follow-up data of the WESDR. We take the selected covariates (discussed in Section 4) into the model, suitable transformations and interaction effects are also considered following our exploratory study described in the previous section. This seems to be a correct model which can explain the retinopathy levels in terms of different covariates. We use all the available data, i.e., the likelihood is given by (3). We impose an additional restriction $\gamma_{1j} = \gamma_{2j}$ for all j, which seems logical as the two eyes should behave in the same way. The random effect model (in the frequentist's sense) is justified from intuitive feelings. Besides the significant covariates taken in the model (after excluding the insignificant ones) there must be several unobserved (most of which are unobservable) factors that can be responsible for retinopathy and also the association between the retinopathy levels between two eyes. As the two eyes of the same human being are considered, several nerves, tissues, same reading habit and work habit, same environmental exposure, etc. are responsible for similar effects on both the eyes. These unobservable factors are expressed in terms of the uncertainty b_i's.

The model is then taken as

$$y^*_{kit} = constant + \log(DuD) \times \beta_{DuD} + GH \times \beta_{GH} + DBP \times \beta_{DBP} + (GH \times DBP) \times \beta_{GH \times DBP} + b_{ki} + \epsilon_{kit}, \quad k = L, R.$$

A Bayesian semiparametric study as in Das and Biswas [9] was carried out with different choices of the prior. The prior for β is taken as a vague prior. This is taken by setting $\beta \sim N_p(0, \Sigma_0)$, where $\Sigma_0 = 10^4 \times$ identity matrix, i.e., by

setting the variances to be large. Some noninformative prior proportional to unity is set for γ. The prior for ρ is the same discussed in subsection 3.1. A Dirichlet process prior is taken for $b_i = (b_{Li}, b_{Ri})^T$'s such that

$$b_i \sim G,$$
$$G \sim DP(\alpha, G_0),$$

with $G_0 \sim N_m(0, \Gamma)$, with some known Γ. When G_0 is known, i.e., G has a prior probability model with known hyperparameters, the posterior distributions can be easily obtained as in Wilks *et al.* [35]. Here we have set $\Gamma =$ identity matrix. The posteriors are observed to be dominated by the data. We primarily take priors of the form discussed in Section 3.

Now we discuss the nature of the conditional posteriors without providing the exact mathematical expressions. The conditional posterior of β is multivariate normal. The conditional posterior of ρ is non-standard, to sample from it one has to use the Metropolis-Hastings algorithm (see Hastings [18]). With some probability $\alpha(\rho, \rho')$ (easily computable) we move to a candidate value ρ' from the present value ρ, and with probability $1 - \alpha(\rho, \rho')$ we stick to ρ. We take $\rho' = \rho + h$, where h is a random zero mean increment. The variance of h is usually taken to be of order $O(\frac{1}{n})$. The conditional posterior of γ_{1j} is found to be uniform over some domain. The conditional posterior of y_{it}^* is truncated bivariate normal, truncated in some rectangular region. Finally, the conditional posterior of b_i is a mixture where one piece is normal and the others are point masses. That is, with some probability we choose $b_i = b_j$ and otherwise we sample from the normal density. For detailed mathematical forms of the conditional posteriors, one can see Das and Biswas [9]. The derivation of the present case is similar to Das and Biswas [9] with a slight possible change in some expressions.

The posterior summary statistics of the relevant parameters are then obtained. From our final analysis, the posterior mean of β_{DuD} is 0.78, that for β_{GH} is 0.52 and for β_{DBP} is 0.32. The posterior mean of $\beta_{GH \times DBP}$ is 0.27. Thus all of DuD, GH and DBP have positive effect to the retinopathy on both the eyes in the sense that presence of any of them increases the retinopathy levels. The interaction of GH and DBP is also affecting the retinopathy in a positive way. The posterior mean of ρ, the polychoric correlation (correlation between two categorical random variables) is observed to be 0.862. Thus, quite a high correlation between the two eyes is present. This justifies the need for such an analysis of multivariate categorical data – one would loose much information if the analysis were carried out marginally for each response, separately. It is also observed that the posterior standard deviations of the regression coefficients are low for most of the cases. The Monte Carlo

error (MC error) is also quite small, these are less than 0.001 for most of the parameters.

In turn we have tried the same analysis using logistic distributions for ϵ_{Lit} and ϵ_{Rit} (see Qui *et al.* [29], for such logistic error distributions). But the results are almost same as that for normal errors. Hence we are not discussing the figures separately.

6 Computational Issues

In such analyses, computations play a major role. The main objection to the earlier frequentist's solutions (without going to the philosophical issues) was their tremendous computational burden. In fact, in some more complicated scenarios, most of those approaches become computationally intractable. Thus, along with the philosophical view point, we look at the computational burden of the Bayesian approaches with great interest.

Qui *et al.* [29] discussed some such computational issues for multivariate ordinal data in some other situations. Their case was much simpler in the sense that they did not consider complicated random effect distributions like the Dirichlet process semiparametric model. Moreover the prior for ρ in our case yielded a non-standard conditional posterior, requiring Metropolis-Hastings algorithm to be carried out.

All the computations of Biswas and Das [2] and Das and Biswas [9] were carried out using MCMC technique which avoids the evaluation of high dimensional numerical integration. The computations in Angers and Biswas [1] was done using a computer program in S-Plus. It took hours for the computation for model selection. The MCMC requires intensive computation and careful assessment of convergence. For computations, the freely available software WinBUGS is used and it eased the computational burden greatly.

WinBUGS is a window version of the software BUGS (Bayesian analysis Using Gibbs Sampler), developed by MRC, Biostatistics Group, Institute of Public Health. For details see *http://www.mrc-bsu.cam.ac.uk/bugs/* or the manual of WinBUGS 1.3 (see also Spiegelhalter *et al.* [32]). To operate the necessary computations, the software requires the likelihood, prior and data, although we have mentioned the distribution of the conditional posteriors only to visualize what is going on behind the tandom. It generates random samples from a series of conditional posterior distributions specified in the Bayesian model according to an MCMC algorithm.

The analysis of the baseline data in Biswas and Das [2] was initially programmed in C. But it took more than 24 hours for the computation. By WinBUGS the computations were done in less than 4 hours. For the DP

prior with baseline and 4-year follow-up data in Das and Biswas [9], WinBUGS took about 16 hours for the computation. After selecting the model and exclusion of several insignificant covariates, for the present analysis WinBUGS took about 12 hours for each computation.

Standard approaches can be considered for assessing convergence (Cowles and Carlin [6]) and for model checking (Gelman *et al.* [14]) and comparisons (Chib [4], Diciccio *et al.* [11]). In the present paper the convergence of Gibbs sampler is ensured by the basic approach of Gelman and Rubin [15]. Starting from some initial values of the parameters we generate 4 chains of values, each chain being generated starting from an over-dispersed distribution and with a sample size of 8000. We delete 4000 replications as "burning" samples to minimize the effect of initial values and retain the values of the next 4000 replications to approximate the posterior distribution. To monitor the convergence we focus our attention to the parameters of interest, namely DuD, GH, DBP and GH × DBP. Following Gelman and Rubin [15], we compute the between and within chain mean squares of the retained values, say B and W respectively, for each of the parameters. Then we find

$$s^2 = (4000 - 1)W/4000 + B/4000, \quad \nu = s^2 + (4 \times 4000)^{-1}B,$$

and finally the potential scale reduction factor $r = \nu/W$. The potential scale reduction factors are nearly 1, and this suggests that the desired convergence is achieved in the Gibbs sampler.

A 4000 updates were burn in as the initial samples, followed by a further 4000 samples which were used to obtain the posterior summary statistics like the posterior mean, median, standard deviation, MC error, 95% probability interval for each of the parameters under consideration.

7 Concluding Remarks

• It could be of interest to analyze the 10-year follow-up data at the same occasion. But we could not access the data.

• In addition to the probit link, a logit link is also tried, of course with not much difference in the results. This result was somewhat expected, as there is not much difference in the tail behavior of the normal and the logistic distributions. With the help of WinBUGS the computation is easily doable, and it only needs a slight change in the program.

• Some nonparametric technique in terms of smoothing spline ANOVA was employed by Wahba *et al.* [33]. We are now trying to employ a suitable

wavelet approach. This is under consideration and will be pursued in a future communication.

• If we had to deal with multivariate data, the number of correlation components in the correlation matrix would be greater than one, namely $\rho_{12}, \rho_{13}, \cdots$. In this case writing $\rho = (\rho_{12}, \rho_{13}, \cdots)^T$, the prior of the k-dimensional vector ρ (say) can be taken to be proportional to a k-variate normal density (with known mean vector and dispersion matrix) truncated in a set $\mathcal{C}$ which is a convex solid body, subset of $[-1, 1]^k$ that leads to a proper correlation matrix. See Chib and Greenberg [5] for details. One can set the variances large to make it an uniform prior. If, instead, Σ were taken as a proper dispersion matrix (unknown), an inverted Wishart prior could be appropriate. The conditional posterior would be inverted Wishart in that case.

• One relevant point may be the case where some component(s) of the multivariate categorical response is (are) not ordinal. In that case one can possibly use the Rasch model for that component(s). The problem of combining it in our present set up is a challenge. Latent variable approach will not be applicable for those components. The theory is under study with some ordinal components and some (non-ordinal) only categorical components. The details will be pursued in a separate communication. Also one interesting situation could be the case where some of the components of y_i, say $y_{1i}, \cdots, y_{si}$ are ordinal categorical, and the remaining components of y_i, say $y_{s+1,i}, \cdots, y_{ki}$ are continuous. See Das *et al.* [10] for a Bayesian approach in such a situation.

• If, in the multivariate case, some of the cells of the k-way ordered classification (two-way, for bivariate case) are empty, the analysis will become more complicated. In a bivariate case, such a situation is observed and analyzed by Weiss [34] in the frequentist's set up. But Weiss [34] observed that the log-likelihood function is not globally concave resulting serious difficulty in estimation. In the Bayesian set up, the likelihood will be simply as (1) with an indicator multiplied for each individual i, which will be 1 if the bivariate observation vector belongs to the non-empty cells, and will be 0 otherwise. A Bayesian analysis in this direction has been done by Das and Biswas [8] which shows that the Bayesian solution does not have the drawbacks of the frequentist's approach.

• Finally, we would like to provide some motivation of the random effect (in the frequentist's sense) semiparametric model. We are, in fact, observing some covariates. But, there is reason to believe that there are numerous physiological factors such as several nerves are affecting the retinopathy of an individual, most of them are unobservable. To model their effects in the retinopathy level, one has to bring some random component (in the fre-

quentist's sense) under consideration. See Das and Biswas [9] for a detailed discussion in this context.

Acknowledgments

The author wishes to thank the two referees for their careful reading and some suggestions which led to some improvement over an earlier version of the manuscript. The author wishes to thank Drs. Ronal Klein and Barbara E. K. Klein of the University of Wisconsin for providing both the baseline and 4-year follow-up data of the WESDR study. The WESDR project was originally supported in part by grant EY 03083 (R. Klein) from the National Eye Institute, NIH.

References

1. J.-F. Angers and A. Biswas, *Technical Report.* (University of Montreal, CRM 2757, 2001).
2. A. Biswas and K. Das, *Statistics in Medicine* **21**, 549 (2002).
3. M.-H. Chen and Q.-M. Shao, *Journal of Multivariate Analysis* **71**, 277 (1999).
4. S. Chib, *Journal of the American Statistical Association* **90**, 1313 (1995).
5. S. Chib and E. Greenberg, *Biometrika* **85**, 347 (1998).
6. M.K. Cowles and B.P. Carlin, *Journal of the American Statistical Association* **89**, 883 (1994).
7. J.R. Dale, *Biometrics* **42**, 909 (1986).
8. K. Das and A. Biswas, *Technical Report* (Applied Statistics Division, Indian Statistical Institute, 2000).
9. K. Das and A. Biswas, *Technical Report.* (Applied Statistics Division, Indian Statistical Institute, 2001).
10. K. Das *et al.*, *Calcutta Statistical Association Bulletin* **49**, 255 (1999).
11. T.J. Diciccio *et al.*, *Journal of the American Statistical Association* **92**, 903 (1997).
12. T.S. Ferguson, *Ann. Statist.* **1**, 209 (1973).
13. A.E. Gelfand *et al.*, *Journal of the American Statistical Association* **85**, 972 (1990).
14. A. Gelman *et al.*, *Statistica Sinica* **6**, 733 (1996).
15. A. Gelman and D.B. Rubin, *Statistical Science* **7**, 457 (1992).
16. S. Geman and D. Geman, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **6**, 721 (1984).

17. R.F. Hamman, *Proceedings of the Diabetes Control Conference,* Atlanta, Centre of Disease Control, **32**, (1982).
18. W.K. Hastings, *Biometrika* **57**, 97 (1970).
19. K. Kim, *Statistics in Medicine* **14**, 1341 (1995).
20. K. Kim *et al.* in *Collected Papers in Honor of Retirement of Professor Chung Han Yung.* Seoul National University, Department of Computer Science and Statistics Alumni Association, eds. J.Y. Park *et al.*, **36**, (1996).
21. R. Klein *et al.*, *American Journal of Epidemiology* **118**, 228 (1983).
22. R. Klein *et al.*, *Arch. Ophthalmol.* **112**, 1217 (1994).
23. R. Klein *et al.*, *Journal of the American Medical Association* **260**, 2864 (1988).
24. R. Klein *et al.*, *Arch. Ophthalmol.* **107**, 237 (1989).
25. R. Klein *et al.*, *Am. J. Epidemiol.* **119**, 54 (1984).
26. K.P. Kleinman and J.G. Ibrahim, *Biometrics* **54**, 921 (1988).
27. G. Molenberghs and E. Lesaffre, *Journal of the American Statistical Association* **89**, 633 (1994).
28. P. Palmberg *et al.*, *Ophthalmology* **88**, 613 (1981).
29. Z. Qui, *Journal of Biopharmaceutical Statistics* **12**, (2002)
30. L.I. Rand, *Am. J. Med.* **70**, 595 (1981).
31. SAS Institute *SAS/STAT User's Guide, Version 6,* 4th ed. (SAS Institute, Inc., Cary, North Carolina, 1989).
32. D.J. Spiegelhalter *et al.*, *Technical report* (University of Cambridge, MRC Biostatistics Unit, 1994).
33. G. Wahba *et al.*, *Ann. Statist.* **23**, 1865 (1995).
34. A.A. Weiss, *Applied Statistics* **42**, 487 (1993).
35. W.R. Wilks *et al.*, *Biometrics* **49**, 441 (1993).
36. J. M. Williamson and K. Kim, *Statistics in Medicine* **15**, 1507 (1996).
37. J.M. Williamson *et al.*, *Journal of the American Statistical Association* **90**, 1432 (1995).
38. J. M. Williamson *et al.*, *Computer Methods and Programs in Biomedicine* **58**, 25 (1999).

SECOND-ORDER MOMENTS AND MUTUAL INFORMATION IN THE ANALYSIS OF TIME SERIES

DAVID R. BRILLINGER
Statistics Department
University of California
Berkeley, CA, 94720-3860
E-mail: brill@stat.berkeley.edu

A statistical network is a collection of nodes representing random variables and a set of edges that connect the nodes. A probabilistic model for such is called a statistical graphical model. These models, graphs and networks are particularly useful for examining statistical dependencies amongst quantities via conditioning. In this article the nodal random variables are time series. Basic to the study of statistical networks is some measure of the strength of (possibly directed) connections between the nodes. The use of the ordinary and partial coherences and of mutual information is considered as a study for inference concerning statistical graphical models. The focus of this article is simple networks. The article includes an example from hydrology.

1 Introduction

Science concerns relationships. The question that usually arises is what is the form of some relationship. A lesser question is how strong is a relationship. The work presented considers the use of partial coherency, and of coefficients of mutual information as measures of the strength of association of connections.

An example involving river flows measured at a succession of dams along the Mississippi River is presented. Here the nodes are in series and the edges are directed. The locations of the dams are provided in Figure 1.

Basic books discussing statistical graphical models include Cox and Wermuth[7],Whittaker [19], Edwards [8], Lauritzen [16]. The paper has the following sections: Mutual Information, Networks, Results, Discussion and Extensions.

2 Mutual Information

2.1 Continuous Case

The field of information theory provides some concepts of broad use in statistics. One of these is mutual information. It is a generalization of the coefficient of determination, $corr\{X,Y\}^2$, and it unifies a variety of problems.

For a bivariate random variable (X,Y) with density function $p(x,y)$ the

mutual information (MI) is defined as

$$I_{XY} = \int_S p(x,y) \log \frac{p(x,y)}{p_X(x)p_Y(y)} \, dxdy \tag{1}$$

where S is the region $p(x,y) > 0$.

As an example, for the bivariate normal the MI is given by

$$I_{XY} = -\frac{1}{2}\log(1-\rho_{XY}^2)$$

where ρ_{XY} is $corr\{X,Y\}$.

The coefficient I_{XY} has the properties of:

1). Invariance, $I_{XY} = I_{UV}$ if the transformation $(X,Y) \to (U,V)$ has the form $U = f(X), V = g(Y)$ with f and g each differentiable 1-1 transforms.

2). Non negativity, $I_{XY} \geq 0$.

3). Measuring independence in the sense that $I_{XY} = 0$ if and only if X and Y are statistically independent.

4). Providing a measure of the strength of dependence in the senses that i) $I_{XY} = \infty$ if $Y = g(X)$, and ii) $I_{XZ} \leq I_{XY}$ if X is independent of Z given Y.

The property 3) that $I_{XY} = 0$ only if X and Y are independent stands in strong contrast to the much weaker correlation property of ρ_{XY}^2.

The estimation of entropy

There are several methods that have been used.

Nonparametric estimate

Suppose one is considering the bivariate random variable (X,Y). Supposing further that $\hat{p}(x,y)$, is an estimate of the density $p(x,y)$, for example a kernel estimate, then a direct estimate of the entropy is

$$\delta^2 \sum_{i,j} \hat{p}(i\delta, j\delta) \log \hat{p}(i\delta, j\delta) \doteq E\{\log p(X,Y)\} \quad \text{for } \delta \text{ small.}$$

In the same way $E\{\log p_X(X)\}$, $E\{\log p_Y(Y)\}$ may be estimated and one can proceed to an estimate of the mutual information via expression (1). References to the type of entropy estimate just described and some statistical

properties include: Joe [15], Hall and Morton [13], Fernandes [9], Hong and White, [14], Granger *et al.* [12].

Difficulties with this form of estimate can arise when $p_X(.)$, $p_Y(.)$ are small. The nonparametric form also runs into difficulty when one moves to higher dimensions.

A sieve type of estimate is presently being investigated for this situation, in particular an orthogonal function expansion employing shrunken coefficient estimates.

Parametric estimates of entropy

If the density $p(x, y|\theta)$ depends on a parameter θ that may be estimated reasonably then an immediate estimate of the entropy is provided by

$$\int p(x, y|\hat{\theta}) \log p(x, y|\hat{\theta}) \, dxdy.$$

Another form of estimate is based on the likelihood function. Suppose one has a model for the random variable (X, Y) including the parameter θ, (of dimension ν). Suppose the model has the property that X and Y are independent when $\theta = 0$. When there are n independent observations the log-likelihood ratio for the hypothesis $\theta = 0$ is

$$\sum_1^n \log \frac{p(x_i, y_i|\theta)}{p_X(x_i)p_Y(y_i)}$$

with expected value

$$nI_{XY}.$$

This suggests the use of the log-likelihood ratio statistic divided by n as an estimate of I_{XY}. A further aspect of the use of this statistic is that its distribution will be approximately proportional to χ^2_ν, where ν is the dimension of θ, when X and Y are independent.

Partial analysis

When networks are being considered the conditional mutual information is also of use. For a trivariate random variable X, Y, Z one can consider

$$I_{XY|Z} = \int\int\int p(x, y, z) \log \frac{p(x, y, z)p(z)}{p(x, z)p(y, z)} \, dxdydz$$

Its value for the trivariate normal is

$$-\frac{1}{2}\log(1 - \rho^2_{XY|Z})$$

with $\rho_{XY|Z}$ the partial correlation of X and Y having removed the linear effects of Z.

2.2 Processes

A disadvantage of MI as introduced above is that it is simply a scalar. As consideration turns to the process case, *i.e.* functions, it seems pertinent to seeks to decompose its value somehow.

1. *Time-side approach*

   The entropy of a process is defined by a suitable passage to the limit for example as

   $$\lim_{T\to\infty} E\{\log\ p(X_1, X_2, ..., X_T)\}$$

   where $p(x_1, ..., x_T)$ denotes the density of order T. To begin one can simply consider the mutual information of the values $Y(t+u), Y(t)$ or of the values $Y(t+u), X(t)$. This leads to a consideration of the coefficients

   $$I_{YY}(u) \quad \text{and} \quad I_{YX}(u)$$

   i.e. mutual information as a function of lag u. References to this idea include: Li [17] and Granger and Lin [11].

2. *Frequency-side approach*

   Similarly it seems worth considering the mutual information at frequency λ of two components of a bivariate stationary time series. This could be defined as the mutual information of the spectral increments $dZ_X(\lambda)$ and $dZ_Y(\lambda)$ of the Cramér representation. Because these variates are complex-valued a 4-variate random variable is involved. In the Gaussian case the MI at frequency λ is

   $$-\ \log(1 - |R_{XY}(\lambda)|^2)$$

   where $|R_{XY}(\lambda)|^2$ is the coherence of X and Y at frequency λ. The overall information rate is

   $$-\int_{-\pi}^{\pi} \log(1 - |R(\omega)|^2)d\omega,$$

   (see Granger and Hatanaka [10]).

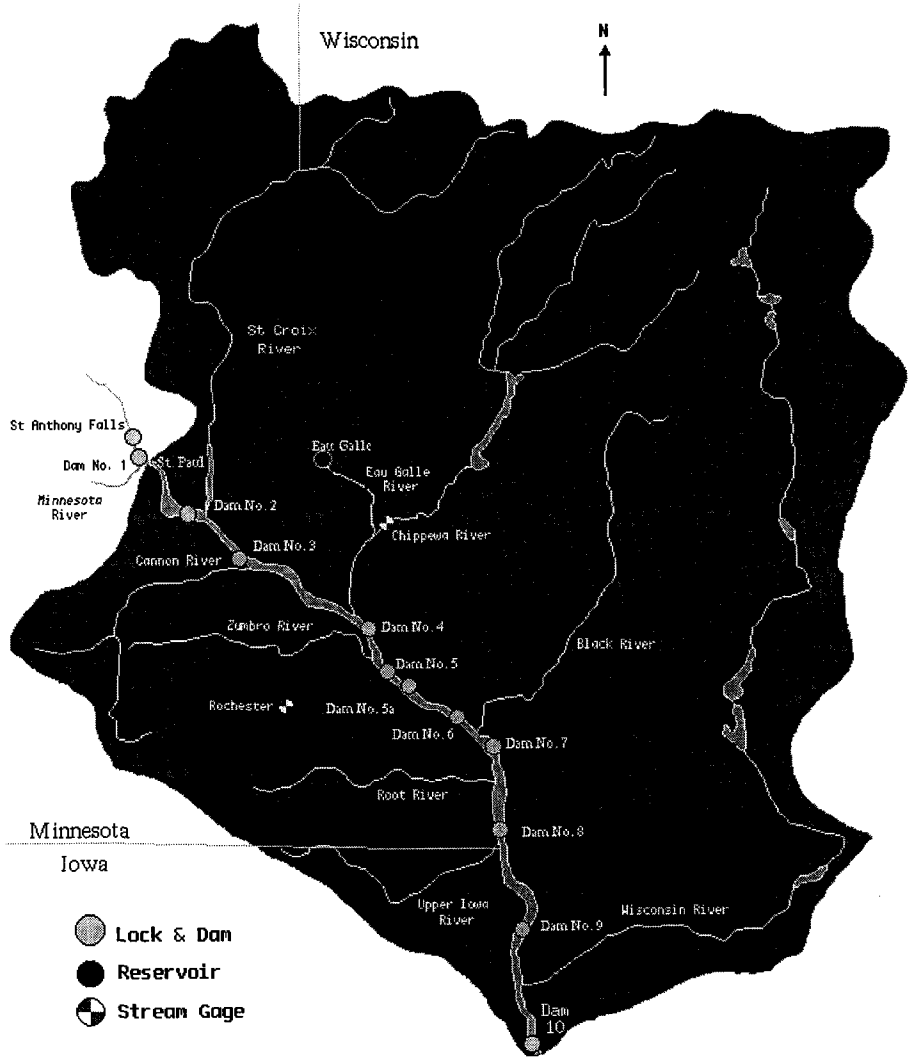


Figure 1. The locations of the 10 dams along the Mississippi River some of whose flow rates are studied.

In the general case for each frequency one might construct an estimate, $\hat{I}_{XY}(\lambda)$, based on kernel estimates of the densities taking empirical FT-values near λ as the data.

One way to estimate the MI, suggested above, is to fit a parametric model and then to use the log-likelihood ratio test statistic for a test of independence.

A novel way, also being pursued, is to use first recurrence time estimates of entropy (see Ornstein and Weiss [18] and Wyner [20]).

3 Networks

In crude terms a *network* is a box (or node) and line (or edge) diagram and some of the lines may be directed. In our work a box corresponds to a random entity, to a random variable, to a time series or to a point process. In studying such models the methods of statistical graphical models provide pertinent methodology. Typically these models are based on conditional distributions. See the books by Edwards [8], Whittaker [19], Lauritzen [16], Cox and Wermuth [7].

If A, B, C represent nodes a question may be as simple as: Is the structure $A \rightarrow B \rightarrow C$ appropriate or is it better described as $(A, B) \rightarrow C$? On the other hand the question may be as complicated as: What is the wiring diagram of the brain?

Figure 1 shows the locations of dams of a network along the Mississippi River. Since the bulk of the water flows south an elementary graphical model for this situation is: *Dam* 1 $\rightarrow$ *Dam* 2 $\rightarrow$ ... $\rightarrow$ *Dam* 10. Of course there are other sources of water, such as entering rivers and rainfall to be taken note of. The figure and the data to be analyzed are taken from

www.mvp-wc.usace.army.mil/projects/lock_dam.html

One reference to the approach of this paper is Brillinger [3].

4 Results

4.1 Mississippi River Flow

The waters of the Mississippi River flow from Minnesota in the north of the United States to the Gulf of Mexico. Flooding along the river has long been a concern, and the U.S. Army Corps of Engineers has constructed a series of locks for flood control and as an aid to navigation. The waters flowing may be viewed as a system added to by precipitation and by flow from entering streams and runoff and reduced by evaporation, absorption and diversion.

The basic data employed in the study to be described are the daily water

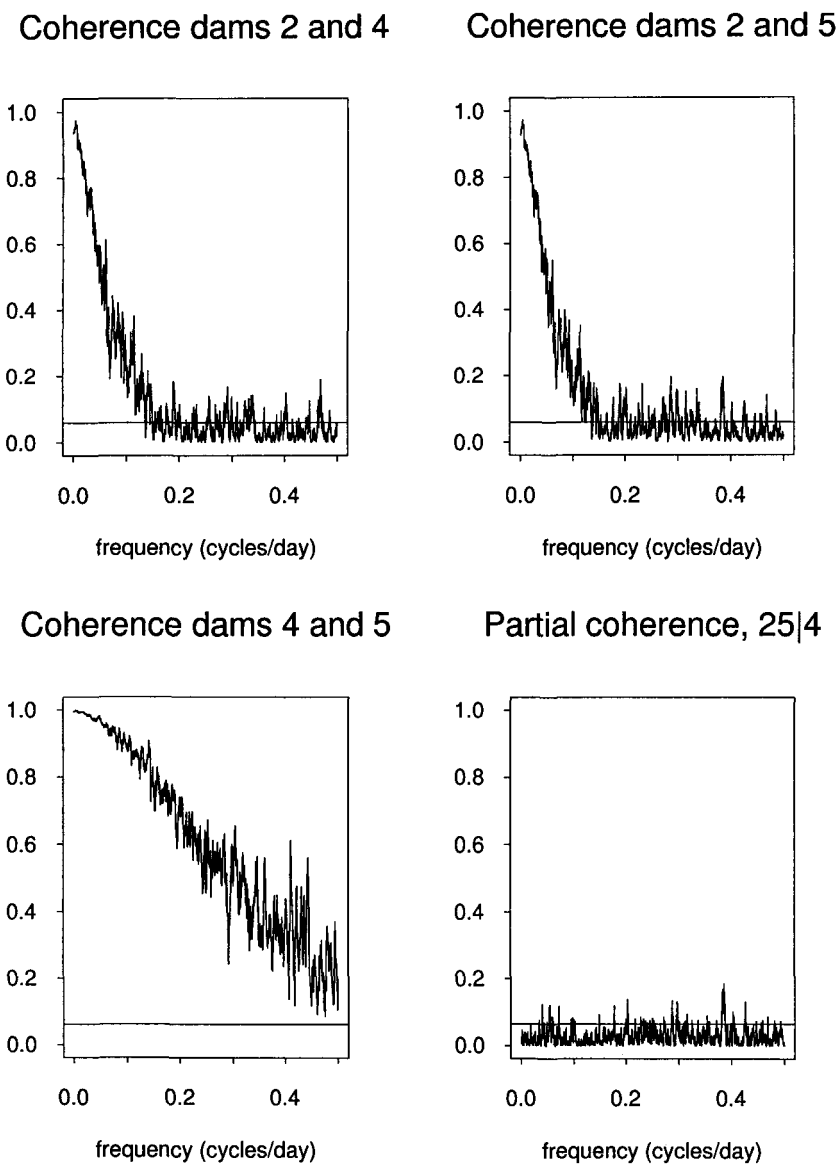


Figure 2. Estimated partial coherence of log(flow rate) at Dams 2 and 5 given those at Dam 4. The horizontal line gives the approximate upper 95% null level.

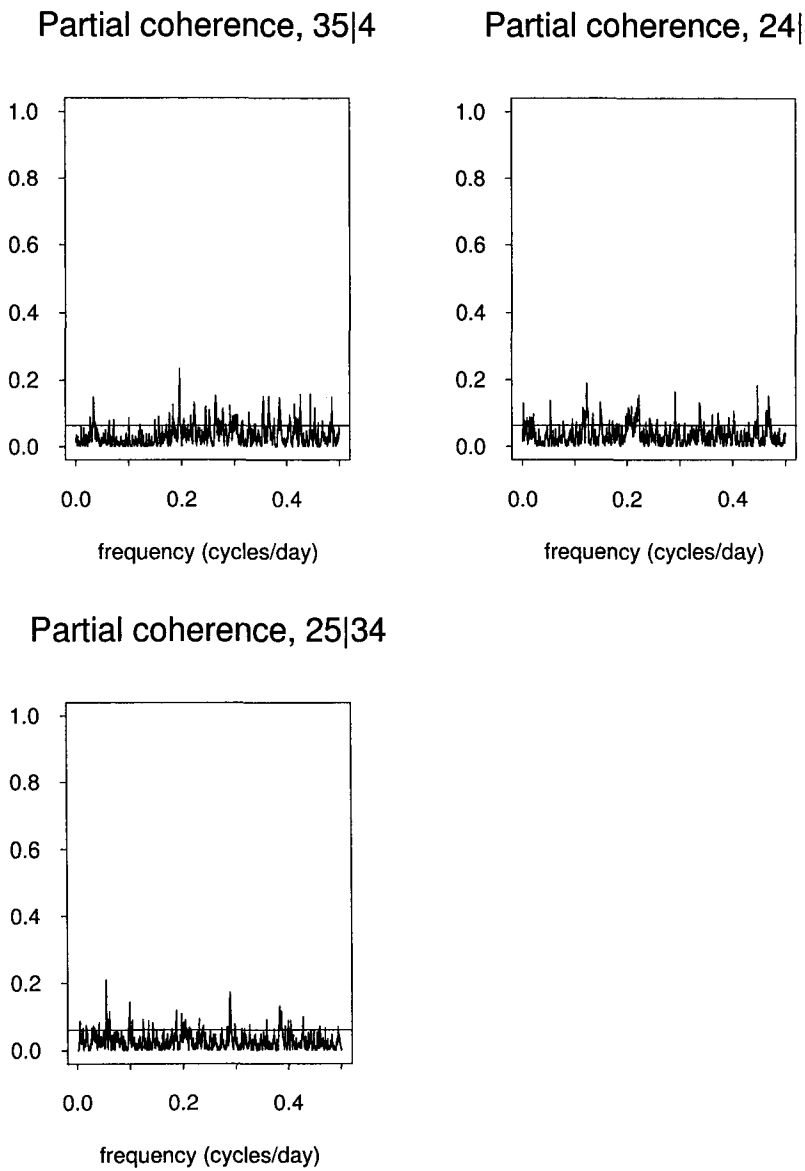


Figure 3. The estimated partial coherence of log(flow rate) at Dams 3 and 5 given Dam 4, Dams 2 and 4 given Dam 3 and Dams 2 and 5 given those at Dams 3 and 4. The horizontal lines give the approximate upper 95% null level.

flow rates as recorded at a succession of dams along the river. The data are daily from 1960 on and were obtained from the WWW site mentioned above. Figure 1, taken from that web site, shows the locations of the dams. Entering streams may be seen. Consider for example the logarithms of the flow rates, $Y_2(t), Y_4(t), Y_5(t)$ at Dams 2, 4, 5. Their locations may be seen in Figure 1. One sees the bulk of the waters passing from Dam 2 to Dam 4 and then onto Dam 5. One sees the Zumbo River entering between Dams 4 and 5 and the three other rivers entering between Dam 2 and Dam 4. The logarithms of the flow rates are taken as the basic variables.

This situation will be studied as providing a useful test bed for studying the effectiveness of the partial coherence and mutual information parameters.

Figure 2 presents the results of a partial coherence analysis focusing on Dams 2, 4, 5. The top right panel provides the estimated coherence functions of Dams 2 and 5. The horizontal line gives the approximate upper 95% null level under the null hypothesis of zero coherence. One sees high coherence at the low frequencies. The bottom right plot is the estimated partial coherence function of Dams 2 and 5 having removed the linear time invariant effects of Dam 4. Once again the horizontal line gives the approximate upper 95% null level under the null hypothesis of zero coherence. One looks for more than about 5% of the values lying above these lines. In the partial coherence case one sees not too much activity. In this situation the partial coherence could have been anticipated to be negligible because of Dam 4's being so close to Dam 5, *i.e.* highly effective in blocking off the direct effects of Dam 2. As an aid to understanding this analysis one can consider the model

$$Y_5(t) = \int a(t-u)Y_4(u)du + \epsilon(t)$$

$$Y_4(t) = \int b(t-u)Y_2(u)du + \eta(t)$$

with ϵ and η noise processes. The partial coherence estimated is then the coherence of the processes ϵ and η.

The analysis is now extended to 4 series. Figure 3 provides the estimated partial coherences of Dams 3 and 5 having removed the effects of Dam 4 and of Dams 2 and 4 having removed the effects of Dam 3 and also that of Dams 2 and 5 having removed the effects of Dams 3 and 4. The required formulas may be found in Brillinger [2]. The networks contemplated in the analysis are the parallel and series ones. Logically the series structure is appropriate.

All three of the partial coherences appear weak. This is consistent with the series structure of the graph as was anticipated. Had there been parallel links

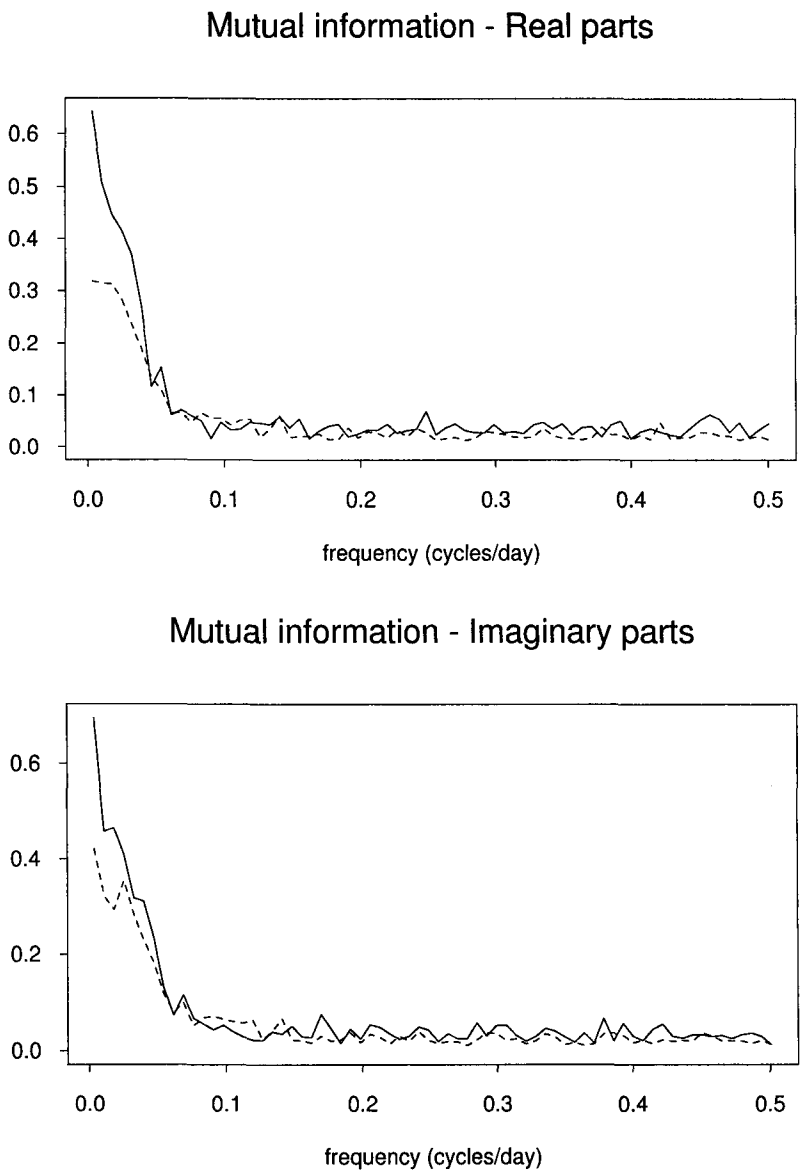


Figure 4. The top panel provides an estimate of the MI of the two real parts of the Mississippi river flows at Dams 2 and 5. The bottom panel similarly refers to the two imaginary parts. The dashed line gives the approximate upper 95% null level.

through Dams 3 and 4 instead of the serial ones then the partial coherences 35|4 and 24|3 would not be expected to be near 0 generally.

As a further example the MI of series 2 and 5 was estimated as a function of frequency λ, actually the MI's of the two real parts of the frequency components and of the two imaginary parts. These estimates are graphed in Figure 4. Approximate 2.326 s.e. limits are obtained by randomly altering the phases of the empirical FT components, following the procedure of Braun and Kulperger [1]. In the plots one notes some apparent association at the lower frequencies. This could arise, for example, from the occurrence of snow or rain storms affecting the segments of the river at the same time.

An estimate of the full MI is currently being developed following the discussion of Section 2 .

A point process analysis of this situation is developed in Brillinger [4]. These data are also considered in Brillinger [5].

5 Discussion and Extensions

The estimated partial coherences of the dams' flow rates had the forms anticipated given the physical knowledge of the situation. The coefficient of mutual information is a unifying concept extending second-order quantities that have restricted applicability. Its being 0 actually implies that the quantities involved are statistically independent. Another important advantage is that the MI pays no real attention to the values of the process. They can be non-negative, integers or proportions for example.

The MI is useful when one wishes to make inferences stronger than: "The hypothesis of independence is rejected" and more of the character "The strength of connection is I."

During the work the plot of the *function* $\hat{I}_{XY}(\lambda)$, appear more useful than simple scalars $\hat{I}_{XY}$. Both parametric model-based estimates and nonparametric estimates of mutual information have been mentioned and computed.

A number of extensions are available and some work is in progress. One can consider the cases of: point processes, spatial-temporal data, local estimates, of learning, of change, of trend, and of comparative experiments. Brillinger [6] contains related ideas and examples from neurophysiology.

One needs to develop the statistical properties of other estimates of MI such as the estimate based on the waiting time and the sieve estimates.

Acknowledgments

Dr. Partha Mitra made some stimulating remarks concerning the use of mutual information. The Referees' remarks led to clarifications and reductions. I thank them.

Part of the material was presented as the Opening Lecture at the XXXIème Journées de Statistique in Grenoble in 1999.

This research was supported by the NSF grants DMS-9704739 and DMS-9971309.

References

1. J. Braun and R. Kulperger, *Commun. Statist.-Theory Meth.* **26**, 1329 (1997).
2. D. R. Brillinger, (1975). *Time Series: Data Analysis and Theory* (Holt-Rinehart, New York. Republished as a SIAM Classic in Applied Mathematics 2001).
3. D. R. Brillinger, *Brazilian Review Econometrics* **16**, 1 (1996).
4. D. R. Brillinger, *Technometrics* **43**, 266 (2001).
5. D. R. Brillinger, *Proc. Interface 2001.*
6. D. R. Brillinger, *Revista Investigacion Operacional* (to appear, 2002).
7. D. R. Cox and N. Wermuth, *Multivariate Dependencies: Models, Analysis, and Interpretation* (Chapman & Hall, London, 1998).
8. D. Edwards, *Introduction to Graphical Modelling* (Springer, New York, 1995).
9. M. Fernandes, *Nonparametric entropy-based tests of independence between stochastic processes* (Preprint, Fundacão Getulio Vargas, Rio de Janeiro, 2000).
10. C. W. J. Granger and M. Hatanaka, *Spectral Analysis of Economic Time Series* (Princeton University Press, Princeton, 1964).
11. C. W. J. Granger and J-L. Lin, *J. Time Series Anal.* **15**, 371 (1994).
12. C. W. J. Granger, E. Maasouni and J. Racine, *A dependence metric for nonlinear time series* (Preprint, Dept. of Economics, UCSD, 2000).
13. P. Hall and S. C. Morton, *Ann. Inst. Statist. Math.* **45**, 69 (1993).
14. Y. Hong and H. White, *Asymptotic distribution theory for nonparametric entropy measures of serial dependence* (Preprint, Economics Department, UCSD, 2000).
15. H. Joe, *Ann. Inst. Statist. Math.* **41**, 683 (1989).
16. S. L. Lauritzen, *Graphical Models* (Oxford University Press, Oxford, 1996).

17. W. Li, *J. Statistical Physics* **60**, 823 (1990).
18. D. S. Ornstein and B. Weiss, *IEEE Inf. Theory* **39**, 78 (1993).
19. J. Whittaker, *Graphical Models in Applied Multivariate Statistics* (Wiley, New York, 1990).
20. A. J. Wyner, *Ann. App. Prob.* **9**, 780 (1999).

SPATIAL ASSOCIATION BETWEEN COMMUNITY AIR POLLUTION AND HEART DISEASE: ANALYSIS OF CORRELATED DATA

SABIT CAKMAK, RICK BURNETT
Health Canada, PL 3104E, Ottawa, On., K1A 1B6, Canada
E-mail: sabit-cakmak@hc-sc.gc.ca

MICHAEL JERRETT
School of Geography and Geology, McMaster University, Hamilton, Canada

MARK S. GOLDBERG
Departments of Epidemiology, Biostatistics, and Occupational Health, McGill University, Canada

ARDEN POPE III
Economics Department, Brigham Young University, Provo, Utah, USA

RENJUN MA
Department of Mathematics and Statistics, University of New Brunswick, Canada

DANIEL KREWSKI
Institute of Population Health, University of Ottawa, Canada

Cohort study designs are often used to assess the association between community based ambient air pollution concentrations and health outcomes, such as mortality, development and prevalence of disease, and pulmonary function. Typically, a large number of subjects are enrolled in the study in each of a small number of communities. Fixed site monitors are used to determine long-term exposure to ambient pollution. The association between community average pollution levels and health is determined after controlling for risk factors of the health outcome measured at the individual level (i.e., smoking). We present a new spatial regression model linking spatial variation in ambient air pollution to health. Health outcomes can be measured as continuous variables (pulmonary function), binary (prevalence of disease), or time to event data (survival or development of disease). The model incorporates risk factors measured at the individual level, such as smoking, and at the community level, such as air pollution. We demonstrate that the spatial autocorrelation in community health outcomes, an indication of not fully characterizing potentially confounding risk factors to the air pollution-health association, can be accounted for through the inclusion of location in the deterministic component of the model assessing the effects of air pollution on health. We present a statistical approach that can be implemented for very large cohort studies. Our methods are illustrated with an analysis of the American Cancer Society cohort to determine whether the prevalence of heart disease is associated with concentrations of sulfate particles.

1 Introduction

Cohort study designs are often used to assess the association between community based ambient air pollution concentrations and health outcomes, such as mortality, development and prevalence of disease, and pulmonary function. Typically, a large number of subjects are enrolled in the study in each of a small number of communities. Practical considerations of implementation motivate this epidemiologic design, in that it is relatively easy to recruit additional subjects in a community compared to identifying communities with appropriate longitudinal pollution monitoring records. Fixed site monitors are used to determine long-term exposure to ambient pollution. The association between community average pollution levels and health is determined after controlling for risk factors of the health outcome measured at the individual level (e.g., smoking). Standard statistical computing software programs (e.g., SAS [10]) can be used for analysis if the assumption of statistical independence between subjects is appropriate.

Health responses, however, often cluster by community, indicating that responses of subjects within the same community are more similar than responses of subjects in different communities. This implies that community itself poses some risk to health. Community-level variables, such as measures of socio-economic status of the community, can be used to model this unexplained risk in addition to individual-level risk factors. Failure to account for all the variation between community health outcomes even after controlling for individual and community level risk factors can lead to downward biased estimates of the uncertainty in the community-level risk factors, including air pollution (see Ware and Strom [13]). Additional bias in the uncertainty of the risk estimates can occur if the community average health outcomes display spatial autocorrelation. That is, health responses for communities close together are more similar than responses for communities farther apart, thus invalidating the use of standard statistical models such as linear, logistic or hazard models. Autocorrelation in the residuals of these models could be due to missing or systematically mis-measured risk factors that are spatially autocorrelated. Failure to account for spatial autocorrelation can yield downward biased estimates of uncertainty in the community-level risk factors and may suggest uncomplete control for potentially confounding community-level factors with the variables of primary interest, such as air pollution (Miron [8]).

We present a regression model in which the residual community health responses are characterized by community-based stochastic variables called "random effects", after controlling for individual and community-level risk factors. The variance of the random effects represents the residual variation

in response between communities. Broader spatial trends in residual health outcomes are modeled by non-parametric smoothers of location in the deterministic component of the model. This approach is analogous to that used in time series studies in which temporal trends in health series are jointly modeled with air pollution (Cakmak *et al.* [2]).

Our modeling framework can accommodate continuous, binary, and time to event health data. We present an approach to statistical inference that can accommodate very large datasets typically encountered in this area. Our methods are illustrated using data obtained from the American Cancer Society (ACS) study of particulate air pollution and health (Pope III *et al.*[9]). We examined the association between the prevalence of heart disease and ambient particulate sulfate concentrations in 144 metropolitan areas in the United States.

2 Spatial Model

Three basic regression models are considred here: linear, logistic, and time to event. First, we start with time to event models.

Suppose that the cohort of interest is stratified on the basis of one or more relavant covariates. Let the instanteneous probability of event at time t, or hazard fuction, for a individual i residing in area s and a member of stratum m is given by $h_i^m(t)$. We propose a space-time model to relate spatial risk factors to longevity. The hazard ${h_i}^m(t)$ for our model is determined by

$$h_0^m(t)e^{\{\zeta(s)+\rho^T X_i^m(t)+\beta^T Z(s)+\eta(s)\}} \tag{1}$$

where $h_0^m(t)$ is the baseline hazard function for the mth strata, $\zeta(s)$ is the two-dimensional term to account for residual spatial variability, ρ is a vector of unknown regression coefficients linking individual risk factors to the hazard function, and β is a vector of unknown regression coefficients linking the spatial level risk factors to the hazard function. Covariate information modulates the baseline hazard function with the regression parameters ρ and β representing the logarithm of the relative risk of death per unit change in the individual and spatial covariates, respectively. The spatial random effects, $\eta(s)$, or frailties, are shared by all individuals in area s. These random effects reflect the difference between the observed hazard function and the hazard function predicted from a statistical model. We assume that the spatial process has zero expectation, variance , $\Theta > 0$ and correlation matrix Ω with dimension S.

Second, we consider logistic model. Let us assume that the response probability is given by π. The odds of positive response $(\frac{\pi}{1-\pi})$ for a individual

residing in area s and a member of stratum m is defined by

$$e^{\{\zeta(s)+\rho^T X_i^m(t)+\beta^T Z(s)+\eta(s)\}} \tag{2}$$

where $\zeta(s)$ is the two-dimensional term to account for residual spatial variability, ρ is a vector of unknown regression coefficients linking individual risk factors to the odds of positive response, β is a vector of unknown regression coefficients linking the spatial level risk factors to odds of positive response, and the spatial random effects process, $\eta(s)$, are shared by all individuals in area s.

Third, we consider linear model. The model assumes that dependence of health responses on covariates occours through linear combination:

$$\zeta(s) + \rho^T X_i^m(t) + \beta^T Z(s) + \eta(s) \tag{3}$$

where $\zeta(s)$ is the two-dimensional term to account for residual spatial variability, ρ is a vector of unknown regression coefficients linking individual risk factors to the health responses, β is a vector of unknown regression coefficients linking the spatial level risk factors to the health responses, and the spatial random effects process, $\eta(s)$, are shared by all individuals in area s.

3 Statistical Estimation

We divide the estimation procedure into two stage. In "Stage One", we work in time domain. Health outcome data is modeled by covariates at the individual level and indicator functions for each community. Community-level covariates, such as air pollution, are not included at this stage. For our hazard model (1), for example, in the time domain we consider

$$h_0^m e^{\{\Sigma_{s=1}^{S-1}\delta(s)I_{(s)}+\rho^T x_i^m(t)\}} \tag{4}$$

where $I(s), s = 1, ..., S-1$ are indicator variables taking the value 1 if the subject resides in area s and zero otherwise. One area (S) is (arbitrarily) assigned as a reference. The unknown parameters represent the logarithm of the relative risk of death for those subjects living in area s compared to those subjects in the reference area S, after controlling for the individual risk factors $x_i{}^{(m)}(t)$. In the time domain the corresponding equations for linear and logistic regression models are straightforward.

Estimates of the community-specific health responses are determined using standard computer software for linear, logistic and Cox proportional hazard survival models. We used SAS [10] procedures because they can accommodate very large sample sizes. For linear and logistic regression models, we do not specify an intercept term. The estimates of the indicator functions can be

interpreted as the average response in a specific community after adjusting for individual-level covariates. Output from stage one is the community-specific adjusted health responses denoted by $\{\hat{\delta}(s), s = 1, ..., S\}$, where s denotes a zero dimensional point in Cartesian (x,y) space representing the location of one of communities under study. Additional output from this stage is the variance-covariance matrix of the $\hat{\delta}(s)$, denoted by $\hat{v}$, which describes the uncertainty in the estimates of the community-specific adjusted health response.

The Cox proportional hazards survival model incorporates a baseline hazard function that is interpreted as the instantaneous probability of an event given all covariates in the model are assigned a zero value. In this case, indicator functions for community are defined with respect to a reference community. A limitation of this procedure is that the uncertainty of the estimate of the reference area is not defined. Because these values are based on comparisons with the same reference area, they are correlated. This correlation increases the estimated uncertainty in the location-specific log-relative risks $\{\hat{\delta}(s)\}$. The induced correlation can be removed by methods developed by Easton *et al.* [4]. This procedure eliminates the covariance between the $\{\hat{\delta}(s)\}$, and defines an associated estimate of uncertainty to the assigned value of zero for $\hat{\delta}(S)$.

In "Stage Two", we work in space domain. Estimates of community health responses are related to risk factors defined at the community level using a linear random effects regression model as follows:

$$\hat{\delta}(s) = \zeta(s) + \beta^T Z(s) + \eta(s) + \epsilon(s) \tag{5}$$

where $\epsilon(s)$is a random process with zero expectation, variance-covariance matrix $\hat{v}$, independent of the spatial random effects process $\eta(s)$.

$\zeta(s)$is the two-dimensional trend term to account for residual spatial variability, and β is a vector of unknown regression coefficients linking the vector of spatial level risk factors, Z(s), to the community-specific health responses.

The spatial random effects, $\eta(s)$, are shared by all individuals in area s. These random effects reflect the difference between the observed and predicted values from our statistical model. We assume that the spatial process $\eta(s)$ is stationary (i.e., expectation does not vary in space), has zero expectation, variance $\Theta > 0$, and correlation matrix Ω with dimension S. The correlation of the random effects between two areas can be modeled by their distance apart or some other characteristic of their locations. Autocorrelation models, or correlation between the same response at different locations, typically assume that closer locations will have attribute values, in this case random effects, that are more similar than values in locations farther apart. Thus, these models are often characterized by functions that decrease monotonically with

distance (Matheron [7]).

Here, $\hat{\delta}(s)$, has expectation $\mu(s) = \zeta(s) + \beta^T Z(s)$ and variance-covariance matrix $\Sigma = \Theta\Omega + \hat{V}$ If the number of subjects and events in each community is large, as is assumed here, the $\{\hat{\delta}(s)\}$ have approximately a multivariate normal distribution with mean vector μ and variance-covariance matrix Σ. The unknown parameter vector, β , and the variance of the random effects, Θ, can be estimated by maximum likelihood methods with the log-likelihood function minimized by Fishers Scoring algorithm (see Burnett *et al.* [1]). The elements of $\hat{V}$ are assumed to be known in this stage.

The random effects will be non-stationary if there is a trend in space. Spatial trends can be accounted for by surface $\zeta(s)$. We consider non-parametric smoothed estimates of ζ using the robust locally-weighted regression (LOESS) smoothers (Cleveland and Devlin [3]) within the generalized additive model (GAM) framework (Hastie and Tibshirani [5]). The complexity of the surface is controlled by a "span" parameter which is the proportion of the data used for the local regression. The larger the span, the smoother the estimated surface.

We have developed a simple method to judiciously select the appropriate span in the LOESS smoother so as to minimize the autocorrelation structure of the random effects. We do this by plotting the correlation of the standardized residuals versus the distance between communities. The span is decreased until there is no apparent association with distance.

The GAM procedure in S-Plus requires that the data are uncorrelated. Within our iterative estimation procedure, we run the GAM with a weight function as the inverse of diagonal elements of Σ. The GAM regression model consists of a LOESS surface of (x,y) and spatial or community-level covariates, $Z(s)$. We then capture the marginal values of this surface in the x and y dimension. These values are used as two new covariates in the random effects linear regression model, in which estimates of the covariate parameters, β, and the dispersion parameters, Θ , are obtained by maximum likelihood methods. We iterate between the GAM step and the maximum likelihood step until little relative change in consecutive estimates of Θ is observed (in this case $< 10^{-4}$).

4 The American Cancer Society Study

Volunteers of the ACS enrolled over 1.2 million people in September of 1982 throughout the United States. Information on history of disease, longevity, and demographic characteristics were obtained. We obtained information on particulate sulfate levels from the Aerometric Information Retrieval System

(AIRS) and the Inhalable Particle Network (IPN) for 1980 and 1981 for 144 Metropolitan Statistical Areas (MSAs) in which ACS subjects were enrolled. Sulfates are secondarily formed particulate aerosols originating from sulfur dioxide emissions and are a major component of fine particulate matter. The sulfate data from AIRS was collected using glass fiber filters, which react in the presence of sulfur dioxide and artifactually inflate the sulfate concentration. The sulfate data obtained from the IPN used teflon filters which are not subject to this artifact problem. Both monitoring networks were operating in 41 MSAs. We calibrated the AIRS sulfate data to the IPN sulfate data using six linear regression models with separate calibrations for three regions of the county and two time periods [April-September and October to March] [6]. Estimates of exposure were obtained by averaging all available sulfate data from all monitors located in a MSA for the years 1980 and 1981, inclusive. The mean concentration of sulfate particles adjusted for the sulfur dioxide artifact across all 144 cities was 6.4 $\mu g/m^3$, with a minimum value of 1.4 $\mu g/m^3$, an interquartile range of 4.2 $\mu g/m^3$, and a maximum value of 15.6 $\mu g/m^3$.

5 Results

We examined the association between concentrations of sulfate particles and the prevalence of doctor diagnosed heart disease (8.6%). The mean age at enrollment for the 540,679 subjects was 56.6 years, 5% of subjects were younger than 40 years, 5% were older than 75 years, 56.6% of subjects were women, and 94% were white. There were 3,755 subjects per community on average, with a mean of 323 subjects with heart disease per community. Beaumont-Port Arthur, Texas provided the least number of subjects (61) and the fewest with heart disease (7), while the most subjects were recruited in Los Angeles (23,151) with the highest number with heart disease (1,970).

The first step in the analysis was to use the logistic regression procedure in SAS to identify all relevant individual risk factors that were associated with heart disease. This model also included indicator functions for the 144 communities to account for any extra between community variation not account for by the individual-level covariates. Thirty-seven risk factors were selected including variables representing age (a cubic polynomial to account for potential non-linear association with disease), gender, age-gender interaction, race (white versus other), quadratic polynomial of the number of cigarettes smoked and years smoking for current and former smokers, age starting smoking for current and former smokers, consumption of beer, wine, and liquor, quadratic polynomial of body mass index (square of height divided by weight), edu-

cational attainment (less than high school, high school, or more than high school), martial status (single, married, other), passive exposure to tobacco smoke, and regular exposure to some air toxics in work or daily life (asbestos, chemicals/acids/solvents, coal or stone dusts, coal tar/pitch/asphalt, diesel engine exhaust, or formaldehyde).

In the next step, we visualized the spatial association between the prevalence of heart disease and sulfate particles using our stage two linear random effects regression model. Here, we regressed the $\{\hat{\delta}(s)\}$ onto the (x,y) geographic coordinates defined by longitude and latitude of the 144 MSAs with a non-parametric smoothed spatial surface $\hat{\zeta}$ (Figure 1, right panel), excluding spatial covariate information such as air pollution (i.e. $Z(s) \equiv 0$). We used the GAM procedure in S-Plus [12] assuming no autocorrelation (i.e. $\Omega = 0$) and a diagonal form for V. [A diagonal covariance matrix is required to use the GAM in S-Plus [12].] We use latitude and longitude for this visualization step because these coordinate definitions are more easily interpretable than the Cartesian (x,y) coordinate specification. We use the Cartesian coordinates in all other formal statistical analyses as the examination of spatial autocorrelation usually relies on Euclidian rather than angular distance measures. This procedure produced a three-dimensional surface of $\{\hat{\delta}(s)\}$ after adjusting for all individual level risk factors.

We found that adjusted prevalence of heart disease was elevated in the Ohio Valley region south of Lake Erie and in the southeast, diminished in the west and south, and moderately elevated in the mountain states. We also spatially modeled concentrations of sulfate particles using a GAM assuming these values were uncorrelated (i.e. $\Sigma = \sigma^2 I$, where σ^2 is the residual variance and I is the identity matrix). Modeled sulfate values centered by their mean concentration are portrayed in left panel of Figure 1. A corresponding elevation in concentrations of sulfate particles exits in the Ohio Valley region and the southeast, with much lower concentrations in the west. Heart disease was elevated in the mountain states, a pattern not observed for sulfates. This visualization suggests a positive association between the two surfaces.

We then fit our spatial linear random effects model with no spatial predictors, no location surface, assuming that the data are uncorrelated ($\Sigma = \Theta I + V$) and determined the standardized residuals from this model. The association between the autocorrelation of these standardized residuals and distance is graphically presented in Figure 2 (panel a) using the correlogram function in the Spatial Module of S-Plus [11]. Correlograms measure autocorrelation as a function of distance between the observations within binned distance groups. This procedure generates average autocorrelations

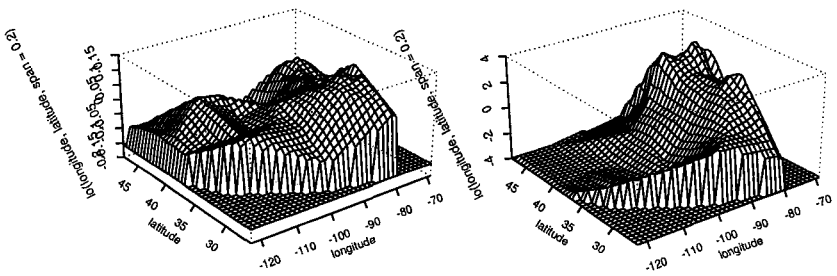


Figure 1. Non-parametric smoothed surface of the prevalence of heart disease by latitude and longitude, adjusted for individual level covariates in American Cancer Society Study with smoothing parameter of 20 percent (left panel). Non-parametric smoothed surface of particulate sulfate concentrations by latitude and longitude with a smoothing parameter of 20 percent (right panel). Note, z-axis represents residuals from generalized additive model.

at 20 roughly equal distances. Autocorrelation peaks at a value of 0.45 for near communities, and is positive but declining for distances under 500kms.. A cyclic pattern is evident in the autocorrelations for distances greater than 500kms.. This pattern could be due to the several regions of elevated prevalence of heart disease (see Figure 1, right panel). The association between autocorrelation and distance is reduced slightly by the inclusion of sulfates (Figure 2, panel b).

We also examined the autocorrelation structure in the standardized residuals for models with a LOESS location surface with spans of 80, 60, 40, and 20% (Figure 2, Panels c-f respectively). A span of 20% was required to eliminate the association between the standardized residuals and distance.

The sensitivity of the sulfate effect, $\hat{\beta}$, and the random effects variance, $\hat{\Theta}$, to the model specification is given in Table 1. The sulfate effect based

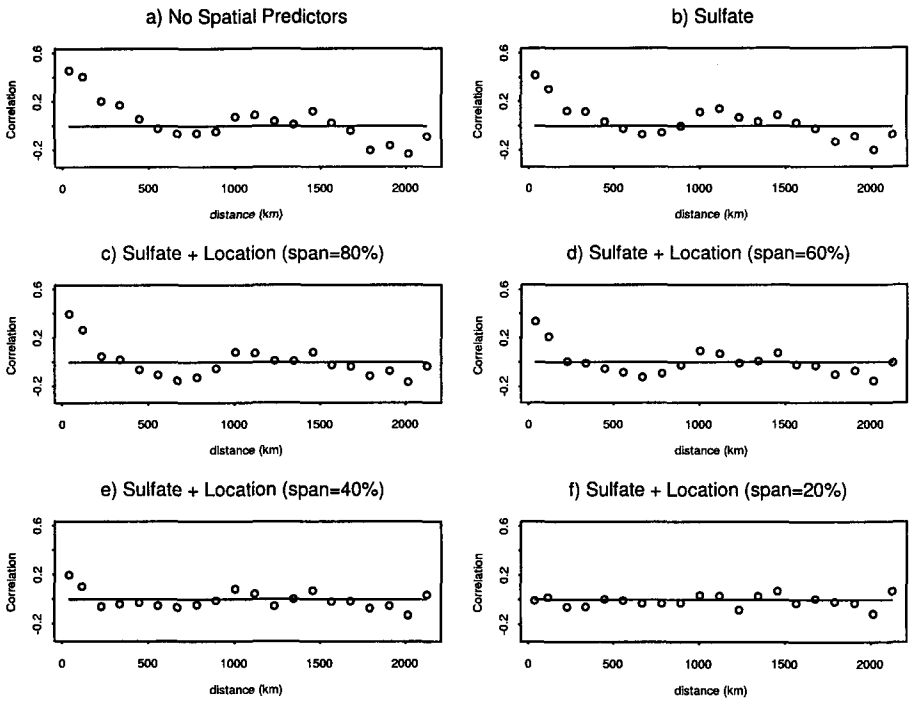


Figure 2. Correlation of standardized residuals by distance between locations for spatial random effects model under six model specifications: a) no spatial predictors and no location surface b) sulfates as the spatial covariate with no location surface, c-f) model includes sulfate and a location surface estimated with a LOESS span of 80-20% respectively. Horizontal line indicates zero values.

on the standard logistic regression model (Model 1) assuming subjects are independent ($\hat{\beta} = 0.0100$) was similar to the corresponding estimate based on our "Two-Stage" spatial random effects model (Model 2) ($\hat{\beta} = 0.0103$) under the identical assumption of independence (i.e. $\Theta \equiv 0$), indicating that the 2-stage estimation approach gave similar results to the logistic regression model for this example. This was most likely due to the large number of subjects and cases of heart disease per MSA. The estimates of the standard errors of the sulfate effect for the two approaches were also similar (Table 1).

There existed strong statistical evidence to support the assumption of

Table 1. Sulfate Effects and random effects variance by model type and span of LOESS smoother of location surface.

Model Type (Model No.)	Span(%)	Sulfate Effect(β) (st. error)	Relative *Risk** (95% Conf. Int.)	Random Effects Variance(Θ))
Logistic (1)	NA	0.0100 (0.0018)	1.043 (1.027, 1.058)	NA
Spatial (2)	NA	0.0103 (0.0019)	1.044 (1.028, 1.060)	0
Spatial (3)	NA	0.0123 (0.0033)	1.053 (1.026, 1.080)	0.0076 0.00041
Spatial (4)	80	0.0132 (0.0036)	1.057 (1.025, 1.089)	0.0062 0.00041
Spatial (5)	60	0.0132 (0.0035)	1.057 (1.026, 1.089)	0.0055 0.00015
Spatial (6)	40	0.0131 (0.0032)	1.057 (1.028, 1.085)	0.0043 0.00012
Spatial (7)	20	0.0110 (0.0031)	1.047 (1.02, 1.07)	0.0022 0.00025

*: Relative risk evaluated at interquartile range of sulfate levels (4.2 $\mu g/m^3$)
NA: not applicable

additional variation in the adjusted prevalence of heart disease between communities (i.e., $\Theta > 0$) based on the likelihood ratio test comparing Models 2 and 3 ($p < 0.0001$). Increasing the complexity of the location surface (i.e., decreasing the span) was also justified in terms of improving the model's fit to the data ($p < 0.0001$) (likelihood ratio tests comparing Model 3 to Models 4 to 7 respectively).

The inclusion of a random effect for location (Model 3) increased the estimate of the sulfate effect ($\hat{\beta} = 0.0123$) but almost doubled the estimated standard error (0.0033) compared to the error obtained from a model assuming independence among subjects (i.e., 0.0019). This suggests that there was more variation in heart disease between communities ($\hat{\Theta} = 0.0076$) than can be fully explained by the within community estimation error,V, and sulfate particles. The sulfate effect estimate was insensitive to inclusion of location surfaces (Models 3-7). The unexplained between community variation, however, was much lower for Model 7 ($\hat{\Theta} = 0.0022$) compared to Model 3 in which no surface

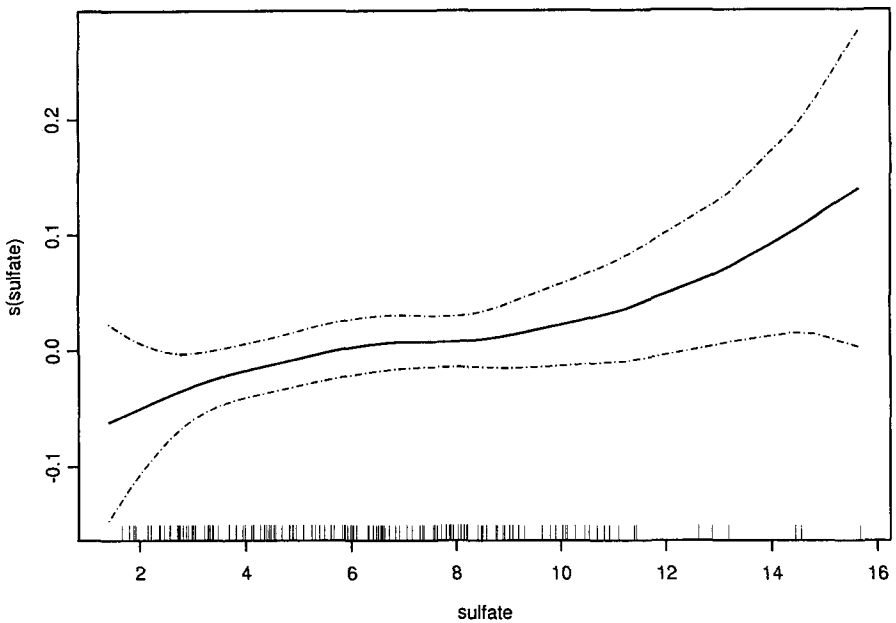


Figure 3. Cubic-spline representation of the association between particulate sulfate concentrations ($\mu g/m^3$) and prevalence of heart disease adjusted for individual level covariates and a smoothed representation of location (LOESS span of 20 percent). Dashed lines represent 95 percent confidence intervals, with tick marks indicating sulfate values.

modeling was undertaken ($\hat{\Theta} = 0.0076$), reflecting the ability of the location surface to explain between community variation not accounted for by sulfate particles.

The form of the relation between sulfate concentrations and the prevalence of heart disease is illustrated in Figure 3. Here, we used a spline function representation of sulfates in the GAM after adjusting for a LOESS surface of heart disease using a span of 20 percent. There was little statistical evidence for a departure from a linear association (p=0.4429).

6 Discussion

We have identified two related issues about the analyses and interpretation of studies linking spatial variation in ambient air pollution and health. First, if the assumption of statistical independence is not valid, the uncertainty in the estimates of effect may be understated. Second, autocorrelation in these residuals may suggest that there exists missing or systematically mis-measured risk factors that may be correlated with air pollution and consequently could confound the pollution-health association.

On the first issue, our model provides more accurate estimates of the uncertainty of estimates of effect. Based on the analysis of the ACS data, while our model gave similar sulfate-heart disease estimates as the standard logistic model, the standard errors of these estimates were higher (Table 1).

With regard to the second issue, we have observed a pattern of spatial autocorrelation in the prevalence of heart disease that cannot be fully explained by ambient particulate sulfate concentrations, even after controlling for a host of risk factors measured at the individual level. Inclusion of a location surface eliminated this spatial correlation pattern and slightly reduced the uncertainty in the sulfate effect estimate from 0.0030 for a LOESS location surface estimated using a 80% span (Model 4) to 0.0029 for a span of 20% (Model 7). However, the sulfate effect was insensitive to adjustment for spatial trends in heart disease suggesting that the association between heart disease and sulfate pollution was strong enough to effectively complete with location in predicting the prevalence of heart disease.

We have extended our modeling approach beyond that reported in our reanalysis [6]. Spatial autocorrelation was modeled by including regional indicator variables into the random effects model. This procedure was ad hoc in that there is no unique method for defining regions. Our new approach of using location surface models allows the data to determine the form and extent of the spatial adjustment. We also removed spatial autocorrelation by pre-filtering both the $\delta(s)$ and the sulfate data for spatial trends using a spatial moving average function. The radius of data inclusion for the moving average term was based on the distance beyond which no evidence of spatial autocorrelation was graphically apparent. Furthermore, the number of communities comprising the moving spatial average varies in space, yielding spatially filtered estimates with varying uncertainty. Finally, all evidence of associations between air pollution and mortality at the regional scale are removed using the pre-filtering approach. Our new modeling method allows air pollution to compete with location to predict health responses. Evidence of regional scale associations between health responses and air pollution will be

captured with this new approach.

Our spatial models can also be used to assist with the selection of appropriate spatial or community-level risk factors. The visualization stage suggested that the upper midwest had elevated prevalence of heart disease and this pattern was not observed in the sulfate data. This would indicate the need for including additional risk factors that cluster in this region. By repeating the visualization and spatial dependence analysis after adding new variables, our method can help to identify variables most likely to explain significance between-community variation.

References

1. R. T. Burnett, W. H. Ross and D. Krewski, *Environmetrics* **6**, 85 (1995).
2. S. Cakmak, R. Burnett and D. Krewski, *J. Expos. Anal. Environ. Epidemiol* **8**, 129 (1998).
3. W. S. Cleveland and S. J. Devlin, *J. Am. Statist. Assoc.* **74**, 829 (1988).
4. D. F. Easton, J. Peto and G. A. G. Babiker, *Statist. in Med.* **10**, 1025 (1991).
5. T. Hastie and R.Tibshirani, *Generalized Additive Models*, (Chapman and Hall, London, 1990).
6. Health Effects Institute, *Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality: A Special Report of the Institute's Particle Epidemiology Reanalysis Project*, (Health Effects Institute, Cambridge MA,2000).
7. G. Matheron, *Eco. Geology* **58**, 1246 (1963).
8. J. Miron, In *Spatial Statistics and Models*, eds. G.L. Gaile and C.J. Willmott (D. Reidel Publishing Company, Boston, 1984).
9. C. A. Pope III, M. J. Thun, M. M. Namboodiri, D. W. Dockery, J. S. Evans, F. E. Speizer and C. W.Heath, *Am. J. Respir. & Crit. Care Med.* **151**, 669 (1995).
10. SAS/STAT Software, *Changes and Enhancements through Release 6.12*, (SAS Institute Inc., Cary, NC, 1997).
11. S+SpatialStats, *User's manual for Windows and UNIX*, (MathSoft, Data Analysis Products Division, Seattle, WA, 1997).
12. S-PLUS, *Programmer's Guide*, (MathSoft, Data Analysis Products Division, Seattle, WA, 2000).
13. J. H. Ware and D. O. Stram, *Can. J. Statist.* **16**, 5 (1988).

ON THE ROBUSTNESS OF RELATIVE SURPRISE INFERENCES TO THE CHOICE OF PRIOR

MICHAEL EVANS
Department of Statistics, University of Toronto, Toronto, Ontario M5S 3G3
E-mail: mevans@utstat.utoronto.ca

TIANLI ZOU
Department of Statistics, University of Toronto, Toronto, Ontario M5S 3G3
E-mail: tianli@utstat.utoronto.ca

Relative surprise inferences are derived in the context of a sampling model for the data and a prior model for the parameter by examining how beliefs change from *a priori* to *a posteriori*. Relative surprise inferences are often different than those that arise from frequency or some standard *a posteriori* Bayesian methods. Given that relative surprise inferences are determined by how the data has caused beliefs to change we would expect that these inferences are more robust to the choice of prior than inferences that are dependent solely on the posterior. This paper is concerned with demonstrating this and quantifying the extent to which robustness is enhanced by taking the relative surprise approach. The connection with Bayes factors is also explored.

1 Introduction

Suppose that we have model $\{f_\theta : \theta \in \Omega\}$ for data $x \in \mathcal{X}$, where each f_θ is a density with respect to support measure μ, and a prior density π with respect to support measure ν for θ. Suppose further that our interest is in making inferences about the marginal parameter $\tau = \Upsilon(\theta)$ where $\Upsilon : \Omega \to \mathcal{T}$. Conditional probability (Bayes Theorem) then prescribes that inferences about τ should be based on the conditional distribution of τ given x; i.e. the posterior distribution of τ. We will denote the prior density of τ with respect to some support measure $\upsilon_{\mathcal{T}}$ on $\mathcal{T}$ by π_Υ and the posterior density of τ with respect to $\upsilon_{\mathcal{T}}$ by $\pi_\Upsilon(\cdot \mid x)$. Throughout this paper all priors are proper.

The actual choice of the densities is important for what follows and this cannot be done arbitrarily; e.g. changing the value of a density on a set of measure zero. For the purposes of this paper we will always take our densities to be limits. So, for example, if P is a probability measure on a space $\mathcal{S}$, absolutely continuous with respect to support measure λ, then the density at s will be taken to be equal to

$$\frac{dP}{d\lambda}(s) = \lim_{n\to\infty} \frac{P(A_n)}{\lambda(A_n)} \tag{1}$$

where $A_n \subseteq S$ is a sequence of sets that converge "nicely" to s. That we can define densities in this way in very general situations follows from definitions found, for example, in Rudin [8] and, as can be seen from Evans, Fraser and Monette [2], this does not restrict our discussion in any meaningful way. Further this restriction on the definition of densities establishes an important connection between the relative surprise approach to inference and the use of Bayes factors as shown at the end of Section 2.

In Section 2 we review some discussion found in Evans [1] concerning the use of surprise and relative surprise for deriving inferences. In addition we establish a relationship between the relative surprise approach to inference and the use of Bayes factors and consider an important example whose robustness properties are subsequently analyzed.

Inferences based on what is called the observed surprise are seen to correspond to some standard Bayesian inferences. These inferences suffer from a lack of invariance; i.e. the inferences depend intrinsically on the parameterization chosen for the model. There may be reasons in a particular application why a parameterization must be fixed; e.g. a loss function is prescribed, but, in general, it seems reasonable to ask that our inferences not be parameterization dependent. Relative surprise inference, as defined in Evans [1], is one example of inference that is independent of the parameterization. This type of inference is also seen to have a very natural interpretation as it is based on how the observed data has changed our beliefs from *a priori* to *a posteriori*.

Given that relative surprise inferences are more data driven that some standard Bayesian inferences one might expect that they are more robust to the selection of the prior. This is the topic of Section 3 and is the main content of the paper. We develop the relevant mathematical theorems for characterizing the local robustness properties of relative surprise and surprise inferences and compare these. Although the arguments are somewhat subtle the results confirm our suspicions that relative surprise inferences exhibit superior robustness properties when considering sensitivity to the choice of prior.

Our approach to studying the robustness properties of our inferences is similar to that taken in Gustafson and Wasserman [6] and Gustafson [4,5]. One key difference, however, is that these authors are concerned with developing robustness diagnostics for applications of Bayesian inference; i.e. determining whether or not a particular model, prior and data combination is robust. While such diagnostics are highly relevant in applications, however, this is not our concern here. We simply want to compare the robustness properties of two different approaches to inference.

In Section 4 we draw some general conclusions and discuss future research on this topic.

2 Surprise and Relative Surprise

While conditional probability prescribes that inferences should be based on the posterior distribution it says nothing about what form these inferences should take. For example, if we wish to quote an estimate of τ then perhaps we might think of using the posterior mean as natural choice. Of course this may not make sense in general as τ need not take its values in a Euclidean space, or the relevant expectation may not exist or the posterior distribution may be strongly skewed or multimodal which could render its mean an inappropriate choice as a representative value. One way to prescribe an inference in such a situation is to specify a loss function and to use a Bayes rule; e.g. posterior expectations are the optimal choice when using squared error loss. Such an approach places a premium on minimizing expected losses over any other considerations.

Another approach is to look for a principle or basic idea that leads necessarily to the form of the inference much as the principle of conditional probability leads to the use of the posterior. If the principle, or basic idea, is compelling then we feel confident that the inferences derived make sense. Of course the assessment of the principle involves examining many applications to see whether or not its application leads to what are generally considered to be sensible inferences.

One principle, or basic idea, is that of *surprise*. In essence we want to measure whether or not a specified value $\tau_0 \in \mathcal{T}$ is surprising in light of the data x. While there are a number of different ways of measuring surprise one fairly natural method for this is given by the observed surprise (OS)

$$\Pi\left(\pi_{\Upsilon}\left(\Upsilon\left(\theta\right) \mid x\right) > \pi_{\Upsilon}\left(\tau_0 \mid x\right) \mid x\right); \tag{2}$$

namely the posterior probability that the posterior density at $\tau = \Upsilon(\theta)$ is greater than at the specified value τ_0. If (2) is near 1 then τ_0 is in a region where the posterior density is relatively small and τ_0 is then said to be surprising in light of the data.

As an example consider the following.

Example 1. *Hypothesis testing.*

Suppose that $\Upsilon = I_{H_0}$ (the indicator function for the set H_0) where $H_0 \subseteq \Omega$ is such that $\Pi(H_0) > 0$ and $\tau_0 = 1$. So in such a case (2) provides an assessment of whether or not the hypothesis H_0 is surprising in light of the data. Taking $\pi_{\Upsilon}(1 \mid x) = \Pi(H_0 \mid x)$ and $\pi_{\Upsilon}(0 \mid x) = \Pi(H_0^c \mid x)$ we have that

the observed surprise equals

$$\Pi\left(\pi_{\Upsilon}\left(\Upsilon\left(\theta\right)\mid x\right) > \pi_{\Upsilon}\left(\tau_0\mid x\right)\mid x\right)$$
$$= \begin{cases} 0 & \text{if } \Pi\left(H_0\mid x\right) \geq \Pi\left(H_0^c\mid x\right) \\ \Pi\left(H_0^c\mid x\right) & \text{if } \Pi\left(H_0\mid x\right) < \Pi\left(H_0^c\mid x\right) \end{cases} \tag{3}$$

and we have evidence against H_0 when $\Pi\left(H_0^c\mid x\right)$ is large. It is easy to see that (3) is a Bayes rule when using 0-1 loss so that the OS approach agrees with the standard Bayesian decision-theoretic answer for this problem.

Sometimes it is argued that posterior probabilities are to be preferred to P-values in assessing a hypothesis but (3) makes it clear that assessing hypotheses using posterior probabilities can also be thought of as falling within the P-value approach. The virtue of a P-value approach as exemplified by (2) is that it does not require $\Pi\left(H_0\right) > 0$ to provide a measure of surprise, although the assessment is different than simply looking at $\Upsilon = I_{H_0}$, so that continuous priors can be used. This has some consequences for the avoidance of Lindley's paradox, as discussed in Evans [1], but this issue does not concern us here.

If we want to estimate τ then Good's principle of least surprise naturally leads us to select the least surprising value of τ_0 and this is a value which minimizes (2). It is obvious then that the least surprise estimator is given by a mode of the posterior density $\pi_{\Upsilon}\left(\cdot\mid x\right)$. Further if we wanted to construct a region containing the true value of τ then perhaps a natural choice is to specify $\gamma \in [0,1]$ and use the γ-surprise region

$$\left\{\tau_0\mid \Pi\left(\pi_{\Upsilon}\left(\Upsilon\left(\theta\right)\mid x\right) > \pi_{\Upsilon}\left(\tau_0\mid x\right)\mid x\right) \leq \gamma\right\};$$

i.e. the set of values for τ_0 that are not surprising at the level γ. It is immediate that such a region corresponds to a highest posterior density (HPD) region for τ.

The above shows that measuring surprise via (2) leads to some standard Bayesian inferences and that the derivation of these does not require the specification of a loss function. It has been pointed out previously in Evans [1], and it may be obvious to the reader, that there is a fundamental difficulty with inferences derived via (2). In particular these inferences will depend upon how we specify the posterior density $\pi_{\Upsilon}\left(\cdot\mid x\right)$, or equivalently, how we specify the support measure $\upsilon_{\mathcal{T}}$. Different choices will lead to very different values for (2), very different estimators and very different γ-surprise regions. In general there does not seem to be an argument that leads us to a canonical choice of support measure. So surprise as a justification for these inferences seems untenable and perhaps we are even lead to doubt the validity of such inferences even though posterior modes and HPD regions are commonly used.

As a solution to this problem Evans [1] proposed to base the derivation of inferences instead on the observed relative surprise (ORS) given by

$$\Pi\left(\frac{\pi_\Upsilon(\Upsilon(\theta)\mid x)}{\pi_\Upsilon(\Upsilon(\theta))} > \frac{\pi_\Upsilon(\tau_0\mid x)}{\pi_\Upsilon(\tau_0)}\,\middle|\, x\right). \tag{4}$$

Here

$$\frac{\pi_\Upsilon(\tau_0\mid x)}{\pi_\Upsilon(\tau_0)} \tag{5}$$

is measuring the change in belief in τ_0 from *a priori* to *a posteriori* and (4) is the posterior probability of a change in belief larger than that observed at τ_0. Again if this number is near 1 we have that τ_0 is a surprising value. In essence a value τ_0 is surprising when the *data* has lead to a bigger increase in belief at other values of $\tau = \Upsilon(\theta)$ compared to the increase (or decrease) at τ_0. A least relative surprise estimate is then obtained by choosing a value of τ_0 that maximizes (4) or equivalently maximizes (5). As above we can also obtain γ-relative surprise regions.

A virtue of all the relative surprise inferences is that they are not dependent on the choice of the prior or posterior densities or equivalently the choice of the support measure $\upsilon_\mathcal{T}$. These inferences are invariant under smooth reparameterizations. As shown in Evans [1] these inferences can also be quite different than standard Bayesian inferences although often they are similar. We consider the context of Example 1.

Example 2. *Hypothesis testing (Example 1 continued).*

We consider again the situation where $\Upsilon = I_{H_0}$, $\Pi(H_0) > 0$ and we take $\pi_\Upsilon(1) = \Pi(H_0)$ and $\pi_\Upsilon(0) = \Pi(H_0^c)$ then the observed relative surprise is given by

$$\Pi\left(\frac{\pi_\Upsilon(\Upsilon(\theta)\mid x)}{\pi_\Upsilon(\Upsilon(\theta))} > \frac{\pi_\Upsilon(\tau_0\mid x)}{\pi_\Upsilon(\tau_0)}\,\middle|\, x\right) = \begin{cases} 0 & \text{if } BF_{H_0} \geq 1 \\ \Pi(H_0^c\mid x) & \text{if } BF_{H_0} < 1 \end{cases} \tag{6}$$

where

$$BF_{H_0} = \frac{\Pi(H_0\mid x)}{1-\Pi(H_0\mid x)} \Big/ \frac{\Pi(H_0)}{1-\Pi(H_0)} \tag{7}$$

is the Bayes factor in favor of H_0. So we get something similar to what is commonly recommended in this context. We will see, however, that the robustness properties of (6) are very different than those of (3). Further (4) has the virtue of providing a measure of surprise even when $\Pi(H_0) = 0$ and again, see Evans [1], this has implications for the avoidance of Lindley's paradox.

Sometimes just the Bayes factor is recommended in hypothesis testing with small values of (7) treated as evidence against H_0. The interpretation of the value a Bayes factors takes is somewhat ambiguous, however, although various calibrations have been suggested; see, for example Kass and Raferty [7]. Perhaps the most direct way of calibrating a Bayes factor is to proceed as we have with the OS and ORS computations and compute the posterior probability of obtaining a value larger than (7) with large values of this probability indicating that H_0 is indeed surprising. For the context of Example 1 we obtain

$$\Pi\left(BF_{\Upsilon(\theta)} > BF_{H_0} \mid x\right) = \begin{cases} 0 & \text{if } BF_{H_0} \geq 1 \\ \Pi\left(H_0^c \mid x\right) & \text{if } BF_{H_0} < 1; \end{cases}$$

i.e. this measure of surprise agrees exactly with the ORS. So in a sense we can think of the ORS as a generalization of the Bayes factor when we choose to calibrate the value of a Bayes factor using the tail probability. Suppose more generally we have that $\mathcal{T} = \{\tau_1, \ldots, \tau_k\}$ with prior probabilities $\pi_1, \ldots, \pi_k$ respectively, and we want to assess the plausibility of the value τ_0. Then

$$BF_\tau = \frac{\Pi\left(\{\tau\} \mid x\right)}{1 - \Pi\left(\{\tau\} \mid x\right)} \Big/ \frac{\Pi\left(\{\tau\}\right)}{1 - \Pi\left(\{\tau\}\right)}$$

and possibly we could assess the surprise at τ_0 by computing

$$\Pi\left(BF_{\Upsilon(\theta)} > BF_{\tau_0} \mid x\right)$$

but when $k \geq 3$ this is generally not equal to the ORS at τ_0. Effectively this approach is saying that the change in belief in τ from a priori to a posteriori should be measured by BF_τ and this seems somewhat less direct than using $\pi_\Upsilon\left(\tau \mid x\right) / \pi_\Upsilon\left(\tau\right)$ as in the ORS. Further it is not at all clear how to extend this approach to the continuous case. Notice, however, that when $\Pi\left(A_n\right) > 0$ for all n, and the A_n converge nicely to $\{\tau\}$ with $\Pi\left(\{\tau\}\right) = 0$, then

$$BF_{A_n} = \frac{\Pi\left(A_n \mid x\right) / \upsilon_\mathcal{T}\left(A_n\right)}{1 - \Pi\left(A_n \mid x\right)} \Big/ \frac{\Pi\left(\{\tau\}\right) / \upsilon_\mathcal{T}\left(A_n\right)}{1 - \Pi\left(A_n\right)} \to \frac{\pi_\Upsilon\left(\tau \mid x\right)}{\pi_\Upsilon\left(\tau\right)}$$

as $n \to \infty$. So in general it is very natural to think of the relative surprise approach as a modification and generalization of the Bayes factor approach. Of course the Bayes factor is restricted to hypothesis testing problems where the parameter space is partitioned into $\Omega = H_0 \cup H_0^C$ with the prior probability of H_0 satisfying $0 < \Pi\left(H_0\right) < 1$. The relative surprise inferences are more general in their application than this, however, both with respect to hypothesis testing, and estimation, prediction and model checking problems; see Evans [1].

3 Robustness to the Prior Under Linear Perturbations

In an obvious sense relative surprise inferences are more data driven than inferences based on surprise. Observe also that when $\Upsilon(\theta) = \theta$ then (4) equals

$$\Pi\left(f_\theta(x) > f_{\theta_0}(x) \mid x\right);$$

i.e. the observed relative surprise is the posterior probability that the likelihood at θ is greater than the likelihood at θ_0. In particular, a least relative surprise estimate of θ is given by a maximum likelihood estimate while relative surprise regions for the full parameter θ correspond to likelihood regions and only depend on the prior through their posterior probability contents. So in a general sense we expect relative surprise inferences to be less dependent on the prior than other Bayesian inferences. It is the purpose of this section to examine this issue more closely and in particular to examine the marginal parameter case; i.e. $\Upsilon(\theta) \neq \theta$.

Our interest then is in how the inferences vary as we perturb the prior Π. Basically we will say one inference method is more robust than another if the rate of change in the inference is smaller at Π for the first inference method than for the second. For example, it is clear immediately that when our goal is to estimate θ then a least relative surprise estimator is more robust to the choice of prior than any other Bayesian inference because it only depends on the likelihood function. Actually this example is somewhat unusual as even the estimation problem becomes more difficult as soon as interest is in a marginal parameter τ. In that case a LRSE will depend on the prior and the extent of the dependence is not obvious. Still the result for the full parameter leaves us hopeful that something similar can be obtained for a marginal parameter as well.

Computational and mathematical issues prevent us from arbitrarily choosing the directions in which we perturb the prior so that definitive comparisons; i.e. results that hold for all directions, seem impossible to achieve at this point. Still we can obtain results for several choices of directions and the results give us considerable insight into the relative robustness of the various inferences. Further, since all the inferences under consideration are based on the observed surprise (2) or observed relative surprise (4), we will compare the robustness of these quantities. Recall that these quantities can be used for assessing whether or not τ_0 is a plausible value for τ.

Perhaps the most commonly used directions are the contamination directions given by prior probability measures of the form

$$\Pi^\epsilon = (1-\epsilon)\,\Pi + \epsilon Q$$

where $\epsilon \geq 0$ and Q is a probability measure on Ω. The prior density of θ with respect to υ is then given by

$$\pi^{\epsilon} = (1-\epsilon)\,\pi + \epsilon q$$

when Q is also absolutely continuous with respect to ν. Actually, because our interest is in the marginal parameter τ, we want to consider only perturbations to the marginal prior distribution so that the conditional distribution of the full parameter given the marginal remains fixed. This kind of perturbation is easily expressed in terms of densities as

$$\begin{aligned}\pi^{\epsilon}(\theta) &= (1-\epsilon)\,\pi_{\Upsilon}(\tau)\,\pi(\theta \mid \tau) + \epsilon q_{\Upsilon}(\tau)\,\pi(\theta \mid \tau) \\ &= [(1-\epsilon)\,\pi_{\Upsilon}(\tau) + \epsilon q_{\Upsilon}(\tau)]\,\pi(\theta \mid \tau) \end{aligned} \tag{8}$$

where $\pi(\theta \mid \tau)$ is the conditional density of θ given $\Upsilon(\theta) = \tau$ with respect to support measure υ_{τ} on $\{\theta : \Upsilon(\theta) = \tau\}$. Then the marginal posterior density, when the prior density is given by (8), is

$$\pi^{\epsilon}_{\Upsilon}(\tau \mid x) = (1-\epsilon_x)\,\pi_{\Upsilon}(\tau \mid x) + \epsilon_x q_{\Upsilon}(\tau \mid x)$$

where

$$\epsilon_x = \frac{\epsilon m_q(x)}{(1-\epsilon)\,m_{\pi}(x) + \epsilon m_q(x)}, \tag{9}$$

$$m_{\pi}(x) = \int_{\Omega} f_{\theta}(x)\,\pi(\theta)\,\upsilon(d\theta)$$

is the prior predictive density for the data when the prior is π with a similar definition for $m_q(x)$,

$$\pi_{\Upsilon}(\tau \mid x) = \frac{\pi_{\Upsilon}(\tau)\,f^{*}_{\tau}(x)}{m_{\pi}(x)},$$

$$q_{\Upsilon}(\tau \mid x) = \frac{q_{\Upsilon}(\tau)\,f^{*}_{\tau}(x)}{m_q(x)}$$

and we write the conditional density with respect to μ of x given that $\Upsilon(\theta) = \tau$ as

$$f^{*}_{\tau}(x) = \int_{\Upsilon^{-1}\{\tau\}} f_{\theta}(x)\,\pi(\theta \mid \tau)\,\upsilon_{\tau}(d\theta).$$

Now letting G denote the posterior distribution function of $\pi_{\Upsilon}(\tau \mid x)$ when $\tau \sim \Pi_{\Upsilon}(\cdot \mid x)$ then the following result, as proved in Evans and Zou [3], yields expressions for the upper and lower Gateaux derivatives of the observed surprise.

Theorem 1. If the posterior distribution of $\pi_\Upsilon(\tau \mid x)$ is continuous when $\tau \sim \Pi_\Upsilon(\cdot \mid x)$ or when $\tau \sim Q_\Upsilon(\cdot \mid x)$, G is differentiable at $\pi_\Upsilon(\tau_0 \mid x)$ with derivative $g(\pi_\Upsilon(\tau_0 \mid x))$ and $q_\Upsilon(\cdot \mid x)$ is continuous and bounded, then the lower Gateaux derivative (and the upper Gateaux derivative) of the observed surprise at τ_0, under linear perturbations of the marginal parameter $\tau = \Upsilon(\theta)$, takes the form

$$\frac{m_q(x)}{m_\pi(x)}\{q_\Upsilon(\tau_* \mid x) - q_\Upsilon(\tau_0 \mid x)\}\, g(\pi_\Upsilon(\tau_0 \mid x)) + \frac{m_q(x)}{m_\pi(x)}\left\{\begin{array}{c} Q_\Upsilon[\pi_\Upsilon(\tau \mid x) > \pi_\Upsilon(\tau_0 \mid x) \mid x] - \\ \Pi_\Upsilon[\pi_\Upsilon(\tau \mid x) > \pi_\Upsilon(\tau_0 \mid x) \mid x] \end{array}\right\} \tag{10}$$

for some τ_*.

Of course when the Gateaux derivative of the observed surprise exists it takes the form given in (10) as well. The expression (10) is not very useful when we need to compute the Gateaux derivative as τ_* is unspecified. In specific examples we are still required to evaluate the limit corresponding to the first part of (10). We will see, however, that (10) is of great value when comparing the robustness of the OS with the robustness of the ORS.

We note that the second term in (10) is always bounded above by

$$\frac{\sup_\tau f_\tau^*(x)}{m_\pi(x)}\{1 - \Pi_\Upsilon[\pi_\Upsilon(\tau \mid x) > \pi_\Upsilon(\tau_0 \mid x) \mid x]\}.$$

The following result, as proved in Evans and Zou [3], suggests that the first term in (10) can be arbitrarily large in absolute value.

Lemma 2. If the density function g of $\pi_\Upsilon(\tau \mid x)$ is continuous from the left at $\pi_\Upsilon(\tau_0 \mid x)$, where τ_0 is the unique mode of $\pi_\Upsilon(\cdot \mid x)$, and $\pi_\Upsilon(\cdot \mid x)$ is continuously differentiable at τ_0 then $g(\pi_\Upsilon(\tau_0 \mid x)) = \infty$.

So Lemma 2 indicates that the first term in (10) could be infinite whenever τ_0 is a posterior mode. That this happens is confirmed by the following example.

Example 3. *Gateaux derivative of the OS can be infinite.*

Suppose that $x \sim \text{Bernoulli}(\theta^{1/2})$ where $\theta \in [0,1]$ is unknown, we take Π to be the Beta$(1/2,1)$ distribution and the support measure ν is Lebesgue measure. Then the posterior density, when we observe $x = 1$, is given by 1; i.e. the posterior distribution is uniform. Suppose we take Q to be a Beta$(3/2,1)$ distribution which leads to the posterior density 2θ when we observe $x = 1$. Now we explicitly evaluate the first term in Gateaux derivative of the observed

surprise at θ_0. First note that $\Pi[\pi(\theta \mid x) > \pi(\theta_0 \mid x) \mid x] = \Pi[1 > 1 \mid x] = 0$ and

$$(1 - \epsilon_x) + 2\epsilon_x\theta > (1 - \epsilon_x) + 2\epsilon_x\theta_0$$

if and only if $\theta > \theta_0$. Therefore

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \left\{ \Pi \left[\begin{array}{c} (1 - \epsilon_x) + 2\epsilon_x\theta > \\ (1 - \epsilon_x) + 2\epsilon_x\theta_0 \end{array} \middle| \, x \right] - \Pi[\pi(\theta \mid x) > \pi(\theta_0 \mid x) \mid x] \right\}$$
$$= \lim_{\epsilon \to 0} \frac{1}{\epsilon} (1 - \theta_0) = \infty$$

and we see that the first term in the Gateaux derivative for the observed surprise is infinite for every θ_0.

In Evans and Zou [3] the following result is established for the Gateaux derivative of the observed relative surprise.

Theorem 3. The Gateaux derivative of the observed relative surprise at τ_0 under linear perturbations of the marginal parameter $\tau = \Upsilon(\theta)$ is given by

$$\frac{m_q(x)}{m_\pi(x)} \left\{ Q_\Upsilon \left[f_\tau^*(x) > f_{\tau_0}^*(x) \middle| \, x \right] - \Pi_\Upsilon \left[f_\tau^*(x) > f_{\tau_0}^*(x) \middle| \, x \right] \right\}. \tag{11}$$

Note that this result implies that the Gateaux derivative of the observed relative surprise is always finite. In particular consider an application of this result to Example 3.

Example 4. *Gateaux derivative of the ORS in Example 3.*

We have that $f_\theta(x) = \theta^{1/2}$ and therefore

$$Q\left[f_\tau^*(x) > f_{\tau_0}^*(x) \middle| \, x \right] = Q_\Upsilon[\theta > \theta_0 \mid x] = \int_{\theta_0}^1 2\theta \, d\theta = 1 - \theta_0^2$$

$$\Pi\left[f_\tau^*(x) > f_{\tau_0}^*(x) \middle| \, x \right] = \Pi[\theta > \theta_0 \mid x] = \int_{\theta_0}^1 d\theta = 1 - \theta_0$$

while $m_\pi(1) = 1/2$ and $m_q(1) = 3/4$. Therefore the Gateaux derivative of the observed relative surprise at θ_0 is given by, using (11),

$$\frac{1}{2}\left(1 - \theta_0^2 - 1 - \theta_0\right) = \frac{1}{2}\theta_0(1 - \theta_0)$$

which is bounded by 1.

We see from Theorem 3 that when τ_0 is a LRSE then the Gateaux derivative of the ORS is 0. Suppose that τ_0 is not a LRSE. If we consider a sequence

of Q measures that converge to a measure degenerate at a LRSE then (11) converges to

$$\frac{\sup_\tau f_\tau^*(x)}{m_\pi(x)} \left\{1 - \Pi_\Upsilon \left[f_\tau^*(x) > f_{\tau_0}^*(x) \mid x\right]\right\} \tag{12}$$

and this is bounded whenever $\sup_\tau f_\tau^*(x) < \infty$. It clear then that (12) is the supremum of (11), when the supremum is taken over Q.

We now establish a result that formally confirms our intuition, at least in the continuous case, that inferences based on the ORS are more robust to the choice of prior than inferences based on the OS. For this we consider the supremum of the absolute values of the Gateaux derivatives over all possible absolutely continuous perturbation measures Q and all possible reparameterizations. By a reparameterization we mean a 1-1 bicontinuously differentiable map defined on $\mathcal{T}$.

Theorem 4. If τ has an absolutely continuous prior distribution on an open subset of a k-dimensional Euclidean space, then the supremum of the absolute value of (10) is greater than or equal to the supremum of the absolute value of (11) when the supremum is taken over all possible absolutely continuous measures Q and all possible reparameterizations.

Proof: When π is absolutely continuous on an open subset of R^k we can reparameterize the problem so that the prior distribution for τ is uniform on $[0,1]^k$. For example, we could use the probability transform based on π to do this.

With this parameterization, if τ_0 is a LRSE then it is also a posterior mode so that (11) is 0 and so is the second term of (10). From Lemma 2, however, we have that the first term of (10) is either 0 or infinite in absolute value. So in the case that τ_0 is a LRSE the supremum, over all possible parameterizations, of the absolute value of (10) is always greater than or equal to the absolute value of (11).

When τ_0 is not a LRSE then reparameterizing so that τ is uniform on $[0,1]^k$ establishes that the second term of (10) equals (11). Further if we take a measure Q that assigns 0 probability to a neighborhood of τ_0 then $q_\Upsilon(\tau_0 \mid x) = 0$ and we have that the first term in (10) is nonnegative. This proves the result.

We have seen from Examples 3 and 4 that the value of (10) can be strictly greater than the value of (11). So there is definitely content to Theorem 4. Taking the supremum over all possible parameterizations seems like the fairest comparison as no parameterization is to be preferred over another unless additional ingredients are introduced into a problem. From the point of

view of robustness one would want to ensure that such a selection did not lead to inferences that had terrible robustness at least when using some traditional Bayesian inferences. Of course this is not an issue when using relative surprise inferences as these are not dependent on the parameterization chosen. We interpret Theorem 4 as strong support for the increased robustness of the ORS over the OS.

Theorem 4 is restricted to contexts where the prior is absolutely continuous. Discrete contexts seem much more difficult to analyze. We consider now, however, an important problem where the marginal prior is discrete.

Example 5. *Hypothesis testing (Examples 1 and 2 continued).*

Consider the problem discussed in Section 1 where we want to test the null hypothesis H_0 and we have $\Pi(H_0) = \pi_0 > 0$. Therefore the posterior probability of H_0 is given by

$$\Pi(H_0 \mid x) = \frac{\pi_0 f^*_{H_0}(x)}{\pi_0 f^*_{H_0}(x) + (1-\pi_0) f^*_{H_0^c}(x)}.$$

Let Q be the measure degenerate at H_0. Under a small perturbation ϵ of the prior probability the posterior probability becomes

$$\Pi^\epsilon(H_0 \mid x) = \frac{((1-\epsilon)\pi_0 + \epsilon) f^*_{H_0}(x)}{((1-\epsilon)\pi_0 + \epsilon) f^*_{H_0}(x) + (1-\epsilon)(1-\pi_0) f^*_{H_0^c}(x)}.$$

Now suppose that $\Pi(H_0 \mid x) > \Pi(H_0^c \mid x)$. Then, by (3), and the fact that $\Pi^\epsilon(H_0 \mid x) > \Pi^\epsilon(H_0^c \mid x)$ for all ϵ small enough, the observed surprise is 0 when ϵ is small enough and so the Gateaux derivative of the observed surprise is also 0.

If $\Pi(H_0 \mid x) = \Pi(H_0^c \mid x)$ then $\pi_0 f^*_{H_0}(x) = (1-\pi_0) f^*_{H_0^c}(x)$ and so

$$\begin{aligned}
&\Pi^\epsilon(H_0 \mid x) \\
&= \frac{((1-\epsilon)\pi_0 + \epsilon) f^*_{H_0}(x)}{((1-\epsilon)\pi_0 + \epsilon) f^*_{H_0}(x) + (1-\epsilon)(1-\pi_0) f^*_{H_0^c}(x)} \\
&= \left\{\frac{(1-\pi_0)}{\pi_0}\right\}\left\{\frac{(1-\epsilon)\pi_0 + \epsilon}{(1-\epsilon)(1-\pi_0)}\right\} \times \\
&\qquad \left\{\frac{(1-\epsilon)(1-\pi_0) f^*_{H_0^c}(x)}{((1-\epsilon)\pi_0 + \epsilon) f^*_{H_0}(x) + (1-\epsilon)(1-\pi_0) f^*_{H_0^c}(x)}\right\} \\
&= \frac{1}{\pi_0}\left(\pi_0 + \frac{\epsilon}{1-\epsilon}\right)\left\{\frac{(1-\epsilon)(1-\pi_0) f^*_{H_0^c}(x)}{((1-\epsilon)\pi_0 + \epsilon) f^*_{H_0}(x) + (1-\epsilon)(1-\pi_0) f^*_{H_0^c}(x)}\right\}
\end{aligned}$$

$$> \frac{(1-\epsilon)(1-\pi_0) f^*_{H_0^c}(x)}{((1-\epsilon)\pi_0+\epsilon) f^*_{H_0}(x) + (1-\epsilon)(1-\pi_0) f^*_{H_0^c}(x)}$$
$$= \Pi^\epsilon (H_0^c \mid x)$$

and the Gateaux derivative is again 0 by (3).

If $\Pi(H_0 \mid x) < \Pi(H_0^c \mid x)$ then, since $\Pi^\epsilon(H_0 \mid x) < \Pi^\epsilon(H_0^c \mid x)$ for all ϵ small enough, the Gateaux derivative of the observed surprise is given by

$$-\lim_{\epsilon \to 0} \frac{1}{\epsilon} \left\{ \frac{((1-\epsilon)\pi_0+\epsilon) f^*_{H_0}(x)}{((1-\epsilon)\pi_0+\epsilon) f^*_{H_0}(x) + (1-\epsilon)(1-\pi_0) f^*_{H_0^c}(x)} - \frac{\pi_0 f^*_{H_0}(x)}{\pi_0 f^*_{H_0}(x) + (1-\pi_0) f^*_{H_0^c}(x)} \right\}$$
$$= -\left(\pi_0 f^*_{H_0}(x) + (1-\pi_0) f^*_{H_0^c}(x)\right)^{-2} (1-\pi_0) f^*_{H_0}(x) f^*_{H_0^c}(x)$$
$$= -\frac{1}{\pi_0} \Pi(H_0 \mid x) \Pi(H_0^c \mid x)$$
$$= -\frac{(1-\pi_0) BF_{H_0}}{(1-\pi_0+\pi_0 BF_{H_0})^2}. \tag{13}$$

Now observe that $BF^\epsilon_{H_0} = BF_{H_0}$ for all ϵ; i.e. the Bayes factor in favor of H_0 is independent of ϵ. This implies that the Gateaux derivative of the observed relative surprise is given by 0 when $BF_{H_0} \geq 1$ and is given by (13) when $BF_{H_0} < 1$.

It would seem then that the behavior of the two measures of surprise is very similar. Consider, however, the situation where $BF_{H_0} \geq 1$, $\Pi(H_0 \mid x) < \Pi(H_0^c \mid x)$ and $\Pi(H_0) = \pi_0$ is very small. In this situation the Gateaux derivative of the ORS is 0 while for the OS (13) is approximately equal to BF_{H_0}. So in situations where the data have lead to a large value of BF_{H_0}, but the posterior probability of H_0 is still less than $1/2$, the observed surprise is very sensitive to perturbations in the prior belief assigned to H_0.

Now consider the situation where $BF_{H_0} < 1$ while $\Pi(H_0 \mid x) \geq \Pi(H_0^c \mid x)$ so that the Gateaux derivative of the OS is 0 and the Gateaux derivative of the ORS is (13). We have that (13) is bounded above by

$$\frac{BF_{H_0}}{1-\pi_0+\pi_0 BF_{H_0}}$$

and, because $BF_{H_0} < 1$, this is an increasing function of π_0 which goes to the value 1 as $\pi_0 \to 1$.

So we have shown that the worst case behavior of the Gateaux derivative of the ORS in this problem is given by the upper bound 1 while the worst case behavior of the Gateaux derivative of the OS, using the relation $BF_{H_0} =$

$f^*_{H_0}(x)/f^*_{H_0^c}(x)$, is given by

$$\sup_x \frac{f^*_{H_0}(x)}{f^*_{H_0^c}(x)}.$$

Based on this we see that the ORS has superior robustness properties to the typical Bayesian inference $\Pi(H_0^c \mid x)$ for testing H_0 versus H_0^c.

As mentioned previously sometimes BF_{H_0} is used in these contexts and this has the appearance of being fully robust because $BF^{\epsilon}_{H_0} = BF_{H_0}$ for all ϵ. Recall, however, the need to calibrate the value of a BF_{H_0} through an equation such as (7). This leads to

$$\Pi(H_0^c \mid x) = \left(1 + \frac{\pi_0}{1-\pi_0} BF_{H_0}\right)^{-1}.$$

Therefore a value of $BF_{H_0} = 20$ when $\pi_0/(1-\pi_0) = 1$ implies $\Pi(H_0^c \mid x) = 1/21 = .048$. On the other hand $BF_{H_0} = 20$ when $\pi_0/(1-\pi_0) = 1/10$ implies $\Pi(H_0^c \mid x) = .5$. The point here is that the interpretation of the value BF_{H_0} does inherently depend on the values of the prior probability π_0 and so is not robust to this choice.

The key concept in determining relative surprise inferences is the ORS and so it makes sense to concentrate on assessing its robustness properties. Ultimately, however, we should look at the robustness properties of the specific inference. With hypothesis testing this is the ORS but with estimation, for example, we need to look at the LRSE. We note that

$$\frac{\pi^{\epsilon}_{\Upsilon}(\tau \mid x)}{\pi^{\epsilon}_{\Upsilon}(\tau)} = \frac{f^*_{\tau}(x)}{(1-\epsilon)\, m_{\pi}(x) + \epsilon m_q(x)}$$

for every $\epsilon \geq 0$ immediately implies the following result.

Theorem 5. The LRSE of τ is constant under perturbations that affect only the marginal prior of τ and so the Gateaux derivative of the LRSE is always 0.

Of course this is not true of the posterior mode which arises from the observed surprise and the principle of least surprise.

Example 6.

Suppose that $x \sim \text{Bernoulli}(\theta^{1/2})$, we take Π to be the Beta$(1, 3/2)$ distribution and the support measure ν is Lebesgue measure. Then the posterior distribution, when we observe $x = 1$, is given by the Beta$(3/2, 3/2)$ distribution and this has its mode at $1/2$. Suppose we take Q to be a Beta$(1, 1/2)$

distribution and this implies that the posterior distribution is Beta(3/2, 1/2) when we observe $x = 1$. Note that

$$\pi^{\epsilon}(\theta \mid x) = (1-\epsilon_x)\frac{\Gamma(3)}{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{3}{2}\right)}\theta^{1/2}(1-\theta)^{1/2} + \epsilon_x \frac{\Gamma(2)}{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{1}{2}\right)}\theta^{1/2}(1-\theta)^{-1/2}$$

and this always has its mode at 1. This is an extreme example of course but it illustrates an important point. In particular the Gateaux derivative of the LRSE is 0 while, in this case the Gateaux derivative of the posterior mode is infinity.

4 Conclusions

In Evans [1] relative surprise inferences were advocated for Bayesian contexts where a loss function was not prescribed. One of the advantages of these inferences is that they are invariant under reparamerizations and this is not the case for more traditional approaches to deriving inferences in Bayesian inference problems. Further the relative surprise inferences are seen to be primarily data driven in that they are based on how the data changes beliefs from *a priori* to *a posteriori* rather than being based on the posterior alone. It might be argued that for a specific application a particular parameterization is paramount but reasons must be supplied for this and these do not seem available in many applications. Further it might be argued that if one has a strong belief in a particular prior then the relative surprise approach weakens the amount of input that the prior has in determining an inference. While this is true, the results of this paper point to the negative side of that argument, namely, the lack of robustness to the choice of the prior for some traditional inferences.

The development here has excluded consideration of improper priors. Strictly speaking improper priors are excluded from the relative surprise formulation as they do not marginalize appropriately. Still we can consider many improper priors as limits of sequences of proper priors and consider the limiting surprise and relative surprise inferences as was done in Evans [1]. In such a case we could consider perturbations to this sequence and study the relative robustness properties of the limiting inferences. This is something we are currently investigating.

This paper has only considered linear perturbations to the marginal prior. More generally we will study other types of perturbations. Further there is the question of the effect of perturbing the conditional prior distribution of the full parameter given the marginal parameter. It seems reasonable to separate out the effect of perturbations to the marginal and conditional. We will consider these questions in further research work.

Acknowledgments

The authors thank the referees for some helpful comments. The authors were partially supported by the Natural Sciences and Engineering Research Council of Canada.

References

1. M. Evans, *Communications in Statistics - Theory and Methodology* **26(5)**, 1125 (1997).
2. M. Evans, D. A. S. Fraser and G. Monette, *Canadian Journal of Statistics* **13**, 137 (1985).
3. M. Evans and T. Zou, *On the robustness of relative surprise inferences to the choice of prior*, (Technical Report No. 0201, Dept. of Statistics, University of Toronto, Toronto, Canada, 2002).
4. P. Gustafson, *Annals of Statistics* **24**, 174 (1996).
5. P. Gustafson, *Journal of the American Statistical Association* **91**, 774 (1996).
6. P. Gustafson, and L. Wasserman, *Annals of Statistics* **23**, 2153 (1995).
7. R. E. Kass and A. E. Raferty, *Journal of the American Statistical Association* **90**, 773 (1995).
8. W. Rudin, *Real and Complex Analysis*, Second Edition (McGraw-Hill, New York, 1974).

USING SURVIVAL ANALYSIS IN PRETERM BIRTH STUDY

CHONG YAU FU

The Institute of Public Health, National Yang Ming University, 155 Li-Long St., Sec. 2, Shih-Pai, Taipei, Taiwan, 112
E-mail:chong@ym.edu.tw

SHIH HUA LIU

Department of Humanities and Science, National Yulin University of Science and Technology, 123 University Road, Touliu, Yunlin, Taiwan, 640

In conventional preterm birth studies the outcome variable was a binary data, namely whether or not the gestational age was less than 37 weeks. An absolute yes or no classification was used. Based on the progressing sense of gestational age, this study investigated the risk of preterm birth among different age groups using survival analysis. Using the key concept of the rank ordering of time, Kaplan-Meier estimate and the Cox model were applied to preterm birth studies. The applied results show that the relative risk of preterm birth for the 12-17, 18-19 and 20-24 year old age groups, varied with gestational age, and was 1.6-8.69, 1.2-5.76 and 0.92-1.52 respectively when compared with 25-29 years old. For the older age groups, namely 30-34, 35-39 and 40+ years old, the relative risks of preterm birth were 1.2, 1.54 and 1.74, respectively, constant in gestation age. Hence, applying survival analysis to preterm birth studies can provide more informative results than other approaches.

1 Introduction

Preterm birth is the leading cause of neonatal death and of morbidity in surviving infants (Rush, Davey and Segall [10]). The development of a complement system in preterm births was closely related to gestational age (Wolach *et al.* [11]). Hence, the duration of gestational age was important for preterm birth.

Previous studies of preterm birth, namely Fu *et al.* [5], Meis *et al.* [9], Rush *et al.* [10] etc., always used a binary data approach, merely considering whether or not gestational age was less than 37 weeks. Naturally, the study results thus failed to reflect the relationship between risk of preterm birth and gestational age.

Chiang [1] constructed an antenatal life table for fetal death and live birth in a prospective pregnancy study. The birth data was actually a history cohort data with the special feature of specific birthdays, the gestational age was a duration time of a fetus and the gestational age of less than 37 weeks was treated as an event. Hence, it should be straightforward to use survival analysis for a preterm birth study.

The censoring of data was a major issue in survival analysis. This study explores the performance of partial likelihood for the Cox model through censored data defined in this preterm birth study. The analytical results were expected to be more informative than those of other studies.

2 Statistical Method

Kaplan-Meier [7] estimate is an empirical method for estimating survival function. The Cox [2] model is commonly used for survival model incorporating covariates based on the partial likelihood function. Crowley and Hu [3] employ time-varying covariates in the Cox model.

Both Kaplan-Meier method and Cox model are based on the concept of the rank ordering of time rather than on time itself. According to this concept, the units of time employed might be days, months, hours, gestational age, and so on.

In this work, gestational age was equivalent to time, and birth at 37 weeks or less was an event. Births at 37 weeks or later were treated as censored data. The preterm birth data were then analyzed through survival analysis. Furthermore, graphs explored the original data and presented the modeled results. This study used the software, SAS 6.12 and S-plus (only for drawing plots).

3 Preterm Birth Study

Preterm birth is the cause of neonatal death, and of morbidity. Wolach *et al.* (1997) reported that the development of a complement system in preterm births is closely related to the gestational age . Hence, the conditions of preterm births varied with gestational age. And, the methodology of survival analysis is able to keep this information. Meis *et al.* [9] reported that the risk for preterm birth in different age groups displayed a U-shape, with younger and older mothers being high risks groups for preterm birth. In this study population, comparing with the 25-29 year age group, the relative risk for preterm birth in different age groups would be estimated. The data in this study came from the Computerized Medical Birth Registry (Lu *et al.* [8]), a data bank collected prospectively from ten Taiwanese hospitals from February 1993 to February 1995. The recruitment criteria included single birth, first parity, gestational age $\geq$ 24 weeks, and birth weight $\leq$ 500g. Totally, 17958 births met these criteria.

Gestational age were estimated from the last menstrual period, and births with gestational age $\geq$ 24 weeks and $\leq$ 37 weeks were considered preterm

births and defined as event data. Those normal births, births with gestational age > 37 weeks, were treated as censored data. Age groups were classified as 12-16, 17-19, 20-24, 25-29(reference group), 30-34, 35-39 and 40+ years old.

From a previous study (Fu *et al.* [5]), we realized that the risk factors for preterm births differed among younger and older age groups. Hence, the data set was divided into two parts, corresponding to younger and older age groups, both sharing a common reference group (25-29 years old). This approach increased the homogeneity each data set and did not introduce extra covariates into the model. However, the estimated coefficient values for the main risk factors (younger and older age) changed very slightly when additional covariates were brought into the models.

4 Results

Table 1 lists the number of preterm births and censored values for $24-37, 38^+$ weeks. Most of the births occurred in the $20-34$ year old age group. Almost all censored data occurred at 38^+ weeks. Figure 1 show the smoothed curves for the hazard of preterm birth over different age groups. These smoothed curves can be divided into three sections, for $24-28$, $28-33$, and $33-37$, weeks in both plots of Fig.1. In the $24-28$ week section, all curves display a constant hazard, except for those of the $12-17$ and ≥ 40 years age groups. Across all age groups, the hazard of preterm birth remained roughly constant at 28-33 weeks, but increased tremendously for gestational age ≥ 33 weeks.

Table 1.The Frequency of Fails (preterm birth) and Censoring, (#failed/#censored), in Gestational Age $24-37$ Weeks by age groups.

Weeks	$12-17$ yrs. $n=163$	$18-19$ yrs. $n=384$	$20-24$ yrs. $n=3,545$	$25-29$ yrs. $n=8,597$	$30-34$ yrs. $n=4,188$	$35-39$ yrs. $n=909$	$40+$ yrs. $n=172$
24^-	0/0	0/0	0/0	0/0	0/0	0/0	0/0
24	3/1	0/0	7/0	6/4	2/0	1/0	1/0
25	0/1	5/0	3/1	7/2	4/1	1/0	0/0
26	0/0	0/0	2/0	7/0	5/0	1/0	1/0
27	0/0	2/0	2/0	12/0	4/0	2/0	0/0
28	1/0	1/0	9/0	12/1	9/2	1/0	0/0
29	0/0	2/0	8/2	10/0	11/0	3/0	1/0
30	3/0	1/0	6/2	16/4	13/2	4/0	0/0
31	1/0	4/0	12/2	32/4	9/1	6/0	0/0
32	2/1	4/0	13/0	34/4	15/1	8/0	2/0
33	3/0	4/0	25/2	53/1	21/0	5/0	1/0
34	2/0	13/0	30/1	54/1	35/1	10/0	1/0
35	6/0	7/0	51/2	85/2	65/0	14/0	5/0
36	8/0	8/0	78/2	213/0	120/0	40/0	5/0
37	12/0	31/0	220/0	586/3	336/1	83/1	21/1
38^+	0/119	0/302	0/3069	0/7448	0/3529	0/729	0/133

Tables 2 and 3, and Fig. 2, summarize the results of Cox model. In table 2, model 2 provided significant time dependent results for younger age groups. Compared to the $25-29$ year old age group, the relative risk of the younger $12-17, 18-19$ and $20-24$ age groups varied with gestational age, and ranged

Table 2. Estimated Cox Models for Gestational Age in preterm births; Age Groups: 12-17, 18-19, 20-24, 25-29(reference group); Standard Errors in Parentheses

Covariates	Model 1	Model 2
A1: $12-17$ *vs.* $25-29$	0.768 (0.159)	5.281 (1.498)
A2: $18-19$ *vs.* $25-29$	0.516 (0.116)	4.631 (1.164)
A3: $20-24$ *vs.* $25-29$	0.005 (0.055)	1.293 (0.711)
A1*Gestational Age	NA	−0.130 (0.044)
A2*Gestational Age	NA	−0.118 0.034
A3*Gestational Age	NA	−0.037 (0.020)

Total sample size: 12, 675; Events: 1,712;
Censored: 10,963 (Censoring Rate = 86.5%)

Table 3. Estimated Cox Models for Gestational Age in preterm births; Age Groups: $25-29$(reference group), $30-34, 35-39, 40^{+}$; Standard Errors in Parentheses

Covariates	Model 3	Model 4
A5: $30-34$ *vs.* $25-29$	0.179 (0.049)	−0.119 (0.686)
A6: $35-39$ *vs.* $25-29$	0.441 (0.081)	1.018 (1.050)
A7: 40^{+} *vs.* $25-29$	0.567 (0.165)	0.833 (2.184)
A5*Gestational Age	NA	0.008 (0.019)
A6*Gestational Age	NA	−0.016 0.030
A7*Gestational Age	NA	−0.008 (0.062)

Total sample size: 13, 849; Events: 1,991;
Censored: 11, 858 (Censoring Rate = 85.6%)

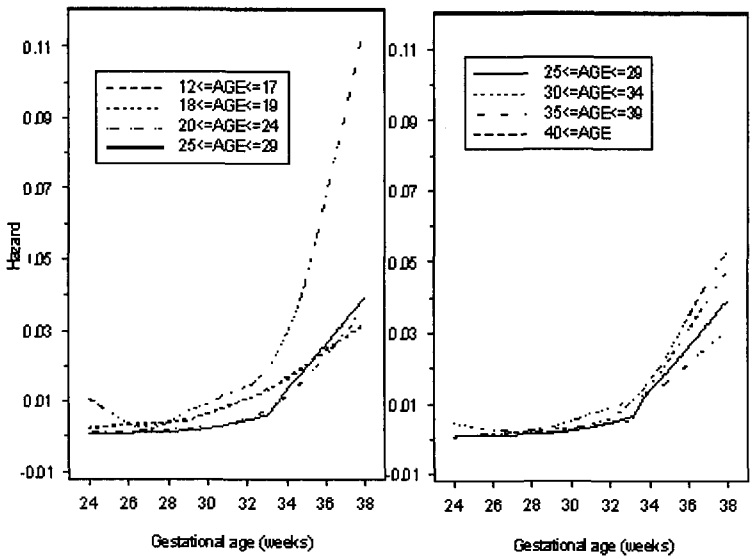


Figure 1: Smooth Hazard Functions for different age groups

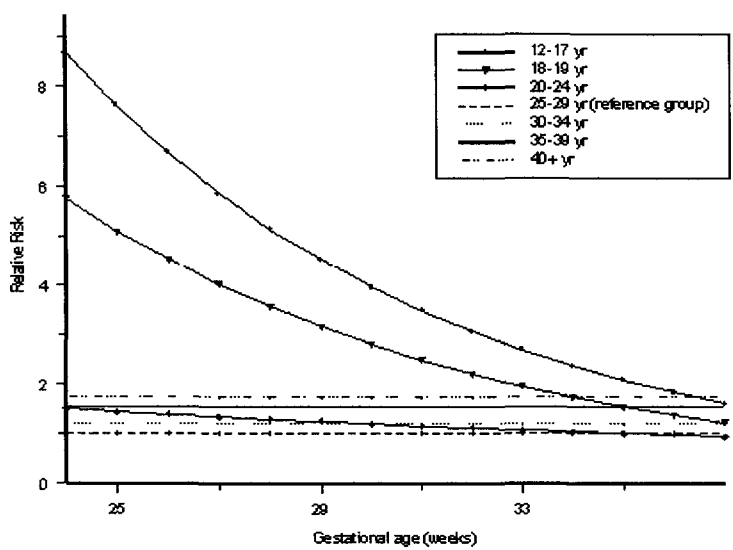


Figure 2: Results from Cox models; age specific R.R. (reference group 25-29 yrs.) for preterm birth in gestational age 24-37 weeks.

from $1.6-8.67, 1.2-5.76$ and $0.92-1.52$, respectively. However, the relative risk for the older $30-34, 35-39$ and 40^+ age groups did not vary significantly with gestational age (model 4 of table 3), and their relative risks were 1.2, 1.54 and 1.74, respectively. Figure 2 displays the relative risk in younger and older age groups comparing with the $25-29$ year old group.

Comparing the empirical results (Fig. 1) with the modeled result (Fig. 2), the common impact of risk of preterm birth were the risk for younger age groups varying with gestational age while for older age groups this risk was constant. Meanwhile, the estimated relative risks were very significant for both models.

5 Discussion

The main methods employed in this application study were the Kaplan-Meier estimate and Cox models. Besides the key issue of censored data, the basic issues included time rank ordering and risk set. The appropriateness or otherwise of the application of the model can be judged based on these issues. This study is an original work applying survival analysis to investigate preterm birth. Gestational ages ≥ 24 weeks and ≤ 37 weeks were considered event data, and their ranked order was analyzed. Meanwhile, a gestational age of greater than 37 weeks was censored data, contributing to the risk set of denominators in parameter estimation, occurring in the survival estimator of Kaplan-Meier or the partial likelihood of the Cox-model. Consequently, the estimated parameters are based on the entire study population.

Based on 17,958 births, the results of log-rank trend testing (Collett [4]) indicated that ranked by age group, the risk for preterm birth was ordered as follows: $(12-17) > (18-19) > (40^+) > (35-39) > (20-24) > (30-34) > (25-29)$, with $p-\text{value} < 0.0001$. This testing was based on the sense of average risk in weeks $24-37$. However, the relative risks of preterm birth varied with gestational age from Cox models for younger age groups. Hence, compared to the results of Cox models and the logrank test, Cox models not only dealt with the time dependent problem, but also estimated their relative risks.

The risk factors for preterm births included maternal age, marital status, education, parity, prenatal care, and so on. Model building process revealed that maternal age was a very strong risk factor for preterm birth. Furthermore, with the very large sample size herein, whether or not other risk factors were included, the coefficient estimations of maternal age made very little difference. Consequently, this investigation only included the maternal age in Cox models.

However, the risk factors for preterm birth differed between the younger and older age groups. Hence, a sufficiently large sample size would offer the benefit of homogeneity in dealing with younger and older age groups separately. Finally, from two separate Cox models, the relative risk of preterm birth varied significantly with gestational age in the younger age group, but not in the older age group. Figure 1 does note this same situation.

Reviewing previous studies, including Meis *et al.* [9], Fraser *et al.* [6], and Wessel *et al.* [12] preterm birth was defined as less than 37 weeks, thus classifying it as a binary response data. This work considered each gestational age using a survival analysis approach and thus incorporated a sense of progressive risk. Hence, these results were more informative than those that used a binary response.

Acknowledgments

I would like to thank Dr. Jen Her Lu for providing data from Computerized Medical Birth Registry and for useful discussion.

References

1. C.L. Chiang , In *The life table and its applications* (Krieger, Florida, 1984).
2. D.R. Cox, *Journal of the Royal Statistical Society, Ser. B* **74**, 187 (1972).
3. J. Crowley and M. Hu, *Journal of American Statistical Association* **78**, 27 (1977).
4. D. Collett, In *Modeling Survival Data in Medical Research*, eds.: C. Chatfield and J.V. Zidek (Chapman and Hall ,London, 1994).
5. C. Y. Fu *et al.*, *Chinese J. of Public Health (Taipei)* **18(3)**, 228 (1999).
6. A.M. Fraser, J.E. Brockert and R.H. Ward, *The New England Journal of Medicine* **332**, 1113 (1995).
7. E.L. Kaplan and P. Meier, *Journal of American Statistical Association* **53**, 457 (1958).
8. J.H. Lu *et al.*, *Medical Information* **19(4)**, 323 (1994).
9. P.J. Meis *et al.*, *American Journal of Obstetrics and Gynecology* **173**, 590 (1995).
10. R.W. Rush , D.A. Davey and M.L. Segall, *British Journal of Obstetrics and Gynecology* **85**, 806 (1978).
11. B. Wolach *et al.*, *Acta Pædiatrica* **86(5)**, 523 (1997).
12. H. Wesselet *al.*, *Acta Obstetricia Gynecologica Scandinavica* **75**, 360 (1996).

ASYMPTOTIC FORMS AND BOUNDS FOR TAILS OF CONVOLUTIONS OF COMPOUND GEOMETRIC DISTRIBUTIONS, WITH APPLICATIONS

JUN CAI

Department of Statistics and Actuarial Sciences, University of Waterloo, Ontario, Canada N2L 3G1

E-mail: jcai@icarus.math.uwaterloo.ca

JOSÉ GARRIDO

Department of Mathematics and Statistics, Concordia University, Montreal, Quebec, Canada H4B 1R6

E-mail: garrido@vax2.concordia.ca

In this paper, we derive exponential and subexponential asymptotic forms for tails of convolutions of compound geometric distributions. General lower and upper bounds for the tails are given, which can be used in some cases to determine a closer tail approximation. Applications of these results are given to the ruin probability in the classical risk process perturbed by a diffusion; previous results are easily derived and a theorem of Veraverbeke [26] is generalized. In addition, two-sided bounds for the ruin probability with large claim sizes are given, extending the bounds of Dickson [10] to the diffusion risk model.

1 Introduction

Let $\{X_i, i \geq 1\}$ be a sequence of *i.i.d.* non-negative random variables with common distribution F and $F(0) = 0$. Further, let N be a geometric random variable with $\Pr\{N = n\} = qp^n$, $n = 0, 1, 2, \ldots$ and $p = 1 - q$, for $0 < q < 1$, which is independent of $\{X_i, i \geq 1\}$. $S_N = \sum_{i=1}^{N} X_i$ is said to be compound geometric, where $S_N = 0$ if $N = 0$. Its distribution function is denoted by $H(x) = \Pr\{S_N \leq x\}$, $x \geq 0$.

Suppose that Y is another non-negative random variable with distribution G and $G(0) = 0$, where Y, N and $\{X_i, i \geq 1\}$ are independent. Then, the convolution $H * G$ of the compound geometric distribution H and distribution G, i.e. the distribution of $S_N + Y$, arises in many applied probability models, such as regenerative processes (Cohen [8], Kalashnikov [18] and Keilson [19]) risk theory (Dufresne and Gerber [11], Sundt and Teugels [23] and Veraverbeke [26]) and queueing theory (Asmussen [1], van Hoorn [25] and Szekli [24]). Many distributions of interest in these works can be expressed in the form $H * G$ of the distribution of $S_N + Y$.

Denote by $W(x) = H * G(x)$, consider $\overline{W}(x) = 1 - W(x)$, the tail or sur-

vival function of $W(x)$. It is well-known (Rényi's theorem) that if F has a finite mean, then for any $x \geq 0$, $\overline{H}(x)/\int_0^\infty \overline{H}(y)\,dy \to e^{-x}$ as $q \to 0$ or equivalently $E(N) \to \infty$. Furthermore, Keilson [19] (see also Kalashnikov [18]) showed that if F has a finite second moment and G has a finite mean, then a similar limit theorem holds for $\overline{W}(x)$, namely for any $x \geq 0$, $\overline{W}(x)/\int_0^\infty \overline{W}(y)\,dy \to e^{-x}$ as $E(N) \to \infty$. However, in many applications, we are interested in the asymptotic behavior and bounds for $\overline{W}(x)$.

The asymptotic behavior and bounds for the tail $\overline{H}(x)$ of the compound geometric distribution function $H(x)$ is well-known. The purpose of this paper is to give a more complete description of the asymptotic behavior for $\overline{W}(x)$, obtain the asymptotic estimates for $\overline{W}(x)$ under various situations, derive bounds for $\overline{W}(x)$ in heavy tailed cases and consider the applications of these results in risk theory (for bounds see *e.g.* Cai and Garrido [4]).

The paper is organized as follows: in Section 2, we derive an exponential asymptotic form for $\overline{W}(x)$ in terms of Lundberg's coefficient using the key renewal theorem.

In Section 3, we consider subexponential asymptotic forms for $\overline{W}(x)$. Here, general lower and upper bounds for $\overline{W}(x)$ are given first, the bounds indicate the possible asymptotic form for $\overline{W}(x)$ and are also used to determine a closer approximation for $\overline{W}(x)$ in some cases.

Section 4 discusses the asymptotic forms for $\overline{W}(x)$ in the intermediate case, when exponential moments exist but Lundberg's coefficient does not.

In Section 5, as an application of the results for $\overline{W}(x)$, we consider the ruin probability in the classical process perturbed by a diffusion (see Rolski *et al.* [22]). The asymptotic estimates of the ruin probability derived by Dufresne and Gerber [11], Gerber [16] and Veraverbeke [26]) are easily obtained. A theorem of Veraverbeke [26] is also generalized.

In addition, two-sided bounds for the ruin probability with large claims are given, thus extending the bound of Dickson [10] to diffusion risk models.

2 Light-tail asymptotics

In general, $\overline{W}(x)$ does not admit an exponential asymptotic form, for example, see Remark 3.5 of this paper. But if conditions similar to those in Cramér-Lundberg's theorem hold, then there exists an exponential asymptotic form for $\overline{W}(x)$, which is stated in the following theorem.

Throughout this paper, $f(x) \sim g(x)$ means that $f(x)/g(x) \to 1$ as $x \to \infty$ and $f(x) = o(g(x))$ means that $f(x)/g(x) \to 0$ as $x \to \infty$ while $m_B(s) = \int_0^\infty e^{sx}\,dB(x)$ denotes the moment generating function of B on $[0, \infty)$.

Theorem 2.1. Let F be non-lattice and κ a constant such that

$$\int_0^\infty e^{\kappa x}\, dF(x) = 1/p\,. \tag{1}$$

If $m_G(\kappa) = \int_0^\infty e^{\kappa x}\, dG(x) < \infty$ then

$$\overline{W}(x) \sim \frac{q\, m_G(\kappa)}{p\kappa\beta} e^{-\kappa x}\,, \quad \text{if } \beta = \int_0^\infty x e^{\kappa x}\, dF(x) < \infty\,, \tag{2}$$

$$= o(e^{-\kappa x})\,, \qquad \text{if } \beta = \infty\,, \tag{3}$$

where, κ in (1) is called Lundberg's coefficient.

Proof. See Willmot and Lin [28], Corollary 9.3.2., pp. 175-6. □

Remark 2.1. When $Y = 0$, G is degenerate at zero, and thus $m_G(\kappa) = 1$, $\overline{W}(x) = \overline{H}(x)$ and Theorem 2.1 is reduced to Cramér-Lundberg's theorem, i.e. under the conditions of Theorem 2.1 about F

$$\overline{H}(x) \sim \frac{q}{p\kappa\beta}\, e^{-\kappa x}\,, \quad \text{if } \beta < \infty\,, \tag{4}$$

$$= o(e^{-\kappa x})\,, \qquad \text{if } \beta = \infty\,. \tag{5}$$

3 Heavy-tail asymptotics

In this Section, we consider the case when F or G are heavy-tailed, in particular, subexponential distributions. First, the two following theorems give general lower and upper bounds for $\overline{W}(x)$, which indicate possible asymptotic forms used to determine, in some cases, a closer approximation for $\overline{W}(x)$.

Theorem 3.1. For any $x \geq 0$,

$$\overline{W}(x) \;\geq\; \frac{p\,\overline{F}(x) + q\,\overline{G}(x)}{p\,\overline{F}(x) + q}\,. \tag{6}$$

Proof. Since $\overline{W}(x) = \Pr\{S_N + Y > x\}$ is the survival function of the random variable $S_N + Y$, $\overline{W}(x)$ is decreasing. By the usual renewal argument, we have

$$\overline{W}(x) \geq q\,\overline{G}(x) + p\,\overline{F}(x) + p\,\overline{W}(x) \int_0^x dF(y)\,, \quad x \geq 0\,,$$

$$= q\,\overline{G}(x) + p\,\overline{F}(x) + p\,\overline{W}(x) F(x)\,,$$

this implies that (6) holds. □

Remark 3.1. Take $Y = 0$ in Theorem 3.1, we get a lower bound for the tail $\overline{H}(x)$ of the compound geometric distribution function $H(x)$, namely

$$\overline{H}(x) \geq \frac{p\overline{F}(x)}{p\overline{F}(x)+q}, \quad \text{for any } x \geq 0. \tag{7}$$

This gives a derivation of a result known for the ruin probability in the classical risk process [*e.g.* see Theorem 3.1 of De Vylder and Goovaerts (1984)].

Theorem 3.2. If F has a finite mean $E(X_1)$ and $G(x)$ has a decreasing density function, then for any $x > 0$

$$\overline{W}(x) \leq \frac{p\overline{F}(x)+q\overline{G}(x)+\delta(x)}{p\overline{F}(x)+q}, \tag{8}$$

where $\delta(x) = pG(x)\{x^{-1}E(X_1) - \overline{F}(x)\} \to 0$ as $x \to \infty$.

Proof. Since $W(x) = H * G(x) = \Pr\{S_N + Y \leq x\}$, by conditioning on the value of Y, we get that for any $x \geq 0$,

$$\overline{W}(x) = \overline{G}(x) + \int_0^x \overline{H}(x-y)\,dG(y) \tag{9}$$

$$= \overline{G}(x) + \int_0^x \overline{H}(x-y)G'(y)\,dy. \tag{10}$$

But, we know that if $f(x)$ and $g(x)$ are integrable functions with different monotonicity on $[a, b]$, $a < b$, then

$$\int_a^b f(x)g(x)\,dx \leq \frac{1}{b-a}\int_a^b f(x)\,dx \int_a^b g(x)\,dx. \tag{11}$$

Thus, by (10), (11) and the fact that $\overline{H}(x-y)$ is increasing in y over $[0, x]$, we get that for any $x > 0$,

$$\overline{W}(x) \leq \overline{G}(x) + \frac{1}{x}\int_0^x \overline{H}(x-y)\,dy \int_0^x G'(y)\,dy \tag{12}$$

$$= \overline{G}(x) + \frac{G(x)}{x}\int_0^x \overline{H}(t)\,dt$$

$$= \overline{G}(x) + \frac{G(x)}{x}\left[E(S_N) - \int_x^\infty \overline{H}(t)\,dt\right]. \tag{13}$$

Since S_N is compound geometric, the distribution H of S_N is a New Worse than Used (NWU) distribution [Lemma 2.1 of Brown (1990)], which implies that H is a NWUE distribution, i.e.

$$\int_x^\infty \overline{H}(t)\,dt \geq E(S_N)\overline{H}(x), \; x \geq 0. \tag{14}$$

Thus, by (13), (14) and (7), we get that for any $x > 0$

$$\begin{aligned}\overline{W}(x) &\le \overline{G}(x) + x^{-1}G(x)E(S_N)[1 - \overline{H}(x)] \\ &\le \overline{G}(x) + x^{-1}G(x)E(S_N)\left[1 - \frac{p\,\overline{F}(x)}{p\,\overline{F}(x) + q}\right] \\ &= \overline{G}(x) + \frac{x^{-1}G(x)E(S_N)q}{p\,\overline{F}(x) + q}\,. \end{aligned} \tag{15}$$

But $E(S_N) = E(\sum_{i=1}^{N} X_i) = E(N)E(X_i) = pE(X_1)/q$. Hence $qE(S_N) = pE(X_1)$. This, together with (15), implies that (8) holds. □

Remark 3.2. The requirement in Theorem 3.2 that G has a decreasing density function is not restrictive. In fact, this condition can be satisfied by many distributions such as the class of the equilibrium distribution function $F_e(x) = \int_0^x \overline{F}(y)\,dy / \int_0^\infty \overline{F}(y)\,dy$ with decreasing density function $f(x) = \overline{F}(x) / \int_0^\infty \overline{F}(y)\,dy$, which often arises in risk theory, reliability, queueing and renewal theory. In addition, the class of decreasing failure rate (DFR) distributions have decreasing density functions since $f(x) = r(x)\,\overline{F}(x)$, where $r(x) \ge 0$ is the failure rate function of F, which is decreasing in this case.

Corollary 3.1. Combining Theorems 3.1 and 3.2 we get that under the conditions and notation of Theorem 3.2, for any $x > 0$

$$\frac{p\,\overline{F}(x) + q\,\overline{G}(x)}{p\,\overline{F}(x) + q} \le \overline{W}(x) \le \frac{p\,\overline{F}(x) + q\,\overline{G}(x) + \delta(x)}{p\,\overline{F}(x) + q}\,. \tag{16}$$

Since $\delta(x) \to 0$ as $x \to \infty$, (16) indicates that

$$\frac{p\,\overline{F}(x) + q\,\overline{G}(x)}{p\,\overline{F}(x) + q}$$

may be an asymptotic form for $\overline{W}(x)$ as $x \to \infty$. Indeed, we show below that under various situations, this is precisely the asymptotic form of $\overline{W}(x)$.

Corollary 3.2. Under Theorem 3.2, if $\lim_{x\to\infty} x\,\overline{G}(x) = \infty$ then

$$\overline{W}(x) \sim \frac{p\,\overline{F}(x) + q\,\overline{G}(x)}{p\,\overline{F}(x) + q}\,. \tag{17}$$

Proof. By (16), we get for any $x > 0$,

$$1 \le \overline{W}(x)\,\frac{p\,\overline{F}(x) + q}{p\,\overline{F}(x) + q\,\overline{G}(x)} \le 1 + \frac{pG(x)\{E(X_1) - x\,\overline{F}(x)\}}{px\,\overline{F}(x) + qx\,\overline{G}(x)}\,. \tag{18}$$

Since F has a finite mean, $x\,\overline{F}(x) \to 0$ as $x \to \infty$. Thus, (18) and $x\,\overline{G}(x) \to \infty$ as $x \to \infty$ imply that

$$\overline{W}(x) \sim \frac{p\,\overline{F}(x) + q\,\overline{G}(x)}{p\,\overline{F}(x) + q}.$$

□

Remark 3.3. There exist classes of distributions that satisfy the conditions of Corollary 3.2, for example the Pareto distribution with density function

$$g(x) = \frac{\alpha c^{\alpha}}{(c+x)^{1+\alpha}}\,, \quad x \geq 0\,, \quad \text{where } 0 < \alpha < 1\,,\ c > 0\,,$$

or more generally, the Burr distribution with distribution function

$$F(x) = 1 - \left(\frac{\lambda}{\lambda + x^{\tau}}\right)^{\alpha}\,, \quad x \geq 0\,, \quad \text{where } \lambda > 0\,, 0 < \alpha \leq 1\,, 0 < \tau \leq 1\,.$$

But, it should be pointed out that the condition $x\,\overline{G}(x) \to \infty$ as $x \to \infty$ is restrictive, since it implies that G has no finite mean.

On the other hand, under the conditions of Corollary 3.2,

$$\frac{p\,\overline{F}(x) + q\,\overline{G}(x)}{p\,\overline{F}(x) + q}$$

gives a closer approximation to $\overline{W}(x)$ than $\overline{G}(x)$ does, as seen by Theorem 3.1 and the fact that for any $x \geq 0$,

$$\frac{p\,\overline{F}(x) + q\,\overline{G}(x)}{p\,\overline{F}(x) + q} \geq \overline{G}(x).$$

We know that if h is an asymptotic form of an unknown function T, i.e. $T(x) \sim h(x)$, then any function g satisfying $g(x) \sim h(x)$ is also an asymptotic form of T. Hence, it is interesting to find a closer asymptotic form among known approximations. As shown above, a closer approximation can be derived by combining bounds and asymptotic forms.

Furthermore, we notice that the only relation between F and G required in Corollary 3.2 is that $\overline{G}(x)$ dominate asymptotically $\overline{F}(x)$. If other relations between them are assumed and subexponentiality (as defined below) is further imposed on F or G, we can derive additional results for $\overline{W}(x)$ as follows.

Definition 3.1. A distribution B on $[0, \infty)$ is said to be subexponential, denoted by $B \in \mathcal{S}$, if $\overline{B^{(2)}}(x) \sim 2\,\overline{B}(x)$.

Subexponential distributions are heavy tailed; typical examples are the Pareto or Lognormal distributions. The following Lemma combines Proposition 1 of Embrechts *et al.* [14] and Theorem 2 of Chistyakov [5].

Lemma 3.1. Suppose that F_1 and F_2 are two distributions on $[0, \infty)$.

(i) If $F_2 \in \mathcal{S}$ and $\overline{F_1}(x) = o(\overline{F_2}(x))$, then $F_1 * F_2 \in \mathcal{S}$ and $\overline{F_1 * F_2}(x) \sim \overline{F_2}(x)$.

(ii) If $F_1 * F_2 \in \mathcal{S}$ and $\overline{F_1}(x) = o(\overline{F_1 * F_2}(x))$ then $F_2 \in \mathcal{S}$ and $\overline{F_2}(x) \sim \overline{F_1 * F_2}(x)$.

(iii) If $F_2 \in \mathcal{S}$, then for any $\varepsilon > 0$, $\lim_{x \to \infty} e^{\varepsilon x} \overline{F_2}(x) = \infty$, i.e. $e^{-\varepsilon x} = o(\overline{F_2}(x))$.

Theorem 3.3. Suppose that $\overline{G}(x) = o(\overline{F}(x))$. The three following assertions are equivalent:

(i) $F \in \mathcal{S}$,

(ii) $W \in \mathcal{S}$,

(iii) $\overline{W}(x) \sim \{p\overline{F}(x) + q\overline{G}(x)\}/\{p\overline{F}(x) + q\} \sim p\overline{F}(x)/q$.

Proof. First, it is clear that $\overline{G}(x) = o(\overline{F}(x))$ implies

$$\frac{p\overline{F}(x) + q\overline{G}(x)}{p\overline{F}(x) + q} \sim \frac{p\overline{F}(x)}{q} . \tag{19}$$

By Corollary 3.2 of Embrechts *et al.* [14], we know that the following three conditions are equivalent:

(a) $F \in \mathcal{S}$, (b) $H \in \mathcal{S}$, (c) $\overline{H}(x) \sim p\overline{F}(x)/q$.

Hence, if $F \in \mathcal{S}$, by (b), (c) and $\overline{G}(x) = o(\overline{F}(x))$, we get that $H \in \mathcal{S}$ and

$$\frac{\overline{G}(x)}{\overline{H}(x)} = \frac{\overline{G}(x)}{\overline{F}(x)} \times \frac{\overline{F}(x)}{\overline{H}(x)} \to 0 \quad \text{as} \quad x \to \infty ,$$

i.e. $\overline{G}(x) = o(\overline{H}(x))$, hence (c) and (i) of Lemma 3.1 imply that $W = G * H \in \mathcal{S}$ and $\overline{W}(x) \sim \overline{H}(x) \sim p\overline{F}(x)/q$.

Conversely, if $W = G * H \in \mathcal{S}$, by Theorem 3.1 and $\overline{G}(x) = o(\overline{F}(x))$:

$$0 \leq \frac{\overline{G}(x)}{\overline{W}(x)} \leq \frac{\{p\overline{F}(x) + q\}\overline{G}(x)}{p\overline{F}(x) + q\overline{G}(x)} = \frac{\{p\overline{F}(x) + q\}\overline{G}(x)/\overline{F}(x)}{p + q\overline{G}(x)/\overline{F}(x)} \to 0 \text{ as } x \to \infty ,$$

this is to say that $\overline{G}(x) = o(\overline{W}(x))$, thus, (ii) of Lemma 3.1 and (c) imply that $H \in \mathcal{S}$ and $\overline{W}(x) \sim \overline{H}(x) \sim p\overline{F}(x)/q$.

So, we have shown that $F \in \mathcal{S} \Leftrightarrow W \in \mathcal{S}$. In addition, the above proof also showed that $F \in \mathcal{S} \Rightarrow \overline{W}(x) \sim p\overline{F}(x)/q$. Thus, in order to complete the proof of Theorem 3.3, we still need to prove that

$$\overline{W}(x) \sim \frac{p}{q}\overline{F}(x) \Rightarrow F \in \mathcal{S} . \tag{20}$$

By definition, $W(x) = \Pr\{S_N + Y \le x\} = \sum_{n=0}^{\infty} qp^n G * F^{(n)}(x)$, and hence

$$\overline{W}(x) = \sum_{n=0}^{\infty} qp^n \overline{G * F^{(n)}}(x) ,$$

which implies that

$$\overline{G * F^{(2)}}(x) = \frac{1}{qp^2}\left[\overline{W}(x) - \sum_{n\neq 2} qp^n \overline{G * F^{(n)}}(x)\right] . \tag{21}$$

Since Y is a non-negative random variable, for any integer $k \ge 1$,

$$\begin{aligned}\overline{G * F^{(k)}}(x) &= \Pr\{Y + X_1 + \cdots + X_k > x\} \\ &\ge \Pr\{X_1 + \cdots + X_k > x\} \;=\; \overline{F^{(k)}}(x) \\ &\ge \Pr\{\max(X_1, \ldots, X_k) > x\} \\ &= 1 - [F(x)]^k = \; \overline{F}(x) \sum_{n=0}^{k-1} [F(x)]^n . \end{aligned} \tag{22}$$

(22) implies that

$$\liminf_{x\to\infty} \frac{\overline{G * F^{(k)}}(x)}{\overline{F}(x)} \;\ge\; \liminf_{x\to\infty} \frac{\overline{F^{(k)}}(x)}{\overline{F}(x)} \;\ge\; k . \tag{23}$$

Clearly, (22) and (23) are also true for $k = 0$, thus by (21) and (22):

$$\frac{\overline{F^{(2)}}(x)}{\overline{F}(x)} \;\le\; \frac{\overline{G * F^{(2)}}(x)}{\overline{F}(x)} \;\le\; \frac{1}{qp^2}\left[\frac{\overline{W}(x)}{\overline{F}(x)} - \sum_{n\neq 2} qp^n \frac{\overline{F^{(n)}}(x)}{\overline{F}(x)}\right] . \tag{24}$$

Thus by (24), (23) and $\overline{W}(x) \sim p\,\overline{F}(x)/q$, we get that

$$\begin{aligned}\limsup_{x\to\infty} \frac{\overline{F^{(2)}}(x)}{\overline{F}(x)} &\le \frac{1}{qp^2}\left[\frac{p}{q} - \sum_{n\neq 2} qp^n \liminf_{x\to\infty} \frac{\overline{F^{(n)}}(x)}{\overline{F}(x)}\right] \\ &\le \frac{1}{qp^2}\left[\frac{p}{q} - \sum_{n\neq 2} nqp^n\right] = \frac{1}{qp^2}\left[\frac{p}{q} + 2qp^2 - \sum_{n=0}^{\infty} nqp^n\right] = 2 , \end{aligned} \tag{25}$$

hence, (23) and (25) imply that $\lim_{x\to\infty} \overline{F^{(2)}}(x)/\,\overline{F}(x) = 2$, i.e. $F \in \mathcal{S}$. □

It is interesting to note that (20) holds and is independent of the condition that $\overline{G}(x) = o(\overline{F}(x))$. In addition, if G is an exponential distribution, using Lemma 3.1(iii) and following the proof of Theorem 3.3, we get directly the following corollary.

Corollary 3.3. If G is an exponential distribution, then Theorem 3.3 holds.

Theorem 3.4. Suppose that $\overline{F}(x) = o(\overline{G}(x))$ and $F \in \mathcal{S}$. The two following assertions are equivalent:

(i) $G \in \mathcal{S}$, (ii) $W \in \mathcal{S}$,

and either one of them implies that:

(iii) $\overline{W}(x) \sim \{p\overline{F}(x) + q\overline{G}(x)\}/\{p\overline{F}(x) + q\} \sim \overline{G}(x)$.

Proof. By (c) in the proof of Theorem 3.3, we know that

$$\frac{\overline{H}(x)}{\overline{G}(x)} = \frac{\overline{H}(x)}{\overline{F}(x)} \times \frac{\overline{F}(x)}{\overline{G}(x)} \to 0 \quad \text{as} \quad x \to \infty ,$$

i.e. $\overline{H}(x) = o(\overline{G}(x))$, hence if $G \in \mathcal{S}$, by Lemma 3.1(i), we get that $W = H * G \in \mathcal{S}$ and $\overline{W}(x) \sim \overline{G}(x)$. But, $\overline{F}(x) = o(\overline{G}(x))$ implies that

$$\frac{p\overline{F}(x) + q\overline{G}(x)}{p\overline{F}(x) + q} \sim \overline{G}(x).$$

Conversely, if $W = H * G \in \mathcal{S}$, by Theorem 3.1 and the proof of Theorem 3.3(c), we get that

$$0 \leq \frac{\overline{H}(x)}{\overline{W}(x)} \leq \frac{\{p\overline{F}(x) + q\}\overline{H}(x)}{p\overline{F}(x) + q\overline{G}(x)} = \frac{\{p\overline{F}(x) + q\}\overline{H}(x)/\overline{G}(x)}{q\overline{F}(x)/\overline{G}(x) + q} \to 0 \text{ as } x \to \infty ,$$

i.e. $\overline{H}(x) = o(\overline{W}(x))$. Then Lemma 3.1(ii) $\Rightarrow$ $G \in \mathcal{S}$ and $\overline{W}(x) \sim \overline{G}(x)$. □

Remark 3.4. We know that

$$\frac{p\overline{F}(x) + q\overline{G}(x)}{p\overline{F}(x) + q} \geq \frac{p\overline{F}(x)}{q} \quad \text{if and only if} \quad q^2\overline{G}(x) \geq [p\overline{F}(x)]^2.$$

Thus, in view of Theorem 3.3(iii), if $F \in \mathcal{S}$, $\overline{G}(x) = o(\overline{F}(x))$ and $q^2\overline{G}(x) \geq [p\overline{F}(x)]^2$ as $x \to \infty$, (for example, $\overline{G}(x) = [\overline{F}(x)]^{3/2}$), then by Theorem 3.1, we know that $\{p\overline{F}(x) + q\overline{G}(x)\}/\{p\overline{F}(x) + q\}$ is a closer approximation for $\overline{W}(x)$ than $p\overline{F}(x)/q$ is. By an argument similar to that in Remark 3.3, we also know that under the conditions of Theorem 3.4 and if $G \in \mathcal{S}$, then $\{p\overline{F}(x) + q\overline{G}(x)\}/\{p\overline{F}(x) + q\}$ is a closer approximation to $\overline{W}(x)$ than $\overline{G}(x)$.

Theorem 3.5. If $\overline{G}(x) \sim \overline{F}(x)$ and $F \in \mathcal{S}$, then

$$\overline{W}(x) \sim \frac{p\overline{F}(x) + q\overline{G}(x)}{p\overline{F}(x) + q} \sim \frac{\overline{F}(x)}{q} \sim \frac{\overline{G}(x)}{q} .$$

Proof. $F \in \mathcal{S}$ implies that $H \in \mathcal{S}$ and $\overline{H}(x) \sim p\overline{F}(x)/q$. Hence,

$$\frac{\overline{G}(x)}{\overline{H}(x)} = \frac{\overline{G}(x)}{\overline{F}(x)} \times \frac{\overline{F}(x)}{\overline{H}(x)} \to \frac{q}{p} \text{ as } x \to \infty .$$

Thus, by Theorem 1 of Cline [6] (see also Theorem 4.3 of Goldie and Klüppelberg [17]) we know that

$$\overline{W}(x) = \overline{G * H}(x) \sim (\frac{q}{p} + 1)\overline{H}(x) = \frac{\overline{H}(x)}{p} \sim \frac{\overline{F}(x)}{q} .$$

On the other hand, $\overline{F}(x) \sim \overline{G}(x)$ implies that

$$\frac{p\overline{F}(x) + q\overline{G}(x)}{p\overline{F}(x) + q} \sim \frac{\overline{F}(x)}{q} \sim \frac{\overline{G}(x)}{q} ,$$

so Theorem 3.5 holds. □

Remark 3.5. The results of this Section also show that the lower bound in Theorem 3.1 is asymptotically exact for $\overline{W}(x)$ as $x \to \infty$, in the subexponential case. Bounds for $\overline{W}(x)$ were considered in Kalashnikov [18] and Willmot and Lin [27]). The bounds for $\overline{W}(x)$ in (5.1) and (5.2) of Kalashnikov [18] are asymptotically exact as $E(N) \to \infty$. The bounds of Willmot and Lin [27] are applicable to the tail of convolutions of more general compound distributions; these are based on a generalized Lundberg's coefficient and NWU distributions (see Willmot and Lin [28] for more details).

Also, we point out that $\overline{W}(x)$ cannot admit exponential asymptotic forms and exponential upper bounds if F or G is subexponential, i.e. no constant $c > 0$ and $\varepsilon > 0$ such that $\overline{W}(x) \sim ce^{-\varepsilon x}$ or $\overline{W}(x) \leq ce^{-\varepsilon x}$, for all $x \geq 0$. For example, if F is subexponential, then we know that $H \in \mathcal{S}$ and $e^{\varepsilon x}\,\overline{H}(x) \to \infty$ as $x \to \infty$, hence $\lim_{x\to\infty} e^{\varepsilon x}\,\overline{W}(x) = \infty$, since $\overline{W}(x) = \overline{H * G}(x) \geq \overline{H}(x)$.

4 Medium-tail asymptotics

We consider here the intermediate case, i.e. when $m_F(s) < \infty$ for some $s > 0$, but Lundberg's coefficient does not exist, as $m_F(s) = 1/p$ can not be satisfied but $m_F(s) < 1/p$ holds (*e.g.* in the inverse Gaussian case). First, recall the definition of the $\mathcal{S}(\alpha)$ class and its properties.

Definition 4.1. A distribution B on $[0, \infty)$ is said to belong the $\mathcal{S}(\alpha)$ class for $\alpha \geq 0$, denoted by $B \in \mathcal{S}(\alpha)$, if

(i) $\lim_{x\to\infty} \overline{B^{(2)}}(x)/\,\overline{B}(x) = 2\,m_B(\alpha) < \infty$,

(ii) $\lim_{x\to\infty} \overline{B}(x-y)/\,\overline{B}(x) = e^{\alpha y}$, for all $y \in \mathbb{R}$.

Clearly, $B \in \mathcal{S}(0) \Leftrightarrow B \in \mathcal{S}$. A class of distributions in $\mathcal{S}(\alpha)$ is the class of generalized inverse Gaussian distribution $N^{-1}(a,b,c)$ with $a < 0$, $b > 0$ and $c \geq 0$, since $N^{-1}(a,b,c) \in \mathcal{S}(c/2)$ (see Embrechts [12] for details). The following proposition recalls some properties of $\mathcal{S}(\alpha)$, which will be used here. Its proof is in Lemma 2.4 and Theorem 2.7 of Embrechts and Goldie [13] and page 268 of Klüppelberg [21] (for details see Rolski *et al.* [22]).

Proposition 4.1. Suppose that $B \in \mathcal{S}(\gamma)$.

(i) For any $\varepsilon > 0$, $e^{\gamma+\varepsilon}\,\overline{B}(x) \to \infty$ as $x \to \infty$,

(ii) If L is a distribution on $[0, \infty)$ and $\lim_{x\to\infty} \overline{L}(x)/\overline{B}(x) = c$, where $0 < c < \infty$, then $L \in \mathcal{S}(\gamma)$,

(iii) If $\gamma > 0$, B has a finite mean m, and B_1 is the ladder height distribution of B [i.e. $B_1(x) = \frac{1}{m}\int_0^x \overline{B}(y)dy$], then $B_1 \in \mathcal{S}(\gamma)$ and $\overline{B_1}(x) \sim \frac{\overline{B}(x)}{(\gamma m)}$.

Theorem 4.1. Suppose that $F \in \mathcal{S}(\gamma)$ for some $\gamma > 0$ and that $m_F(\gamma) = \int_0^\infty e^{\gamma x} dF(x) < 1/p$. If $\overline{G}(x)/\overline{F}(x) \to \alpha$ as $x \to \infty$, then $W \in \mathcal{S}(\gamma)$ and

$$\overline{W}(x) \sim \frac{q\{p\,m_G(\gamma) + \alpha[1 - p\,m_F(\gamma)]\}}{[1 - p\,m_F(\gamma)]^2}\,\overline{F}(x)\,. \tag{26}$$

Proof. Since $0 < p\,m_F(\gamma) < 1$, there exists some $\varepsilon > 0$ such that $0 < p[m_F(\gamma) + \varepsilon] < 1$, thus $\sum_{n=0}^\infty qp^n[m_F(\gamma)+\varepsilon]^n < \infty$. Hence, by Theorem 2.13 of Cline [7], we get that $H \in \mathcal{S}(\gamma)$ and

$$\overline{H}(x) \sim c\,\overline{F}(x) \tag{27}$$

where $c = \sum_{n=1}^\infty nqp^n[m_F(\gamma)]^{n-1} = pq[1 - pm_F(\gamma)]^{-2}$. Thus,

$$\frac{\overline{G}(x)}{\overline{H}(x)} = \frac{\overline{G}(x)}{\overline{F}(x)} \times \frac{\overline{F}(x)}{\overline{H}(x)} \to \frac{\alpha}{c} \quad \text{as} \quad x \to \infty\,.$$

By Theorem 1 of Cline [6], we get that

$$\lim_{x\to\infty} \frac{\overline{W}(x)}{\overline{H}(x)} = \lim_{x\to\infty} \frac{\overline{H * G}(x)}{\overline{H}(x)} = m_G(\gamma) + \frac{\alpha}{c} m_H(\gamma)\,. \tag{28}$$

But,

$$\begin{aligned} m_H(\gamma) = E(e^{\gamma S_N}) &= \sum_{n=0}^\infty qp^n E\left[e^{\gamma \sum_{i=1}^n X_i}\right] \\ &= \sum_{n=0}^\infty qp^n[m_F(\gamma)]^n = q[1 - pm_F(\gamma)]^{-1}\,. \end{aligned}$$

Hence, by (27) and (28), we get that

$$\overline{W}(x) \sim [m_G(\gamma) + \frac{\alpha}{c} m_H(\gamma)]\, \overline{H}(x) \sim [cm_G(\gamma) + \alpha m_H(\gamma)]\, \overline{F}(x)$$
$$= \{pq[1 - pm_F(\gamma)]^{-2} m_G(\gamma) + \alpha q[1 - pm_F(\gamma)]^{-1}\}\, \overline{F}(x)\ ,$$

i.e. (26) holds. $W \in \mathcal{S}(\gamma)$ follows from (26) and (ii) of Proposition 4.1. □

5 Ruin probabilities in a diffusion risk model

To illustrate a possible application of the above results, consider the classical risk process perturbed by a Wiener process, i.e. the surplus $R(t)$ at time t is

$$R(t) = x + ct - S(t) + Z(t)\ ,$$

where $x \geq 0$ is the initial risk reserve, $c > 0$ is the premium rate, $S(t)$ is the compound Poisson process representing the total claims at t [with rate $1/d > 0$, and the independent individual claim sizes with common distribution function B and $B(0) = 0$], and $\{Z(t)\}$ is a Wiener process, independent of $\{S(t)\}$, with infinitesimal drift 0 and infinitesimal variance $2D > 0$. Assume $c > \lambda/d$, where $\lambda = \int_0^\infty \overline{B}(x)\, dx$ is the expected claim size, then the relative security loading $q = 1 - \lambda/(cd)$ is such that $0 < q < 1$.

Let $\psi(x)$ denote the probability of ultimate ruin, starting with initial reserve x :

$$\psi(x) = \Pr\{\inf_{t \geq 0} R(t) < 0\}\ .$$

Denote $\varphi(x) = 1 - \psi(x)$. Dufresne and Gerber [11] (see also Veraverbeke [26]) have shown that for any $x \geq 0$,

$$\varphi(x) = \sum_{n=0}^{\infty} qp^n F^{(n)} * G(x) = H * G(x)\ ,$$

or, equivalently,

$$\psi(x) = \sum_{n=0}^{\infty} qp^n\, \overline{F^{(n)} * G}(x) = \overline{H * G}(x)\ , \tag{29}$$

where $H(x) = \sum_{n=0}^{\infty} qp^n F^{(n)}(x)$, $G(x) = 1 - e^{-cx/D}$, $F(x) = G * B_1(x)$, $B_1(x) = \frac{1}{\lambda}\int_0^x \overline{B}(y)dy$, for $x \geq 0$, and $p = \lambda/(cd)$, $q = 1 - p = \frac{1-\lambda}{(cd)}$.

Suppose that ξ and η are independent random variables with distributions G and B_1, respect., then $\xi + \eta$ has distribution $F = G * B_1$ and for any $s \in \mathbb{R}$,

$$m_F(s) = \int_0^\infty e^{sx} dF(x) = E[e^{s(\xi+\eta)}] = E(e^{s\xi})E(e^{s\eta}) = m_G(s) m_{B_1}(s). \tag{30}$$

Thus, if there exists a constant R such that

$$m_F(R) = m_G(R)\, m_{B_1}(R) = \frac{1}{p} \tag{31}$$

then (31) implies that $R < c/D$,

$$m_G(R) = E(e^{R\xi}) = \frac{c}{D}\int_0^\infty e^{Rx} e^{-\frac{cx}{D}}\, dx = \frac{c}{c - RD} < \infty\,, \tag{32}$$

and

$$E(\xi e^{R\xi}) = \frac{c}{D}\int_0^\infty x e^{Rx} e^{-\frac{cx}{D}}\, dx = \frac{cD}{(c-RD)^2} = \frac{D}{c-RD}\, m_G(R)\,. \tag{33}$$

By (31) and (32), we get that

$$E(e^{R\eta}) = m_{B_1}(R) = \frac{1}{p m_G(R)} = \frac{c - RD}{pc}\,.$$

Hence

$$\begin{aligned}\beta &= \int_0^\infty x e^{Rx} dF(x) = E[(\xi+\eta)e^{R(\xi+\eta)}] = E(\xi e^{R\xi})E(e^{R\eta}) + E(e^{R\xi})E(\eta e^{R\eta}) \\ &= m_G(R)\left[\frac{D}{pc} + \frac{1}{\lambda}\int_0^\infty x e^{Rx}\overline{B}(x)dx\right]\,. \end{aligned} \tag{34}$$

5.1 Asymptotics for the diffusion risk model

We know that if G and B_1 have density functions, so does $F = G * B_1$. This implies that F is non-lattice. Thus by Theorem 2.1, we get the following exponential formula for the ruin probability $\psi(x)$.

Corollary 5.1. Suppose that $m_F(R) = 1/p$. If $\int_0^\infty x e^{Rx}\, \overline{B}(x)\, dx < \infty$ then

$$\psi(x) \sim \frac{q m_G(R)}{pR\beta} e^{-Rx} = (1 - \frac{\alpha}{cd})\left[\frac{RD}{c} + \frac{R}{cd}\int_0^\infty x e^{Rx}\, \overline{B}(x)\, dx\right]^{-1} e^{-Rx}$$

and if $\int_0^\infty x e^{Rx}\, \overline{B}(x)\, dx = \infty$, then

$$\psi(x) = o(e^{-Rx})\,. \tag{35}$$

This Corollary includes Theorem 4.1 of Gerber [16], the results in Section 7 of Dufresne and Gerber [11] and in Section 3 of Veraverbeke [26].

Since $G(x) = 1 - e^{-cx/D}$ is an exponential distribution function, by (i) and (ii) of Lemma 3.1, we know that $B_1 \in \mathcal{S} \Leftrightarrow F = G*B_1 \in \mathcal{S}$. Furthermore, by Corollary 3.3 and Lemma 3.1, we know that

$$B_1 \in \mathcal{S} \Rightarrow \psi(x) \sim \frac{p}{q}\overline{F}(x) = \frac{\lambda}{cd-\lambda}\overline{F}(x)$$
$$\sim \frac{\lambda}{cd-\lambda}\overline{B_1}(x) = \frac{1}{cd-\lambda}\int_x^\infty \overline{B}(y)\,dy\,.$$

In addition, using the fact $\overline{F}(x) = \overline{G*B_1}(x) \geq \overline{B_1}(x)$ and following the proof of (20), it is clear that

$$\psi(x) \sim \frac{p}{q}\overline{B_1}(x) = \frac{\lambda}{cd-\lambda}\overline{B_1}(x) \Rightarrow B_1 \in \mathcal{S}\,,$$

thus, Corollary 3.3 implies the following theorem.

Theorem 5.1. For the classical risk process perturbed by a diffusion, the following conditions are equivalent:

(i) $B_1 \in \mathcal{S}$, (ii) $\varphi \in \mathcal{S}$, (iii) $\psi(x) \sim \frac{1}{cd-\lambda}\int_x^\infty \overline{B}(y)\,dy$.

Remark 5.1. Theorem 5.1 generalizes Theorem 1 of Veraverbeke [26], where it is shown that (i) and (ii) are equivalent and that either one of them implies (iii). Conditions enabling $B_1 \in \mathcal{S}$ can be found, expressed in terms of B, in Embrechts and Omey [15] and Klüppelberg [20].

Finally, for the intermediate case, suppose that for some $\gamma > 0$, $B \in \mathcal{S}(\gamma)$ and $m_F(\gamma) = m_G(\gamma)m_{B_1}(\gamma) < 1/p$, then by Proposition 4.1, we get that $B_1 \in \mathcal{S}(\gamma)$ and $\overline{B_1}(x) \sim \overline{B}(x)/(\gamma\lambda)$, in addition,

(1) $m_G(\gamma) = (c/D)\int_0^\infty e^{\gamma x}e^{-cx/D}\,dx = c/(c - D\gamma) < \infty$,

(2) $m_{B_1}(\gamma) = \int_0^\infty e^{\gamma x}\,dB_1(x) = (1/\lambda)\int_0^\infty e^{\gamma x}\,\overline{B_1}(x)\,dx < \infty$ and $\gamma < c/D$,

hence, there exists some $\varepsilon > 0$ such that $\gamma + \varepsilon < c/D$, thus $m_G(\gamma+\varepsilon) = c/[c - (\gamma+\varepsilon)D] < \infty$, and by Proposition 4.1, we get that

$$\frac{\overline{G}(x)}{\overline{B_1}(x)} = \frac{e^{\gamma+\varepsilon}\,\overline{G}(x)}{e^{\gamma+\varepsilon}\,\overline{B_1}(x)} \to 0 \quad \text{as} \quad x \to \infty\,.$$

Thus, by Theorem 1 of Cline [6] and (ii) of Proposition 4.1, we get

$$\frac{\overline{F}(x)}{\overline{B_1}(x)} = \frac{\overline{G*B_1}(x)}{\overline{B_1}(x)} \to m_G(\gamma) \quad \text{as} \quad x \to \infty$$

and

$$F \in \mathcal{S}(\gamma) \text{ and } \overline{F}(x) \sim m_G(\gamma)\,\overline{B_1}(x) \sim \frac{m_G(\gamma)}{\gamma\mu}\,\overline{B}(x)\,.$$

Thus, $\overline{G}(x) = o(\overline{F}(x))$ and Theorem 4.1 imply that $\varphi \in \mathcal{S}(\gamma)$ and

$$\psi(x) \sim \frac{pqm_G(\gamma)}{[1-pm_F(\gamma)]^2}\,\overline{F}(x) \tag{36}$$

$$\sim \frac{pq[m_G(\gamma)]^2}{\lambda\gamma[1-pm_F(\gamma)]^2}\,\overline{B}(x) \tag{37}$$

$$= \frac{1}{cd\gamma}(1-\frac{\lambda}{cd})\left[1-\frac{\gamma D}{c}-\frac{1}{cd}\int_0^\infty e^{\gamma y}\,\overline{B}(y)\,dy\right]^{-2}\overline{B}(x)\,, \tag{38}$$

thus, (38) yields Theorem 2 of Veraverbeke [26].

5.2 Ruin probability bounds for the diffusion risk model with heavy tails

Under condition (31), the exponential bounds for the ruin probability $\psi(x)$ have been derived by Dufresne and Gerber [11]. Here, we use a generalized condition of Dickson [10] to derive bounds for $\psi(x)$ with heavy claim size tails. General upper and lower bounds for $\psi(x)$ follow directly by Theorems 3.1 and 3.2 since G is exponential and has a decreasing density.

Given $t > 0$, suppose that R_t satisfies

$$m_G(R_t)\int_0^t e^{R_t\,y}\,dB_1(y) = \frac{1}{p}\,, \tag{39}$$

or equivalently,

$$\int_0^t e^{R_t\,y}\,dB_1(y) = \frac{c-R_tD}{cp}\,. \tag{40}$$

Lemma 5.1. For any claim size distribution B with $B(0) = 0$, there exists a unique solution $R_t \in (0,\, c/D)$ to equation (39).

Proof. Let $h(x) = m_G(x)\int_0^t e^{xy}dB_1(y) - \frac{1}{p}$. Since $H(0) = B_1(t) - \frac{1}{p} < 0$ and

$$\lim_{x\uparrow c/D} m_G(x) = \lim_{x\uparrow c/D}\frac{c}{c-xD} = \infty\,,$$

this implies that $\lim_{x\uparrow c/D} h(x) = \infty$. Thus, the existence of the unique root of h follows from the fact that h is continuous and strictly increasing. □

Since B_1 is continuous in here, condition (39) is clearly equivalent to

$$m_G(R_t)\int_0^\infty e^{R_t\,y}\,dB_t(y) = \frac{1}{pB_1(t)}\,, \tag{41}$$

where

$$B_t(x) = \begin{cases} B_1(x)/B_1(t) & \text{if } 0 \le x < t \\ 1 & \text{if } x \ge t \end{cases} \tag{42}$$

Let

$$\varphi_t(x) = \sum_{n=0}^{\infty} q_t\, p_t^n\, F_t^{(n)} * G(x) \tag{43}$$

or equivalently,

$$\psi_t(x) = 1 - \varphi_t(x) = \sum_{n=1}^{\infty} q_t\, p_t^n\, \overline{F^{(n)} * G}(x)\ , \tag{44}$$

where $p_t = pB_1(t)$, $q_t = 1 - p_t$ and $F_t = B_t * G$. That is to say, $\psi_t(x)$ is the ruin probability in the diffusion risk model with corresponding parameters p_t, q_t, F_t and G.

By induction we get that for any $0 \le x \le t$, $B_t^{(n)}(x) = B_1^{(n)}(x)/[B_1(t)]^n$, hence for any $0 \le x \le t$,

$$\begin{aligned} \varphi_t(x) &= \sum_{n=0}^{\infty} q_t\, p_t^n B_t^{(n)} * G^{(n)} * G(x) \\ &= \frac{q_t}{q} \sum_{n=0}^{\infty} q\, p^n F^{(n)} * G(x) = \frac{q_t}{q}\, \varphi(x)\ . \end{aligned}$$

Thus, for any $0 \le x \le t$,

$$\psi_t(x) = 1 - \varphi_t(x) = 1 - \frac{q_t}{q}\, \varphi_d(x) = 1 - \frac{q_t}{q}\, [1 - \psi_d(x)]\ .$$

This implies the following property.

Lemma 5.2. For any $0 \le x \le t$,

$$\psi(x) = \frac{p\,\overline{B}_1(t)}{q + p\,\overline{B}_1(t)} + \frac{q\,\psi_t(x)}{q + p\,\overline{B}_1(t)}\ . \tag{45}$$

Using Lemma 5.2 above, we can prove the following result.

Theorem 5.2. Suppose R_t satisfies (39), then for any $0 \le x \le t$,

$$\frac{p\,\overline{B}_1(t)}{q + p\,\overline{B}_1(t)} \le \psi(x) \le \frac{p\,\overline{B}_1(t)}{q + p\,\overline{B}_1(t)} + \frac{q\,e^{-R_t\,x}}{q + p\,\overline{B}_1(t)}\ . \tag{46}$$

In particular, for any $x > 0$,

$$\frac{p\,\overline{B}_1(x)}{q + p\,\overline{B}_1(x)} \le \psi(x) \le \frac{p\,\overline{B}_1(x)}{q + p\,\overline{B}_1(x)} + \frac{q\,e^{-R_x\,x}}{q + p\,\overline{B}_1(x)}\ . \tag{47}$$

Proof. The lower bound in (46) follows from $\psi_t(x) \geq 0$ and (45). On the other hand, if a constant R satisfies

$$m_G(R) \int_0^\infty e^{Ry}\, dB_1(y) = \frac{1}{p}\,, \tag{48}$$

then Dufresne and Gerber [16] show that for any $x \geq 0$,

$$\psi(x) \;\leq\; e^{-Rx}\,. \tag{49}$$

Thus, apply (49) to $\psi_t(x)$ with condition (41), to get for any $x \geq 0$,

$$\psi_t(x) \;\leq\; e^{-R_t x}\,.$$

This, together with (45), implies that the upper bound in (46) holds. Taking $x = t$ in (46), gives (47). □

Remark 5.2. **(a)** Since

$$\frac{p\overline{B}_1(x)}{q + p\overline{B}_1(x)} \;\sim\; \frac{p\overline{B}_1(x)}{q}\,,$$

by Theorem 5.1, we know that the lower bound in (47) is asymptotically exact, for large x, if B_1 is a subexponential distribution.

(b) If $D = 0$, the diffusion risk model is reduced to the compound Poisson risk model, thus the bound of Dickson [10] is derived as a special case of Theorem 5.2 and is improved upon.

Other applications of the results in sections 2-4 have been investigated, notably to $M/G/k$ queues (see for example Asmussen [2]).

Acknowledgments

We are grateful to the anonymous referees for their constructive comments. We would also like to acknowledge the funding obtained for this research by the Natural Sciences and Engineering Council of Canada (NSERC) operating grant OGP0036860.

References

1. S. Asmussen, *Applied Probability and Queues* (John Wiley & Sons, New York, 1987).
2. S. Asmussen, *Ruin Probabilities* (World Scientific, Singapore, 2000).
3. M. Brown, *An. Prob.* **18**, 1388 (1990).
4. J. Cai and J. Garrido, *J. App. Prob.* **36**, 1058 (1999).
5. V.P. Chistyakov, *Th. Prob. App.* **9**, 640 (1964).

6. D.B.H. Cline, *Prob. Th. Rel. Fields* **72**, 529 (1986).
7. D.B.H. Cline, *J. Aus. Math. Soc.* **A43**, 347 (1987).
8. J.W. Cohen, *On Regenerative Processes in Queueing Theory* (Springer-Verlag, Berlin, 1976).
9. F. De Vylder and M.J. Goovaerts, *Ins.: Math. Econ.* **3**, 121 (1984).
10. D.C.M. Dickson, *Scand. Act. J.* , 131 (1994).
11. F. Dufresne and H. Gerber, *Ins.: Math. Econ.*10 51 1991.
12. P. Embrechts, *J. App. Prob.* **20**, 537 (1983).
13. P. Embrechts and C.M. Goldie, *Stoc. Proc. App.* **13**, 263 (1982).
14. P. Embrechts, C.M. Goldie and N. Veraverbeke, *Z. Wahr. Verw. Geb.* **49**, 335 (1979).
15. P. Embrechts and E. Omey, *J. App. Prob.* **21**, 80 (1984).
16. H. Gerber, *Skand. Akt.* 2051970.
17. C.M. Goldie and C. Klüppelberg, in *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, ed. R. Adler, R. Feldman and M.S. Taqqu (Birkhäuser, Boston, 1998).
18. V. Kalashnikov, *Topics on Regenerative Processes*, CRC Press, Boca Raton, Florida, 1994).
19. J. Keilson, *Ann. Math. Stat.* **37**, 886 (1966).
20. C. Klüppelberg, *J. App. Prob.* **25**, 132 (1988).
21. C. Klüppelberg, *Prob. Th. Rel. Fields* **82**, 259 (1989).
22. T.Rolski, H.Schmidli, V. Schmidt and J. Teugels, *Stochastic Processes for Insurance and Finance* (John Wiley & Sons, New York, 1999).
23. B. Sundt and J.L. Teugels, *Ins.: Math. Econ.* **16**, 7 (1995).
24. R. Szekli, *Stochastic Ordering and Dependence in Applied Probability,* Lect. Notes Statist. 97 (Springer-Verlag, New York 1995).
25. M.H. van Hoorn, *Algorithms and approximations for queueing systems*, CWI Tract# 8 (CWI, Amsterdam, 1984).
26. N. Veraverbeke, *Ins.: Math. Econ.*, **13**, 57 (1993).
27. G.E. Willmot and X. Lin, *Ins.: Math. Econ.*, **18**, 29 (1996).
28. G.E. Willmot and X. Lin, *Lundberg Approximations for Compound Distributions with Insurance Applications*, Lec. Notes Stat. 156 (Springer-Verlag, New York, 2001).

IMPROVED FINITE-SAMPLE INFERENCE IN OVERIDENTIFIED MODELS WITH WEAK INSTRUMENTS

NIKOLAY GOSPODINOV

Department of Economics, Concordia University, 1455 de Maisonneuve Blvd. West Montréal, Québec, H3G 1M8 Canada

E-mail: gospodin@vax2.concordia.ca

This paper investigates the finite-sample properties of the class of generalized empirical likelihood estimators in possibly overidentified models with weakly identified parameters. These nonparametric likelihood estimators satisfy exactly the moment conditions and automatically remove the bias that arises from a lack of centering of the moment conditions. The inference procedure suggested in the paper does not involve any explicit estimation of the variance-covariance matrix. The confidence sets for the parameters of interest are constructed by inverting the χ^2 acceptance region of the criterion test.

1 Introduction

Moment condition models arise naturally from dynamic economic theory with optimizing agents. Since the seminal paper by Hansen [13], the generalized method of moments (GMM) has become the predominant framework for estimating the structural parameters of these models. Under some general regularity conditions, the GMM estimator is consistent, asymptotically normal and efficient for the given set of moment conditions. Unfortunately, it has been found that the small-sample properties of the conventional GMM estimators (in particular, the two-step GMM) are rather poor.

In this paper, we investigate the properties of the class of generalized empirical likelihood estimators of moment condition models. Members of this class are the empirical likelihood-based GMM of Qin and Lawless [27], Imbens [15] and Imbens, Spady and Jonhson [16], and the maximum entropy-based GMM of Kitamura and Stutzer [17] and Imbens, Spady and Johnson [16]. This class also includes the continuously-updated GMM as a special case (Imbens, Spady and Johnson [16]; Newey and Smith [22]) which explains the superior small-sample performance of this estimator over the traditional two-step GMM found in Hansen, Heaton and Yaron [14]. These nonparametric likelihood estimators minimize the distance between the empirical distribution function and a distribution function that exactly satisfies the moment conditions.

One of the most attractive properties of nonparametric likelihood estimators is that they tend to remove some important sources of bias that give rise to poor finite-sample properties of the GMM estimator and GMM-based

test statistics. The first source of bias arises from the fact that the first-order conditions of the standard two-step GMM estimator (Hansen [13]), evaluated at the true values of the parameters, are non-zero. This bias is exacerbated if the number of instruments increases (Kocherlakota [18]). Altonji and Segal [1] and Angrist, Imbens and Krueger [3] proposed some *ad hoc* methods for reducing the magnitude of the bias. Donald and Newey [8] showed that, for the continuously-updated GMM of Hansen, Heaton and Yaron [14], the first-order conditions are exactly satisfied at the true values of the parameters and this source of bias is automatically removed. In fact, for all members of the class of the generalized empirical estimators, the moment conditions are exactly centered at zero by construction.

Second, the estimation of the weighting matrix can be another important source of bias due to the non-zero finite sample correlation between the elements of the variance-covariance matrix of the parameters and the errors (Altonji and Segal [1]). This source of bias is present for both the two-step and continuously-updated GMM estimators but disappears for the empirical likelihood (EL) estimator of Qin and Lawless [27]. Newey and Smith [22] showed that the bias of the empirical likelihood estimator of Owen [24] and Qin and Lawless [27] is the same as the bias for the infeasible optimal GMM where the optimal linear combination coefficients do not have to be estimated.

Third, the small-sample properties of the GMM estimators and test statistics can be seriously affected by the choice of instruments that are only weakly correlated with the endogenous variables (Stock and Wright [29]). In this case, the finite sample distributions of the GMM estimators and the test statistics may depart substantially from their asymptotic distributions. Stock and Wright [29] proposed an alternative reparameterization of the moment conditions and obtained asymptotic representations with improved finite sample properties. In their framework, however, the weakly identified parameters are not consistently estimable. Fortunately, one could still conduct asymptotically valid inference by inverting criterion-based tests since their limiting χ^2-distribution at the true values of the parameters is preserved.

In this paper, we show that the nonparametric likelihood estimators are robust in the presence of weakly identified parameters. Most importantly, the criterion-based inference procedure does not involve any explicit estimation of variance-covariance matrices. Unlike the Wald test, the confidence sets constructed by inverting the criterion test, also satisfy the requirement of infinite expected volume in the completely unidentified model (Dufour [9]). Finally, the class of generalized empirical likelihood estimators is transformation invariant and the obtained confidence sets are transformation respecting.

The rest of the paper is structured as follows. Section 2 discusses two ap-

proaches to estimating moment condition models that give rise to the GMM and the nonparametric likelihood estimators. The asymptotic validity of the confidence interval construction by criterion test inversion is shown in Section 3. The Monte Carlo experiment in Section 4 studies the finite-sample properties of the different estimators and their corresponding confidence intervals in a linear instrumental variable model with weakly identified parameters. In Section 5, nonparametric likelihood estimators are applied to estimating the return to education. Section 6 summarizes the conclusions.

2 General Approach to Estimating Moment Condition Models

Let $E[g(x,\theta)|F] = \int g(x,\theta)dF = 0$ be an $m \times 1$ vector of population moment conditions implied by economic theory, where $(x_1, x_2, ...)$ are independent random vectors in $\mathbf{R}^p$ with unknown continuous distribution function F, θ is a $k \times 1$ vector of unknown parameters from Θ and $g(.)$ is a given function $\{g(x,\theta) : \mathbf{R}^p \times \mathbf{R}^k \to \mathbf{R}^m\}$ with $m \geq k$.

Suppose that we restrict the family of possible distribution functions to the space of multinomial distributions with finite support on the observed data, denoted by Φ. Also, let F_n denote the empirical measure of the sample $\{x_i\}_{i=1}^n$ from F that places probability mass n^{-1} on each data point and P_n be another probability measure that assigns multinomial weights $p_1, p_2, ..., p_n$ to each of the observations. Below, we consider two versions of the analogy principle discussed in Manski [20]. The first version selects an estimator that minimizes the distance of the moment conditions from zero (GMM estimators). The second version selects an estimator that minimizes the distance between the empirical measure and a measure P_n that satisfies exactly the moment conditions (nonparametric likelihood estimators).

2.1 GMM Estimators

The conventional GMM estimator minimizes the distance of the sample counterparts of these moment conditions from zero using the quadratic form

$$Q_n(\theta) = g_n(\theta)' W_n(\theta) g_n(\theta), \tag{1}$$

where $g_n(\theta) = E[g(x,\theta)|F_n] = \int g(x,\theta)dF_n$ and $W_n(\theta)$ is a positive definite weighting matrix. Then, $\widehat{\theta} = \arg\min_{\theta\in\Theta} Q_n(\theta)$. The properties of the GMM estimator depend crucially on the choice of the weighting matrix. The optimal GMM estimator sets $W_n(\theta) = \left[\frac{1}{n}\sum_{i=1}^n g_i(\theta)g_i(\theta)'\right]^{-1}$, where $g_i(\theta) = g(x_i,\theta)$. If a preliminary consistent (but not necessary efficient) estimator $\widetilde{\theta}$ of θ is

used in the estimation of W_n, we have the two-step GMM estimator

$$\widehat{\theta}_{2step} = \arg\min_{\theta\in\Theta} \left[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)\right]' \left[\frac{1}{n}\sum_{i=1}^{n} g_i(\widetilde{\theta})g_i(\widetilde{\theta})'\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)\right]. \quad (2)$$

If we substitute $\widehat{\theta}_{2step}$ in the weighting matrix and repeat this until convergence of both θ and W_n, the obtained estimator is the iterated GMM estimator (Hansen, Heaton and Yaron [14]) described by the equations

$$\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial g_i(\widehat{\theta}_{igmm})}{\partial\theta'}\right)\right]' \left[\frac{1}{n}\sum_{i=1}^{n} g_i(\widehat{\theta}_{igmm})g_i(\widehat{\theta}_{igmm})'\right]^{-1} g_n(\widehat{\theta}_{igmm}) = 0_k. \quad (3)$$

Finally, the continuously-updated GMM estimator proposed by Hansen, Heaton and Yaron [14] does not require a preliminary estimate of θ and directly minimizes the criterion function

$$\widehat{\theta}_{cu} = \arg\min_{\theta\in\Theta} \left[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)\right]' \left[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)g_i(\theta)'\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)\right]. \quad (4)$$

The estimator is the solution to a (typically nonlinear) system of k first-order conditions

$$\left(\frac{\partial g_n(\widehat{\theta}_{cu})}{\partial\theta'}\right)' W_n(\widehat{\theta}_{cu})g_n(\widehat{\theta}_{cu}) - g_n(\widehat{\theta}_{cu})'W_n(\widehat{\theta}_{cu})\frac{\partial W_n(\widehat{\theta}_{cu})}{\partial\theta'}W_n(\widehat{\theta}_{cu})g_n(\widehat{\theta}_{cu}) = 0$$

or

$$\left[\frac{1}{n}\sum_{i=1}^{n}\left(1+\Lambda' g_i(\widehat{\theta}_{cu})\right)\frac{\partial g_i(\widehat{\theta}_{cu})}{\partial\theta'}\right]' \left[\frac{1}{n}\sum_{i=1}^{n} g_i(\widehat{\theta}_{cu})g_i'(\widehat{\theta}_{cu})\right]^{-1} g_n(\widehat{\theta}_{cu}) = 0, \quad (5)$$

where $\Lambda = -\left[\sum_{i=1}^{n} g_i(\widehat{\theta}_{cu})g_i(\widehat{\theta}_{cu})'\right]^{-1} g_n(\widehat{\theta}_{cu})$.

Although these estimators are asymptotically equivalent, their finite sample properties may differ (see for example Hansen, Heaton and Yaron [14]).

2.2 Nonparametric Likelihood Estimators

A second approach is to obtain a value of θ that minimizes a distance between probability measures rather than the distance of the moment conditions from zero. This data driven approach selects from the set of distributions that satisfy exactly the moment conditions a probability measure P_n closest to the

empirical measure F_n defined by the Cressie and Read [6] power divergence criterion

$$D_\rho(F_n, P_n) = \frac{2}{\rho(1+\rho)} \sum_{i=1}^{n} p_i \left[(np_i)^\rho - 1\right], \tag{6}$$

where ρ is a fixed scalar parameter which determines the shape of the criterion function. Cressie and Read [6] proposed the family of power divergence statistics as goodness-of-fit tests. Here, we use the Cressie-Read divergence criterion for estimation purposes. The estimator is defined as the solution to

$$\min_{P_n \in \Phi, \theta \in \Theta} D_\rho(F_n, P_n) \tag{7}$$

$$\text{subject to } E\left[g(x,\theta)|P_n\right] = \int g(x,\theta) dP_n = 0. \tag{8}$$

This form of the analogy principle maps the empirical distribution function onto the space of feasible distribution functions and chooses the probability measure that is most likely to have generated the observed data, subject to the moment conditions (Manski [20]). The solution to the above constrained optimization problem is a straightforward application of the Lagrange multiplier principle.

This framework embeds several interesting special cases (see Kitamura and Stutzer [17]; and Imbens, Spady and Johnson [16]). The first two cases can also be interpreted as discrete versions of the forward and backward Kullback-Leibler discrepancy between the empirical measure and P_n. If we let ρ approach 0, the estimator is the solution to the problem

$$\min_{p,\theta} -\frac{2}{n} \sum_{i=1}^{n} \ln np_i \tag{9}$$

$$\text{subject to } \sum_{i=1}^{n} p_i g(x_i, \theta) = 0 \text{ and } \sum_{i=1}^{n} p_i = 1. \tag{10}$$

This is the empirical likelihood estimator of Owen [24,25,26] and Qin and Lawless [27] obtained as the root of the system of equations

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} g_i(\widehat{\theta}_{EL}) / \left(1 + \widehat{\lambda}' g_i(\widehat{\theta}_{EL})\right) \\ \\ \frac{1}{n}\sum_{i=1}^{n} \widehat{\lambda}' \left(\frac{\partial g_i(\widehat{\theta}_{EL})}{\partial \theta'}\right) / \left(1 + \widehat{\lambda}' g_i(\widehat{\theta}_{EL})\right) \end{pmatrix} = 0_{m+k},$$

where λ is a vector of Lagrange multipliers on the moment conditions.

The $(m+k) \times 1$ parameter vector $(\widehat{\theta}_{EL}, \widehat{\lambda})'$ can then be used to compute the vector of probability weights $p_i = \left[n\left(1 + \widehat{\lambda}' g_i(\widehat{\theta}_{EL})\right)\right]^{-1}$ for $i = 1, 2, ..., n$. This gives an efficient estimate of the distribution function F given the set of moment conditions. One drawback of this method is that the form of the loss function resists heavy downweighting of specific data points which may be a problem in the presence of outliers.

If $\rho \to -1$, we obtain the maximum entropy, exponential tilting or KLIC (Kullback-Leibler information criterion) estimator of Efron [10] and DiCiccio and Romano [7] and discussed in Kitamura and Stutzer [17], Imbens [15] and Imbens, Spady and Johnson [16]). It is computed as the minimizer of the KLIC function

$$\min_{p,\theta} 2 \sum_{i=1}^{n} p_i \ln np_i \tag{11}$$

subject to (10). The maximum entropy estimator solves the system of first-order conditions

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} g_i(\widehat{\theta}_{ET}) \exp\left(\widehat{\lambda}' g_i(\widehat{\theta}_{ET})\right) \\ \\ \frac{1}{n}\sum_{i=1}^{n} \widehat{\lambda}' \left(\frac{\partial g_i(\widehat{\theta}_{ET})}{\partial \theta'}\right) \exp\left(\widehat{\lambda}' g_i(\widehat{\theta}_{ET})\right) \end{pmatrix} = 0_{m+k}.$$

Finally, if $\rho \to -2$, we obtain the Euclidean likelihood estimator of Owen [26] given by the argument that minimizes

$$\min_{p,\theta} \frac{1}{n} \sum_{i=1}^{n} (n^2 p_i^2 - 1) \tag{12}$$

subject to (10). The solution is obtained from

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} g_i(\widehat{\theta}_{EU}) \left[1 + \widehat{\lambda}' \overline{g}_i(\widehat{\theta}_{EU})\right] \\ \\ \frac{1}{n}\sum_{i=1}^{n} \widehat{\lambda}' \left(\frac{\partial g_i(\widehat{\theta}_{EU})}{\partial \theta'}\right) \left[1 + \widehat{\lambda}' \overline{g}_i(\widehat{\theta}_{EU})\right] \end{pmatrix} = 0_{m+k},$$

where $\overline{g}_i(\widehat{\theta}_{EU}) = g_i(\widehat{\theta}_{EU}) - \frac{1}{n}\sum_{i=1}^{n} g_i(\widehat{\theta}_{EU})$. From the first m equations, $\widehat{\lambda} = -\left[\frac{1}{n}\sum_{i=1}^{n} \overline{g}_i(\widehat{\theta}_{EU}) \overline{g}_i(\widehat{\theta}_{EU})'\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} g_i(\widehat{\theta}_{EU})\right]$. Then, by substituting for $\widehat{\lambda}$ and $\widehat{p}_i = n^{-1}\left[1 + \widehat{\lambda}' \overline{g}_i(\widehat{\theta}_{EU})\right]$ in the last k equations, we get

$$\left[\sum_{i=1}^{n} \widehat{p}_i \left(\frac{\partial g_i(\widehat{\theta}_{EU})}{\partial \theta'}\right)\right]' \left[\frac{1}{n}\sum_{i=1}^{n} \overline{g}_i(\widehat{\theta}_{EU}) \overline{g}_i(\widehat{\theta}_{EU})'\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} g_i(\widehat{\theta}_{EU})\right] = 0_k.$$

It is interesting to see that this system of first-order conditions is almost identical to (5) for the continuously-updated GMM estimator and very similar to (3) for the iterated GMM estimator. Hence, the Euclidean likelihood estimator can be interpreted as a continuously-updated GMM estimator and an optimally weighted iterated GMM estimator.

Note also that the objective functions (9) and (11) implicitly impose the constraint $p_i \geq 0$ which validates the interpretation of p_i as probability weights. For the Euclidean estimator, we either have to inspect the positivity of the probability weights each time or introduce a nonnegativity constraint explicitly into the minimization problem. Owen [26] argues that in small samples, the negativity of the estimated weights may be advantageous for confidence interval construction.

3 Inference in Moment Condition Models

Consider the estimators discussed in Section 2. Let θ_0 denote the true value of the parameter vector and suppose that the following regularity conditions are satisfied.

Assumption A1. Assume that $W_n \xrightarrow{p} W$, *where* W *is a nonstochastic symmetric positive definite matrix;* $g(x_i, \theta)$ *is continuous in* θ; $E\left[\sup_{\theta \in \Theta} |g(x_i, \theta)|\right] < \infty$; $\sup_i E\left[g(x_i, \theta) g(x_i, \theta)'\right] < \infty$ *for all* θ *and* Θ *is a compact subset of* R^k.

Assumption A2. There is a unique θ_0 *such that* $E\left[g(x_i, \theta_0)\right] = 0$ *and* $E\left[g(x_i, \theta)\right] \neq 0$ *for all* $\theta \neq \theta_0 \in \Theta$.

Assumption A3. Assume that $M = E\left(\frac{\partial g(x_i, \theta_0)}{\partial \theta'}\right)$ *is of full rank* k; $\frac{\partial g(x_i, \theta)}{\partial \theta'}$ *is continuous in* θ *and* $E\left[\sup_{\theta \in N(\theta_0)} \left|\frac{\partial g(x_i, \theta)}{\partial \theta'}\right|\right] < \infty$ *for some neighborhood of* θ_0, $N(\theta_0)$.

Theorem 1. *Under Assumptions A1-A2,*

$$\widehat{\theta}_\rho \xrightarrow{p} \theta_0 \text{ as } n \to \infty$$

$$\sqrt{n}(\widehat{\theta}_\rho - \theta_0) \xrightarrow{d} N\left(0, \Omega\right),$$

where $\Omega = (M' V^{-1} M)^{-1}$ and $V = E\left(g(x_i, \theta_0) g(x_i, \theta_0)'\right)$.

Proof. See Imbens [15], Qin and Lawless [27], and Newey and Smith [22].

Theorem 2. *Let* $\theta = (\alpha, \beta)'$, *where* $\alpha \in \Theta_1$ *and* $\beta \in \Theta_2$ *are* $p \times 1$ *and* $(k - p) \times 1$ *vectors, respectively. Then, under Assumptions A1-A3,*

(i) *test for overidentifying restrictions*

$$nD_\rho(\widehat{\theta}_\rho) \xrightarrow{d} \chi^2_{(m-k)} \text{ as } n \to \infty$$

(ii) *test of* $H_0 : \theta = \theta_0$

$$n\left[D_\rho(\theta_0) - D_\rho(\widehat{\theta}_\rho)\right] \xrightarrow{d} \chi^2_{(k)} \text{ as } n \to \infty$$

(iii) *test for a subset* α *with* $H_0 : \alpha = \alpha_0$

$$n\left[D_\rho(\alpha_0, \widetilde{\beta}_\rho) - D_\rho(\widehat{\alpha}_\rho, \widehat{\beta}_\rho)\right] \xrightarrow{d} \chi^2_{(p)} \text{ as } n \to \infty$$

where $D_\rho(\theta_0)$ *for* $\rho = -2, -1$ *and* 0 *is the criterion function defined in (12), (11) and (9),* $\widehat{\theta}_\rho = (\widehat{\alpha}_\rho, \widehat{\beta}_\rho)$ *is the unrestricted nonparametric likelihood estimator that minimizes* $D_\rho(\theta)$ *and* $\widetilde{\beta}_\rho$ *is the minimizer of* $D_\rho(\alpha, \beta)$ *subject to* $\alpha = \alpha_0$.

Proof. See Imbens [15], Qin and Lawless [27], and Newey and Smith [22].

The results in Theorems 1 and 2 show that we can conduct asymptotically valid inference such as testing for overidentifying restrictions and constructing confidence intervals by inverting the χ^2 acceptance region of the criterion test. The $100\eta\%$ confidence set for the parameter of interest θ is then given by the set of values of θ satisfying

$$C_\eta(x) = \{\theta \in \Theta : D_\rho(\theta) \leq q_\eta\},$$

where q_η is the $100\eta^{th}$ quantile of the distribution of $D_\rho(\theta)$. Equivalently, $C_\eta(x) = \{\theta \in \Theta : x \in A(\theta)\}$, where $A(\theta)$ is the acceptance region of the test $D_\rho(\theta)$. The endpoints of the confidence set are the infimum and the supremum over $C_\eta(x)$, respectively. In particular, the two-sided, equal-tailed confidence interval with nominal coverage η is given by $C_\eta(x) = [\theta_L, \theta_U]$, where the confidence limits are defined to satisfy $\theta_L = \inf\{\theta \in \Theta : \Pr(D_\rho(\theta) \leq q_\eta | H_0) \geq \eta\}$ and $\theta_U = \sup\{\theta \in \Theta : \Pr(D_\rho(\theta) \leq q_\eta | H_0) \geq \eta\}$.

For the weak instrument case, Stock and Wright [29] parameterized the moment condition as a function of the sample size and developed an alternative limiting theory which yields a better approximation to the finite-sample distributions of the estimator and corresponding test statistics. In particular, Stock and Wright [29] replace Assumption 2 with the assumption that $E[g(x_i, \theta)] = n^{-1/2}m(\theta)$ uniformly in $\theta \in \Theta$, where $m(\theta)$ is continuous in θ and bounded on Θ with $m(\theta_0) = 0$. Under this assumption, the GMM and nonparametric likelihood estimators are no longer consistent ($\widehat{\theta}_\rho - \theta_0 = O_p(1)$)

although the χ^2 asymptotic approximation for the distributions of the test for overidentifying restrictions and a test of a subvector of weakly identifiable parameters is still valid.

4 Monte Carlo Study

The poor small sample performance of the two-step GMM estimator in linear homoscedastic instrumental variables models with weakly identified parameters has been well documented in Nelson and Startz [21], Maddala and Jeong [19], Bound, Jaeger and Baker [4] and Staiger and Stock [28] among others. In this section, we assess the robustness of the nonparametric likelihood estimators in the presence of weak instruments.

The structure of the Monte Carlo experiment is similar to the one considered by Angrist, Imbens and Krueger [3]. It is designed to study the finite sample bias of the different estimators and the size properties of hypothesis tests and the test for overidentifying restrictions with a large number of irrelevant instruments. The data are generated from the model

$$\begin{aligned} y_i &= \theta_0 + \theta_1 x_i + e_i, \\ x_i &= \gamma_0 + \textstyle\sum_{j=1}^{m-1} \gamma_j z_{ij} + u_i, \end{aligned} \tag{13}$$

where $z \sim N(0, I)$, $\begin{pmatrix} e_i \\ u_i \end{pmatrix} = chol(\Sigma)\xi_i$, $\xi_i \sim iid(\mathbf{0}, I)$, $\Sigma = \begin{pmatrix} 0.25 & 0.20 \\ 0.20 & 0.25 \end{pmatrix}$, $chol(\Sigma)$ denotes Cholesky decomposition of Σ, $\theta_0 = 0$, $\theta_1 = 1$, $\gamma_0 = 0$, $\gamma_1 = 0.15$ and $\gamma_l = 0$ for $l = 2, ..., m-1$.

The optimal two-step GMM estimator in this setting is asymptotically equivalent to the two-stage least squares (2SLS) estimator with $W_n = (Z'Z)^{-1}$

$$\widehat{\theta}_{2step} = (X'P_zX)^{-1}(X'P_zy),$$

where $X = (1, x)$, $Z = (1, z)$ and $P_z = Z(Z'Z)^{-1}Z$.

We also consider the limited information maximum likelihood (LIML) estimator which is given by

$$\widehat{\theta}_{LIML} = (X'(I - kM_z)X)^{-1}(X'(I - kM_z)y),$$

where $M_z = I - P_z$, k is the smallest characteristic root of $(\overline{Y}'\overline{Y})(\overline{Y}'M_z\overline{Y})^{-1}$ and $\overline{Y} = (y\ X)$.

The confidence intervals for the OLS, 2SLS and LIML estimators are constructed by inverting the Wald test using the χ^2 critical values. The confidence intervals for the empirical likelihood (EL) and the Euclidean likelihood

(Euclid) estimators are obtained by inverting the χ^2 acceptance regions of the corresponding criterion-based tests discussed above. The results for the exponential tilting (KLIC) estimator are very similar to the results for the EL estimator and are not reported. The number of Monte Carlo replications is 10,000.

Table 1. Monte Carlo results for model (13) with T=500 and Gaussian errors.

estimator	quantiles around $(\widehat{\theta}_1 - \theta_1)$					test for	coverage
	0.10	0.25	0.50	0.75	0.90	OIR	rate of CI
$m-k=1, \gamma_1=.15, \gamma_2=0$							
OLS	0.6991	0.7154	0.7336	0.7528	0.7695	-	0.0000
2SLS	-0.1826	-0.0693	0.0348	0.1224	0.1919	0.1147	0.8809
LIML	-0.2297	-0.1028	0.0165	0.1184	0.2045	0.1219	0.8592
EL	-0.2362	-0.1120	0.0005	0.0940	0.1655	0.1055	0.8909
Euclid	-0.2354	-0.1124	0.0011	0.0939	0.1656	0.1019	0.8927
$m-k=5, \gamma_1=.15, \gamma_2=...=\gamma_6=0$							
OLS	0.6987	0.7157	0.7339	0.7526	0.7692	-	0.0000
2SLS	-0.1098	-0.0148	0.0830	0.1640	0.2297	0.1443	0.8080
LIML	-0.2260	-0.0975	0.0193	0.1190	0.2037	0.1178	0.8617
EL	-0.2427	-0.1175	-0.0011	0.0959	0.1732	0.1118	0.8791
Euclid	-0.2402	-0.1179	-0.0022	0.0951	0.1720	0.1039	0.8830
$m-k=10, \gamma_1=.15, \gamma_2=...=\gamma_{11}=0$							
OLS	0.6985	0.7149	0.7338	0.7520	0.7690	-	0.0000
2SLS	-0.0145	0.0663	0.1484	0.2220	0.2822	0.1837	0.6287
LIML	-0.2291	-0.1020	0.0186	0.1253	0.2104	0.1132	0.8470
EL	-0.2394	-0.1154	-0.0002	0.1009	0.1769	0.1256	0.8745
Euclid	-0.2408	-0.1161	-0.0010	0.1002	0.1770	0.1059	0.8854

First, we assess the effect of increasing the number of redundant moment restrictions on the magnitude of the bias of the estimators and the size properties of the corresponding criterion tests. Tables 1 reports the 0.10, 0.25, 0.50, 0.75 and 0.90 quantiles of the distribution of $(\widehat{\theta}_1 - \theta_1)$ as well as the empirical size of the test for overidentifying restrictions (OIR) with nominal level 0.1 and the coverage properties of the 90% confidence intervals for θ_1 in a model with Gaussian errors.

Newey and Smith [22] showed that in model (13) with symmetric errors $bias(\widehat{\theta}_{LIML}) = bias(\widehat{\theta}_{EL}) = bias(\widehat{\theta}_{EU}) = -\delta/n$, where $\delta = \Omega\sigma_{ex}/\sigma_e^2$. Thus, the LIML, EL and Euclidean estimators are higher-order asymptotically

equivalent and their bias does not depend on the number of instruments. By contrast, $bias(\widehat{\theta}_{2step}) = (m-2)\delta/n$ which increases linearly with the number of instruments.

To investigate the finite-sample sensitivity of the bias of the estimators with respect to the number of irrelevant instruments, we consider three cases of overidentification: $m-1=2$ (1 overidentifying restriction), $m-1=6$ (5 overidentifying restrictions) and $m-1=11$ (10 overidentifying restrictions). Tables 1 contains the results for a sample size of 500 which is commonly encountered in economic applications.

As expected, the OLS estimator is severely upward biased. The 2SLS is slightly biased when $m-k=1$, but its bias starts to approach the bias of the OLS estimator as m increases. Also, the size properties of the test based on the 2SLS deteriorate significantly as the number of instruments gets large. The magnitude of the bias of the LIML estimator is small and insensitive to the degree of overidentification of the model which is consistent with the theoretical results.

The bias of the nonparametric likelihood methods is negligible and it is practically unchanged as the number of the overidentifying restrictions increases. The confidence intervals based on the nonparametric likelihood estimators slightly undercover with the coverage rate of the Euclidean likelihood being closest to the nominal level. Similar results were obtained for sample size $T=150$ but these results are not reported due to space limitations. It is interesting to note that the dominance of the Euclidean over the EL estimator in terms of coverage rates is more pronounced for the smaller sample size. This requires further theoretical investigation of the higher-order properties of these tests using Edgeworth expansions.

The higher-order asymptotic equivalence of the LIML, EL and Euclidean estimators derived by Newey and Smith [22] is valid only for models with symmetric errors. Newey and Smith [23] show that all members of the class of nonparametric likelihood estimators except EL have an additional bias term coming from the estimation of the variance-covariance matrix Ω. To investigate the sensitivity of the results to fat tailed and asymmetric distributions, we also report results from weakly identified models with t-distributed errors with 4 degrees of freedom and χ^2-distributed errors with 1 degree of freedom. In addition, we vary the correlation of the endogenous explanatory variable with the instruments.

The simulation results in Table 2 show that the nonparametric likelihood estimators provide reliable inference in the presence of weak instruments regardless of the distribution of the errors. Similar to the results in Table 1, the Euclidean estimator dominates in terms of coverage rate with empiri-

cal levels within 3 percentage points from the nominal level. Although the bias of the empirical and Euclidean likelihood estimators could be significant for asymmetric errors (for instance, the bottom left corner of Table 2), it is still considerably smaller than the 2SLS and LIML estimators. In summary, the nonparametric likelihood estimators appear to possess good finite-sample properties in overidentified models with weak instruments.

Table 2. Monte Carlo results for model (13) with $T = 500$ and $m - k = 1$.

estimator	$\gamma_1 = .05, \gamma_2 = .02$			$\gamma_1 = .05, \gamma_2 = .05$			$\gamma_1 = .05, \gamma_2 = .1$		
	median bias	test OIR	cov.rate of CI	median bias	test OIR	cov.rate of CI	median bias	test OIR	cov.rate of CI
$\xi_i \sim N(0, I)$									
OLS	.7909	-	.0000	.7844	-	.0000	.7621	-	.0000
2SLS	.1300	.1352	.8182	.0877	.1266	.8405	.0285	.1075	.8854
LIML	.0458	.0995	.8552	.0232	.1022	.8645	.0286	.1319	.8462
EL	.0149	.0859	.8946	.0085	.0967	.8893	-.0040	.0957	.8891
Euclid	.0140	.0852	.8946	.0077	.0963	.8903	-.0037	.0950	.8903
$\xi_i \sim t(4)$									
OLS	.7900	-	.0000	.7841	-	.0000	.7602	-	.0000
2SLS	.1332	.1390	.8190	.0799	.1261	.8479	.0333	.1124	.8870
LIML	.0443	.1006	.8502	.0184	.1108	.8715	.0337	.1396	.8344
EL	.0105	.0930	.8842	.0018	.1010	.8836	.0031	.1064	.8825
Euclid	.0083	.0890	.8943	.0009	.0957	.8924	.0026	.1016	.8907
$\xi_i \sim \chi^2(1)$									
OLS	.7957	-	.0000	.7922	-	.0000	.7795	-	.0000
2SLS	.2661	.1527	.7453	.1595	.1444	.8043	.0618	.1168	.8559
LIML	.2136	.1720	.7314	.1008	.1693	.7799	.0333	.1463	.8495
EL	.1126	.0893	.8545	.0318	.0955	.8644	.0032	.1019	.8757
Euclid	.1027	.0825	.8702	.0287	.0882	.8776	.0027	.0940	.8908

The finite-sample properties of the constructed confidence intervals for the nonparametric likelihood methods can be further improved by bootstrap methods. For the efficient bootstrap suggested by Brown and Newey [5] and Hall and Presnell [12], the data can be resampled using the implied probability weights $\widehat{p}_i$ from the estimation problem rather than the empirical measure ($p_i = n^{-1}$ for all i) as in the conventional bootstrap. Also, the asymptotic validity of the conventional bootstrap requires explicit recentering of the moment conditions (Hall and Horowitz [11]) whereas for the efficient bootstrap the moment conditions, evaluated at the true parameter vector, are centered at zero by construction.

5 Empirical Illustration: Return to Education

Estimating the return to education is of central interest to labour economists. It shows the predicted percentage increase in wage for an additional year of education. Since education is believed to be endogenous, Angrist and Krueger [2] suggested the quarter of birth as an instrument for education. However, Bound, Jaeger and Baker [4] challenged the results obtained by Angrist and Krueger [2] arguing that the two-step GMM could be severely biased in the presence of a large number of weak instruments.

Here we use the Angrist-Krueger data set which consists of a random sample from 1980 census of 329,500 men who were born between 1930 and 1939. See Angrist and Krueger [2] for a detailed description of the data and model specification. Following Angrist and Krueger [2], 30 instruments are constructed by interacting quarter and year of birth. Then we draw random subsamples of 500 observations from the original sample without replacement. The results in Table 3 are obtained from 5,000 repetitions and report the median of the parameter estimates and their corresponding standard errors.

Table 3. Estimation results for the return of education.

	OLS	2SLS	EL	Euclid
parameter estimate	0.0708	0.0714	0.0723	0.0653
standard error	0.0087	0.0437	0.0404	0.0373

The estimated return to schooling for all methods is in the range of 6.5% and 7.3%. This is a bit surprising since the weak instruments employed in the estimation are expected to produce a large upward bias in the OLS and 2SLS estimates. This does not seem to be the case and all the estimates do not appear significantly different from one another. It is also interesting to note the smaller standard errors for the nonparametric likelihood estimators compared to the two-step GMM estimator.

6 Concluding Remarks

This paper shows the robustness of nonparametric likelihood estimators of moment condition models to the presence of weak instruments and nonnormal errors. The computational procedure does not involve any explicit bias correction or estimation of variance-covariance matrices. The confidence intervals are obtained directly from the criterion function by inverting its asymptotic acceptance region.

One interesting finding that emerges from the study is the existence of noticeable differences in the coverage properties within the class of generalized empirical likelihood estimators. Since the criterion-based test statistics are asymptotically equivalent, higher-order expansions are necessary to appraise the statistical significance of these results.

Acknowledgments

I would like to thank Gordon Fisher (the Associate Editor) and an anonymous referee for helpful comments and suggestions. Financial support from the FRDP of Concordia University is gratefully acknowledged.

References

1. J.G. Altonji and L.M. Segal, *Journal of Business and Economic Statistics* **14**, 353 (1996).
2. J.D. Angrist and A.B. Krueger, *Quarterly Journal of Economics* **106**, 979 (1991).
3. J.D. Angrist, G.W. Imbens and A.B. Krueger, *Journal of Applied Econometrics* **14**, 57 (1999).
4. J. Bound, D. Jaeger and R. Baker, *Journal of American Statistical Association* **90**, 443 (1995).
5. B.W. Brown and W.K. Newey, GMM, *Efficient bootstrapping and improved inference*(Mimeo, Department of Economics, Rice University, 2001).
6. N. Cressie and T. Read, *Journal of Royal Statistical Society* **B 46**, 440 (1984).
7. T. DiCiccio and J. Romano, *International Statistical Review* **58**, 59 (1990).
8. S.G. Donald and W.K. Newey, *Economics Letters* **67**, 239 (2000).
9. J.-M. Dufour, *Econometrica* **65**, 1365 (1997).
10. B. Efron, *Canadian Journal of Statistics* **9**, 139 (1981).
11. P. Hall and J.L. Horowitz, *Econometrica* **64**, 891 (1996).
12. P. Hall and B. Presnell, *Journal of the Royal Statistical Society* **B 61**, 143 (1999).
13. L.P. Hansen, *Econometrica* **50**, 1029 (1982).
14. L.P. Hansen, J. Heaton and A. Yaron, *Journal of Business and Economic Statistics* **14**, 262 (1996).
15. G.W. Imbens, *Review of Economic Studies* **64**, 359 (1997).
16. G.W. Imbens, R.H. Spady and P. Johnson, *Econometrica* **66**, 333 (1998).

17. Y. Kitamura and M. Stutzer, *Econometrica* **65**, 861 (1997).
18. N.R. Kocherlakota, *Journal of Monetary Economics* **26**, 285 (1990).
19. G.S. Maddala and J. Jeong, *Econometrica* **60**, 181 (1992).
20. C.F. Manski, *Analog Estimation Methods in Econometrics* (Chapman and Hall, New York, 1988).
21. C.R. Nelson and R. Startz, *Econometrica* **58**, 967 (1990).
22. W.K. Newey and R.J. Smith, *Asymptotic bias and equivalence of GMM and GEL estimators* (Mimeo, Department of Economics, MIT, 2000).
23. W.K. Newey and R.J. Smith, *Higher order properties of GMM and generalized empirical likelihood estimators* (Mimeo, Department of Economics, MIT, 2001).
24. A. Owen, *Biometrika* **75**, 237 (1988).
25. A. Owen, *Annals of Statistics* **18**, 90 (1990).
26. A. Owen, *Annals of Statistics* **19**, 1725 (1991).
27. J. Qin and J. Lawless, *Annals of Statistics* **22**, 300 (1994).
28. D. Staiger and J.H. Stock, *Econometrica* **65**, 557 (1997).
29. J.H. Stock and J. Wright, *Econometrica* **68**, 1055 (2000).

PROBABILITY OF CORRECT SELECTION OF GAMMA VERSUS GE OR WEIBULL VERSUS GE BASED ON LIKELIHOOD RATIO STATISTIC

RAMESHWAR D. GUPTA
Department of Applied Statistics and Computer Science
The University of New Brunswick, Saint John
Canada, E2L 4L5
E-mail: rdg@unbsj.ca

DEBASIS KUNDU AND ANUBHAV MANGLICK
Department of Mathematics
Indian Institute of Technology Kanpur
Pin 208016 India.
E-mail: kundu@iitk.ac.in

This paper proposes the use of likelihood ratio statistic in choosing between gamma and GE models or between Weibull and GE models. Probability of correct selections are obtained using Monte Carlo simulations for various sample sizes and for various model parameters. Simulation results indicate that it is easier to distinguish between GE and Weibull models than between GE and gamma models. Two real life data sets are analyzed. Interestingly, in both cases although in the literature gamma or Weibull model was used but based on the maximum likelihood values we select GE as the 'best' fitted model among these three distributions.

1 Introduction

Recently a new two-parameter distribution named as Generalized Exponential (GE) distribution or Exponentiated Exponential distribution has been introduced and studied quite extensively by two of the authors Gupta and Kundu [6,7,8,9,10]. The GE family has the following distribution function

$$F_{GE}(x;\alpha,\lambda) = (1-e^{-\lambda x})^{\alpha}; \qquad \alpha,\lambda > 0,$$

density function

$$f_{GE}(x;\alpha,\lambda) = \alpha\lambda(1-e^{-\lambda x})^{\alpha-1}e^{-\lambda x}; \qquad \alpha,\lambda > 0, \tag{1}$$

survival function

$$S_{GE}(x;\alpha,\lambda) = 1-(1-e^{-\lambda x})^{\alpha}; \qquad \alpha,\lambda > 0,$$

and the hazard function

$$h_{GE}(x;\alpha,\lambda) = \frac{\alpha\lambda(1-e^{-\lambda x})^{\alpha-1}e^{-\lambda x}}{1-(1-e^{-\lambda x})^{\alpha}}; \qquad \alpha,\lambda > 0.$$

Here α and λ are the shape and scale parameters respectively. It is observed in Gupta and Kundu [6,7] that the two-parameter GE distribution can be used quite effectively in analyzing many lifetime data particularly in place of two-parameter gamma or two-parameter Weibull distribution. The two-parameter GE distribution can have increasing, decreasing and constant failure rates depending on the shape parameter (Gupta and Kundu [6]). It is also observed by Gupta and Kundu [9] that the behavior of the two-parameter GE distribution is very much similar to a gamma distribution in many respects and in many cases the two-parameter GE model provides a *better fit* than the two-parameter gamma model or the two-parameter Weibull model in terms of maximum likelihood or minimum chi-square. Therefore, it is quite important to choose between a Weibull and GE models or between a gamma and GE models to analyze real life data sets. It is found to be very difficult to discriminate between these three models because all these three models are quite flexible and they overlap with each other in the sense that exponential distribution is a special case to all of them. Although, these three models may provide similar data fit for moderate sample sizes but it is still desirable to select the correct or more nearly correct model, since the inferences based on the model will often involve tail probabilities where the affect of the model assumption will be more critical. Therefore, even if large sample sizes are not available, it is still important to make best possible decision based on whatever data are available.

In this paper we propose to use the logarithm of the ratio of the maximum likelihood functions in choosing two overlapping distributions. The idea was originally proposed by Cox [3,4] in discriminating between two models and Bain and Englehardt [1] used it in choosing between gamma and Weibull models. It is observed that the probability of correct selection (PCS) depends only on the shape parameter of the distribution from which the data are coming. Since it is not possible to obtain the exact distributions of the likelihood ratio statistics, we obtain PCSs by using extensive Monte Carlo simulations for various sample sizes and for various model parameters. We use two real data sets to illustrate how the proposed methods can be used in practice.

Rest of the paper is organized as follows. In section 2, we provide the selection method based on likelihood ratio statistic. The PCSs are presented in section 3. We analyze two real data sets in section 4 and finally we conclude the paper in section 5.

2 Likelihood Ratio Statistic

Suppose $X_1, \ldots, X_n$ are independent and identically distributed random variables from any one of the three distributions. The density function of a GE random variable with the shape parameter α and the scale parameter λ will be denoted by (1). For both gamma and Weibull distributions we denote the shape parameter by β and the scale parameter by θ. The density function of a gamma random variable with the shape parameter β and the scale parameter θ will be denoted by $f_{GA}(x)$ as

$$f_{GA}(x) = \frac{\theta^\beta}{\Gamma(\beta)} x^{\beta-1} e^{-\theta x}; \qquad x, \beta, \theta > 0, \tag{2}$$

and similarly the density function of a Weibull random variable with the shape parameter parameter β and the scale parameter θ will be denoted by

$$f_{WE}(x) = \beta \theta^\beta x^{\beta-1} e^{-(x\theta)^\beta}.$$

Let us define

$$L_{GE}(\alpha, \lambda) = \prod_{i=1}^{n} f_{GE}(x_i), \qquad L_{GA}(\beta, \theta) = \prod_{i=1}^{n} f_{GA}(x_i)$$

and

$$L_{WE}(\beta, \theta) = \prod_{i=1}^{n} f_{WE}(x_i).$$

Now first consider choosing between gamma and GE models. The natural logarithm of the likelihood ratio statistic $T_1 = \ln(L_1)$, where

$$L_1 = \frac{L_{GE}(\hat{\alpha}, \hat{\lambda})}{L_{GA}(\hat{\beta}, \hat{\theta})}$$

and

$$T_1 = n \left[\ln(\hat{\alpha}\hat{\lambda}\bar{X}) - \hat{\beta} \ln(\tilde{X}\hat{\theta}) - \frac{(\hat{\alpha} - 1)}{\hat{\alpha}} + \ln(\Gamma(\hat{\beta})) - \bar{X}(\hat{\lambda} - \hat{\theta}) \right]. \tag{3}$$

Here $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, $\tilde{X} = (\prod_{i=1}^{n} X_i)^{\frac{1}{n}}$. Moreover, $\hat{\alpha}$ and $\hat{\lambda}$ are maximum likelihood estimators (MLEs) of α and λ if the data are assumed to come from a GE distribution and in this case (Gupta and Kundu [7]) they have the following relation

$$\hat{\alpha} = -\frac{n}{\sum_{i=1}^{n} \ln(1 - e^{-\hat{\lambda} X_i})}.$$

Similarly $\hat{\beta}$ and $\hat{\theta}$ are MLEs of β and θ if the data are assumed to come from a gamma distribution and in this case they satisfy the following equation

$$\hat{\theta} = \frac{\hat{\beta}}{\bar{X}}.$$

The proposed procedure is as follows. Choose a GE model if $T_1 > 0$, otherwise choose a gamma model. It is clear from (3) that if the data are coming from a GE population, the distribution of T_1 is free of λ and depends only on α and similarly if the data are coming from a gamma population, the distribution of T_1 is independent of θ and it depends only on β. The above assertions can be proved very easily. First let us assume that the data are coming from $GE(\alpha, \lambda)$. It implies that the distribution of λX_i is independent of λ. Moreover it follows from the Theorem 7.8.5 of Bain and Englehardt [2] that the distributions of $\frac{\hat{\lambda}}{\lambda}$ and $\frac{\hat{\theta}}{\lambda}$ are independent of λ. Since the distributions of $\hat{\alpha}$ and $\hat{\beta}$ are independent of λ, therefore the result follows immediately from the expression of T_1. Similar argument can be given when the data are coming from a gamma distribution then the distribution of T_1 is independent of θ and depends only on β. It implies that in the first case PCS is free from λ and depends only on α, whereas in the second case PCS is free from θ and depends only on β.

Now we consider the likelihood ratio statistic for discriminating between GE and Weibull models. Similarly as before, the natural logarithm of the likelihood ratio $T_2 = \ln(L_2)$, where

$$L_2 = \frac{L_{GE}(\hat{\alpha}, \hat{\lambda})}{L_{WE}(\hat{\beta}, \hat{\theta})}$$

and

$$T_2 = n\left[\ln\left(\frac{\hat{\alpha}\hat{\lambda}\tilde{X}}{\hat{\beta}}\right) - \frac{\hat{\alpha}-1}{\hat{\alpha}} - \hat{\lambda}\bar{X} - \hat{\beta}\ln(\hat{\theta}\tilde{X}) + 1\right]. \tag{4}$$

Here $\bar{X}$ and $\tilde{X}$ are same as defined before and $\hat{\alpha}$ and $\hat{\lambda}$ are MLEs of α and λ for GE distribution and similarly $\hat{\beta}$ and $\hat{\theta}$ are MLEs of β and θ for Weibull distribution. In the case of Weibull distribution, $\hat{\theta}$ and $\hat{\beta}$ satisfy the following relation

$$\hat{\theta} = \left(\frac{n}{\sum_{i=1}^{n} X_i^{\hat{\beta}}}\right)^{\frac{1}{\hat{\beta}}}.$$

We use the similar discrimination procedure as before, *i.e.* if $T_2 > 0$, choose GE distribution, otherwise choose Weibull distribution. In this case also it

Table 1. Probability of correct selection between GE and gamma distributions and when the data are generated from a GE distribution

$n \downarrow \alpha \rightarrow$	0.50	0.75	1.0	2.0	4.0	8.0	16.0
20	0.6610	0.5798	0.4942	0.4633	0.5081	0.5503	0.6012
40	0.6342	0.5870	0.5004	0.4812	0.5563	0.6155	0.6566
60	0.6282	0.5847	0.5039	0.5100	0.5841	0.6472	0.7031
80	0.6134	0.5822	0.5047	0.5194	0.6059	0.6740	0.7421
100	0.6134	0.5813	0.5059	0.5321	0.6242	0.7045	0.7641
200	0.6033	0.5760	0.4945	0.5699	0.6947	0.7860	0.8601

can be shown as before that if the data come from a GE population then the distribution of T_2 depends on α and independent of λ and if the data come from a Weibull population, then the distribution of T_2 depends only on β and independent of θ. It indicates that in this case also the PCS is independent of the corresponding scale parameter. It is difficult to compute the exact distributions of T_1 and T_2, therefore, we use Monte Carlo simulations to compute the probability of correct selections in the next section.

3 Probability of Correct Selections

In this section we use Monte Carlo simulations to compute the PCSs for different shape parameters and for different sample sizes. We consider different shape parameters namely 0.50, 0.75, 1.0, 2.0, 4.0, 8.0, 16.0 and also different sample sizes namely 20, 40, 60, 80, 100, 200. All the probabilities are calculated based on 10,000 replications. The exact details are provided below:

Case 1: Between GE and Gamma

First we generate a sample of size n from a $GE(\alpha, 1)$. From the given sample we obtain MLEs of α and λ (GE parameters), similarly we obtain MLEs of β and θ (gamma parameters). We compute T_1 and observe whether it is negative or positive. We replicate the process 10,000 times and compute the percentage of times T_1 is positive and that provides the probability of the correct selection. The same way we estimate the PCS between gamma and GE models when the data are coming from a Gamma(β, 1). We generate the sample from a Gamma(β, 1) and obtain the MLEs of α, λ, β and θ. In this case we compute the percentage of times T_1 is negative out of 10,000 replications. Results are reported in Tables 1 and 2.

Table 2. Probability of correct selection between GE and gamma distributions when the data are generated from a gamma distribution

$n \downarrow \beta \rightarrow$	0.50	0.75	1.0	2.0	4.0	8.0	16.0
20	0.3655	0.4246	0.4977	0.5829	0.6104	0.6347	0.6653
40	0.4086	0.4317	0.5104	0.5948	0.6281	0.6900	0.7365
60	0.4370	0.4398	0.4999	0.5942	0.6607	0.7197	0.7849
80	0.4555	0.4498	0.4964	0.6051	0.6745	0.7544	0.8276
100	0.4519	0.4597	0.4998	0.6081	0.6943	0.7775	0.8542
200	0.5039	0.4855	0.4991	0.6319	0.7534	0.8598	0.9317

Table 3. Probability of correct selection between GE and Weibull distributions when the data are Generated from a GE distribution.

$n \downarrow \alpha \rightarrow$	0.50	0.75	1.0	2.0	4.0	8.0	16.0
20	0.6610	0.5643	0.5025	0.5731	0.6675	0.7283	0.7699
40	0.7258	0.5982	0.4976	0.6316	0.7520	0.8357	0.8811
60	0.7663	0.6157	0.4849	0.6829	0.8154	0.8929	0.9301
80	0.7856	0.6309	0.5042	0.7139	0.8582	0.9224	0.9582
100	0.8128	0.6430	0.5097	0.7382	0.8825	0.9496	0.9752
200	0.8849	0.7012	0.4933	0.8295	0.9614	0.9911	0.9981

Case 2: Between GE and Weibull

In this case first we compute the PCS between GE and Weibull distributions and when the data are coming from a GE distribution. We generate a sample of size n from a $GE(\alpha, 1)$ and compute MLEs of α, λ (GE parameters) and β, θ (Weibull parameters). Replicate the process 10,000 times and observe the percentage of times T_2 is positive. Exactly the same way, we compute the PCSs between GE and Weibull distributions and when the data are coming from a Weibull distribution. The results are reported in Tables 3 and 4 [a].

Some of the points are quite clear from the Tables 1, 2, 3 and 4. First of all in all four cases when $\alpha = 1.0$, the PCS is close to .5 as it should be for all sample sizes. Because when $\alpha = 1.0$, all three distributions become exponential distribution therefore, between any two distributions the probability of choosing any particular one is 0.5. It is also observed that in all cases (except for $\alpha = 0.50$ or 0.75 in Table 1) as sample size n increases the PCS increases

[a] When $\beta = 4$, the Weibull distribution becomes almost symmetric. In this case it is too time consuming to compute MLEs of the GE parameters. In Table 4, for $\beta = 4$, the results are based on 1000 replications.

Table 4. Probability of correct selection between GE and Weibull distributions when the data are generated from a Weibull distribution.

$n \downarrow \beta \rightarrow$	0.50	0.75	1.0	2.0	4.0	8.0	16.0
20	0.6655	0.5155	0.5037	0.7061	0.8142	0.8738	1.0000
40	0.7159	0.5844	0.5045	0.7823	0.9185	0.9624	1.0000
60	0.7768	0.6169	0.5046	0.8395	0.9573	0.9868	1.0000
80	0.8233	0.6457	0.5069	0.8757	0.9768	0.9947	1.0000
100	0.8653	0.6700	0.5047	0.9006	0.9874	0.9979	1.0000
200	0.9492	0.7536	0.5085	0.9660	0.9992	1.0000	1.0000

for all values of $\alpha \neq 1$. Moreover as $|\alpha - 1|$ increases the PCS increases from 0.5 in most of the cases. Surprisingly in Table 1, when the sample size 20 or 40 and the $\alpha = 2.0$, the PCSs are less than .5. It indicates that when the data come from GE and the shape parameter is greater than one but not very far from one then the likelihood ratio statistic can not distinguish properly whether the data are coming from a GE distribution or from a gamma distribution if the sample size is not very large. Similar pattern is also observed in Table 2, when the shape parameter is 0.50 or 0.75. It indicates that for certain ranges of the shape parameter it is really very difficult to distinguish between gamma and GE distributions. In Table 1, when $\alpha = 0.50$ or 0.75, the PCSs decrease as n increases. In Table 2, when $\alpha = 0.50$ and 0.75, the PCS are less than 0.5. Both these findings are quite counter intuitive and we can not justify them. We feel it might be due to the estimation procedures. It is known that when the shape parameters is less than one then the parameter estimations are quite difficult for both GE and gamma distributions. The likelihood functions are very flat and therefore the maximum likelihood estimators are quite unstable. The results might be the reflection of that. One of the referees has mentioned that it might be due to discretization of the cumulative distribution function.

Now comparing the results of Tables 1 and 2, it clear that the PCSs of Table 2 are more than the corresponding PCSs of Table 1 when the shape parameter is more than one. It clearly indicates that for the shape parameter greater than one, the likelihood ratio statistic can distinguish better between gamma and GE distributions if the data are coming from a gamma distribution than vice verse. It can be justified as follows. As the shape parameter increases the gamma density becomes symmetric but the GE density remains skewed as the shape parameter goes to infinity. Therefore, if the data are coming from a gamma distribution and the shape parameter is large then

naturally PCS becomes higher.

Comparing the results between Tables 3 and 4, it is also observed that PCSs of Table 4 are more than the corresponding PCSs of Table 3. Therefore, between GE and Weibull distributions also the likelihood ratio statistic can distinguish better if the data are coming from a Weibull distribution than the other way. The same reason as above holds here also. Finally comparing the results between Tables 1, 2 and Tables 3, 4, it is clear that using the likelihood ratio statistic it is much easier to distinguish between Weibull and GE models than between gamma and GE models. It might be due to the fact that the Weibull density goes to symmetry much faster than the gamma density as the shape parameter increases. We believe that it should be true for any other statistics also and more work is needed in this direction.

4 Data Analysis

In this section we apply the above procedure in two data sets. Both these data sets have been used by several authors in the literature.

Data Set 1: Lieblein and Zelen [11] provided the following sample of size $n = 23$ to illustrate the use of Weibull model. These data indicate the endurance in millions of revolutions of deep-groove ball bearings. Thoman, Bain and Antle [12] mentioned that Weibull distribution should be used where as Bain and Englehardt [1] proposed to use gamma distribution. The data are as follows:
17.88, 28.92, 33.00, 41.52, 42.12, 45.60, 48.80, 51.84, 51.96, 54.12, 55.56, 67.80, 68.64, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40.

For this data $\bar{X} = 72.12$ and $\tilde{X} = 63.46$. The MLEs of the GE parameters are $\hat{\alpha} = 5.2825$, $\hat{\lambda} = 0.0323$, gamma parameters are $\hat{\beta} = 4.0309$, $\hat{\theta} = 0.0558$ and Weibull parameters are $\hat{\beta} = 2.1033$, $\hat{\theta} = 0.0122$. Also $ln(L_{GE}(\hat{\alpha}, \hat{\lambda})$ = -112.9762, $ln(L_{GA}(\hat{\beta}, \hat{\theta})$ = -113.0272 and $ln(L_{WE}(\hat{\beta}, \hat{\theta})$ = -113.6886. Therefore, T_1 = 113.0272-112.9762 = .0510 and T_2 = 113.6886-112.9762 = 0.7124. Since both T_1 and T_2 are positive, therefore GE is preferable than gamma or Weibull distributions based on the likelihood ratio statistic. Moreover, from the tables it can be said that the PCS between gamma and GE is between 51% to 55% and the PCS between Weibull and GE is between 67%-73%.

Data Set 2: The following data represents the relief times of 20 patients receiving an analgesic. The data is obtained from Gross and Clark [5] (page; 105), where they fitted gamma distribution. The data are as follows:
1.1, 1.4, 1.3, 1.7, 1.9, 1.8, 1.6, 2.2, 1.7, 2.7, 4.1, 1.8, 1.5, 1.2, 1.4, 3.0, 1.7, 2.3, 1.6, 2.0.

For this data set $\bar{X} = 1.90$ and $\tilde{X} = 1.80$. The MLEs of the GE parameters are as follows: $\hat{\alpha} = 36.6437$, $\hat{\lambda} = 2.2348$ and $\ln(L_{GE}(\hat{\alpha}, \hat{\lambda}))$ = -16.2605.

The MLEs of the gamma parameters are $\hat{\beta} = 9.6630$ and $\hat{\theta} = 5.0864$ and $\ln(L_{GA}(\hat{\beta}, \hat{\theta})) = -17.8186$. Similarly we obtain the MLEs of the Weibull parameters as $\hat{\beta} = 2.7875$ and $\hat{\theta} = 0.4695$ and $\ln(L_{WE}(\hat{\beta}, \hat{\theta})) = -20.5864$. Therefore, $T_1 = -16.2605 + 17.8186 = 1.5581$ and $T_2 = -16.2605 + 20.5864 = 4.3259$. Since T_1 and T_2 both are positive therefore in both cases we prefer GE models. The PCS between gamma and GE is more than 60% and the PCS between Weibull and GE is more than 77%.

5 Conclusions

In this paper we consider the likelihood ratio statistics to choose between GE and gamma model or between GE and Weibull models. It is very easy to use in practice. We could not obtain the exact distributions of the likelihood ratio statistics so we use Monte Carlo simulations to compute the probability of correct selections. The likelihood ratio statistics depend only on the shape parameters. It is observed that as the shape parameter deviates from one it becomes easier usually to distinguish between two distributions. It is also observed that it is much easier to distinguish between GE and Weibull distributions than between GE and gamma distributions. Two data sets are analyzed. Interestingly, in both cases although in the literature gamma or Weibull model was used but we select GE as the 'best' fitted model based on the maximum likelihood values. Another interesting question is to determine the minimum sample size n to discriminate between two distribution functions (either GE and gamma or GE and Weibull) for a given PCS. Work is in progress in that direction and it will be published elsewhere.

Acknowledgments

Part of the work of the first author was supported by a grant from the Natural Sciences and Engineering Research Council. The authors would like to thank two anonymous referees for some constructive suggestions.

References

1. L.J. Bain and M. Englehardt, *Communications in Statistics Theory and Methods* **9**, 375 (1980).
2. L.J. Bain and M. Englehardt, *Statistical Analysis of Reliability and Life-testing Models; Theory and Methods* (Marcel and Dekker Inc., New York, 1991)

3. D.R. Cox, *Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability* **5**, 105 (1961).
4. D.R. Cox, *Journal of the Royal Statistical Society Ser. B* **24**, 406 (1962).
5. A.J. Gross and V.A. Clark *Survival Distributions; Reliability Applications in the Biomedical Sciences* (John Wiley and Sons, New York, 1975).
6. R.D. Gupta and D. Kundu, *Australian and New Zealand Journal of Statistics* **41**, 173 (1999).
7. R.D. Gupta and D. Kundu, *Biometrical Journal* **43**, 117 (2001).
8. R.D. Gupta and D. Kundu, Jour. Stat. Comp. Simul **69**, 315 (2001).
9. R.D. Gupta and D. Kundu, *Closeness of gamma and generalized exponential distributions* (Technical Report, Department of Mathematics, Statistics and Computer Science, The University of New Brunswick, New Brunswick, Canada, 2001).
10. R.D. Gupta and D. Kundu, *Jour. Appl. Stat. Sc.* (Accepted, 2002).
11. J. Lieblein and M. Zelen, *Jour. Res. Nat. Bur. Stand.* **47**, 273 (1956).
12. D.R. Thoman, L.J. Bain and C.J. Antle, *Technometrics* **11**, 445 (1969).

SURVEY METHODOLOGY FOR STUDYING SUBSTANCE USE PREVENTION PROGRAMS IN SCHOOLS

SHELTON M. JONES, BRIAN C. SUTTON AND KERRIE E. BOYLE
RTI International, P. O. Box 12194,
Research Triangle Park, NC 27709-2194, USA
E-mail: smj@rti.org; suttonb@pharmaresearch.com; keb@rti.org

According to the 1998 National Household Survey On Drug Abuse, almost 10 percent of youths in the United States between ages 12-17 used illicit drugs SAMSHA (Substance Abuse and Mental Health Services Administration [11]). In addition, 18 percent of 12-17 year olds were current cigarette users, over eight percent were current users of marijuana, and almost one percent of youths were reported to be current users of cocaine. These percentages demonstrate a need for continuous substance use prevention programs. Although such programs have existed in U. S. school systems for some time, their extent and prevalence have been unclear (Ringwalt *et al.* [10]; Kann *et al.* [7]). The School-Based Substance Use Prevention Programs Study (SSUPPS) is designed to study substance use prevention activities currently available in middle schools. In addition to a school sample component, SSUPPS includes a public school district component for those districts associated with the public school sample. This paper presents survey design methodology for obtaining statistically valid national estimates of substance use prevention programs among the conventional public and private middle schools in the United States. Also, presented is information on sample selection and sample weighting procedures.

1 Introduction

Most researchers tend to agree that the best way to prevent substance use among adolescents is to reduce demand (Dusenbury and Falco [3]; Eigen and Rowden [4]). Substance abuse prevention programs for many adolescents are sponsored and maintained by schools with Federal assistance from the United States Safe and Drug-Free Schools and Communities Act (SDFSCA). These programs vary widely, and include both classroom curricula and non-classroom activities. The School-Based Substance Use Prevention Programs Study (SSUPPS) is designed (1) to determine if these prevention programs differ across schools with middle school grades by school type, enrollment size, urbanicity, and students socio-economic status, and (2) to determine if the established programs reflect evidence of effectiveness provided by prevention research. This paper presents the sample design established to provide descriptive information about the prevalence and characteristics of school-based programs. Probability sampling of middle schools in the U.S. is the basis for our stratified systematic sampling design (see Cochran [1]) of both public and

private schools. For the list of acronyms used, please see the appendix.

2 Defining the Target Population

SSUPPS is a survey of conventional schools with middle school grades and public school districts for the 1998/1999 academic year. Middle schools are defined as conventional schools with one of the following grade combinations: (1) only fifth and sixth grades, (2) only a sixth grade, or (3) a seventh or an eighth grade. Specifically, this study provides information on substance use prevention programs available for students approximately 10-13 years of age. Excluded from the target population are non-conventional schools or educational units (such as special education schools, alternative education schools, public middle schools with fewer than 20 students enrolled, State Department of Education units, charter schools, Bureau of Native-American Affairs units, adult education schools, Department of Defense units, vocational technical schools, and middle schools with no eligible student enrollment for the 1998/1999 academic year). A public school district is an educational unit of authority that operates one or more public schools.

3 Sampling Frame Considerations

A sampling frame is defined as the list or mechanism used to identify population elements for the selection of a sample. There are three conventional sources of sampling frames for school data. The Common Core of Data or CCD (U.S. Department of Education [15]) is the primary public school data source of the National Center for Education Statistics (NCES). Two other sources are owned by Market Data Retrieval or MDR (Dun & Bradstreet Corporation [2]) and QED (Quality Education Data, Inc.[9]). The MDR includes all educational levels from preschool through college in the United States and Canada. Both the CCD and the MDR have proven to be reliable sources of information on public school data (Hamann [5]). A single data source was preferred to reduce school multiplicity and therefore eliminate the need to merge multiple files to capture the entire target population. The QED includes information on urbanicity and poverty that can be beneficial for stratification. In addition, QED provides educational data useful for all types of schools: public, Catholic, and non-Catholic schools from pre-kindergarten through 12th grade. Therefore, due to the need to survey both public and private schools, as well as the need for stratification, the QED database was chosen as the sampling frame. It should be noted that since the sampling frame for this study was constructed, there have been important developments in school

lists. NCES now maintains, in addition to the CCD, a database of non-public schools by means of the Private School Universe Survey (PSS). Both CCD and PSS are accessible on the NCES website (http://nces.ed.gov). Therefore, if this research was conducted in the future, we would recommend a combination of CCD and PSS. Once finalized, the sampling frame contained a total of 37,841 eligible conventional middle schools (because the classifications of ineligibles are not mutually exclusive, the ineligible counts are not additive). This frame comprised 22,891 public middle schools, 6,031 Catholic schools, and 8,919 non-Catholic private middle schools (Table 1).

3.1 Explicit and Implicit Stratification

The sampling frame was stratified explicitly and implicitly to control the distribution of the sample and to reduce sampling variation among survey statistics (see Kish [8], pp. 75-112). Explicit stratification was used to provide domain estimation for public schools at different levels of urbanicity, school size, and poverty index. Urbanicity was defined at three levels provided by QED: urban (area within the central city), suburban (area surrounding the central city), and rural (area outside any metropolitan area). Three levels of size were defined based on the estimated total number of eligible students enrolled in middle school grades: small (fewer than 200 eligible students), medium (200-600 eligible students), and large (more than 600 eligible students). The poverty index for public schools was determined from the Orshansky percentage, which is the percent of students who fall below the federal governments poverty guidelines. The Orshansky percentage is defined at the district level and is a relative indicator of community wealth/poverty when compared to other school districts. Three poverty index levels were formed: low (schools with an Orshansky percentage 10% or less), medium (11-40 %), and high (greater than 40 %). The various combinations of strata for public schools formed a total of 27 levels of explicit stratification (Table 1). Due to the small population of non-public schools and because of secondary interest, non-public schools were only stratified by Catholic and non-Catholic to assure that the non-public school population received some representation. Additionally, state identification was used to implicitly stratify the public and non-public school sampling frames.

4 Sample Selection

A school sample component was designed to survey a probability sample of teachers responsible for teaching substance use prevention programs at the

school. In addition, a separate survey was conducted of Safe and Drug-Free School District Coordinators. The district survey was conducted only among those districts that oversaw schools in the sample. The sample design consisted of selecting a stratified systematic sample of middle schools within each explicit design stratum as described by Cochran [1](pp. 226). A sample of schools or teachers and district coordinators was selected to participate in a mail survey on substance use prevention programs at the school or district. The sample design is illustrated in Figure 1.

4.1 The Middle School Sample

The initial middle school sample of 3,430 schools accommodated two data collection waves. Wave 1 consisted of a random subsample of 2,852 school selections for possible participation in the study. The remaining schools comprising Wave 2 were not surveyed due to time constraints relative to data collection. Although strict proportional allocation was not used, adequate representation in each stratum made it possible for the public school sample to cover a broad range of different types of middle schools. The non-public school sample of 472 schools was proportionally allocated to Catholic and other private schools for a sample of 189 Catholics and 283 non-Catholics.

4.2 Under-Coverage or Under-Representation

When the frame was constructed, QED was one of the most reliable data sources for public and non-public schools. Public schools are updated four times per year, and both public and non-public schools are re-verified every two years. However, the list of non-Catholic private schools is not as complete as that of public schools. This is primarily because of the difficulty in securing information on non-Catholic private schools, since many tend to be small and difficult to track from year to year. Because of this limitation, non-Catholic private schools are under-represented beyond the under-sampling of non-public schools mentioned above.

4.3 Inclusion of Public School Districts

This study primarily surveyed teachers who had knowledge of substance use prevention programs within schools with middle school grades. Due to secondary interest in district level inference, the survey design was expanded to accommodate a comprehensive survey of public school districts. Each district was included in the sample based on the number of schools associated with the district and selected from the stratum.

5 Sample Weighting and Results

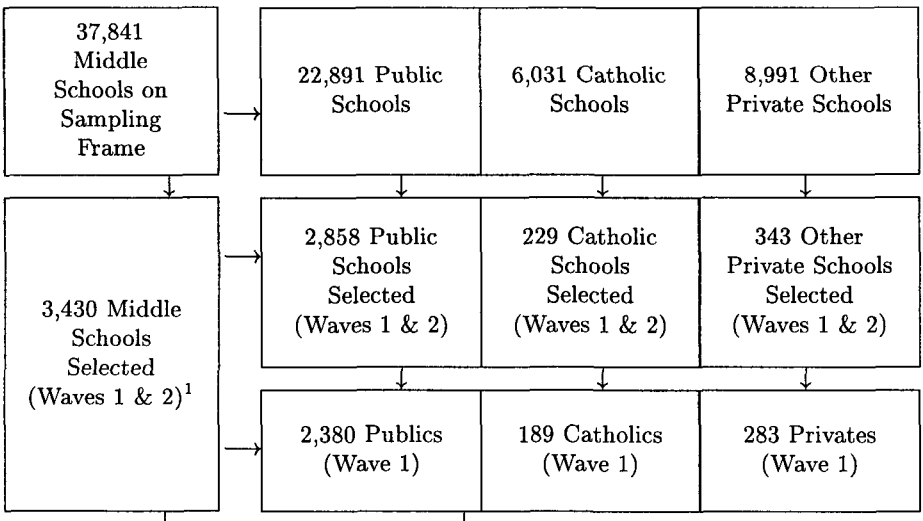
Sample weights for schools were computed as the inverse of the school sample selection probabilities. The district weights were determined as a function of a multiplicity adjustment estimator (Sirken [14]). Some sample schools and district coordinators refused to participate in the study. This type of nonresponse is referred to as unit nonresponse because information is missing for the entire sampling unit. The initial sampling weights of the responding units were adjusted to compensate for the missing data arising from the nonrespondents. A common weighting procedure referred to as sample-based adjustment cell weighting is described by Jones and Chromy [6]. With the school sample, adjustment cell weighting was employed to partition the school respondents into adjustment cells or weighting classes. Within each class or stratum, responding teachers were weighted up in an attempt to compensate for the nonresponding teachers within each stratum. These respondents weights were adjusted by the sum of the stratum weights for the respondents plus the nonrespondents divided by the sum of the respondents weights. The nonresponse adjusted weights are recommended for use in all weighted statistical analyses with software packages such as, SUDAAN (Shah, Barnwell, and Bieler [13]) and SAS (SAS Institute Inc. [12]).

5.1 School Weights

For our design, each school within each explicit stratum was selected with the same probability. Let N_h denote the number of schools on the frame within stratum h, and n_h denote the number of schools selected within stratum h ($h = 1, 2, ..., 29$) based on frame information, which occasionally is not accurate.

It follows that the school sample selection probability (π_{hi}) and the initial sampling weight (w_{hi}) for school i within stratum h are respectively given by: $\pi_{hi} = n_h/N_h$ and $w_{hi} = N_h/n_h$.Due to non-response, not all schools selected for study agreed to participate. The initial sampling weights of responding teachers were modified as follows.

Let n_h^e denote the actual number of eligible schools determined during data collection and n_h^r denote the number of responding schools within stratum h. The weighting class adjustment (λ_h) is defined as: $\lambda_h = n_h^e/n_h^r$. Hence, the teacher non-response adjusted sampling weight (w_{hi}^a) for school i in stratum h is: $w_{hi}^a = w_{hi} * \lambda_h$.



Metro Type	Orshansky Score[2]	Eligible Grade Size[3] Total	Small	Medium	Large
Urban	Low	82	10	33	39
	Medium	352	39	134	178
	High	222	20	83	119
	Total	655	69	250	336
Subarb-an	Low	332	42	152	138
	Medium	442	70	202	170
	High	167	35	75	57
	Total	941	147	429	365
Rural	Low	139	68	57	14
	Medium	468	322	125	21
	High	177	130	41	6
	Total	784	520	223	41
Total	Low	553	120	242	191
	Medium	1,261	431	461	369
	High	566	185	199	182
	Total	2,380	736	902	742

[1] The initial sample included a subsample of 2,852 schools in Wave 1 and 578 schools in Wave 2. Wave 1 was surveyed, but due to time constraints, wave 2 was not.

[2] The Orshansky scores are categorized as: Low - less than or equal to 10%, Medium - 11 to 40%, and High - 41% and higher.

[3] Eligible grade sizes are estimates of the total number of students enrolled at the school for all Middle School specific grades (5-8). The size distinctions are: Small - fewer than 200 students, Medium - 200 to fewer than 600 students, and Large - 600 and more students.

Figure 1. The SSUPPS Sample

5.2 *District Weights*

Since schools were sampled within strata based on school level characteristics, it was possible that a district contained multiple schools across different strata. Thus, it was sometimes impossible to match each district with a unique stratification level. Therefore, district weights were created by expanding or replicating the district-level file so that the number of observations equaled the number of sampled eligible schools. These weights were post-stratified to the sampling frame totals by three Orshansky levels (less than 10%, 11% to 40%, and greater than 41%). Finally, the weights for responding districts were adjusted for district level nonresponse. Given the definitions in Section 5.1, let M_j be the number of eligible schools on the frame within district j, and m_{hj} denote the number of times district j was included within stratum h, then the basic sampling weight $(w_{(p)j})$ for district j is:

$$w_{(p)j} = \frac{1}{M_j} \sum_h \frac{N_h}{n_h} m_{hj}. \tag{1}$$

Denoting the number of sample districts in Orshansky level o, $(o = 1, 2, 3)$ by n_o and the number of eligible schools selected in district j by n_j, equation (1) provides the sampling weight $(w_{(v)j})$ for weighted estimation as: $w_{(v)j} = w_{(p)j}/n_j$. The post-stratified weight $(w^{ps}_{(v)j})$ for district j is: $w^{ps}_{(v)j} = w_{(v)j}(n_o / \sum_o w_{(v)j})$, where $\sum_o w_{(v)j}$ = sum of initial district weights for point and variance estimation by Orshansky levels $(o = 1, 2, 3)$. The non-response adjusted weight $(w^a_{(v)j})$ for district j is :

$$w^a_{(v)j} = w^{ps}_{(v)j} \left(\frac{\sum_o w^{eps}_{(v)j}}{\sum_o w^{rps}_{(v)j}} \right), \tag{2}$$

where $\sum_o w^{eps}_{(v)j}$ = sum of post-stratified weights for eligible districts, and $\sum_o w^{rps}_{(v)j}$ = sum of post-stratified weights for responding districts.

5.3 *Sampling Results*

Table 1 shows various unweighted eligibility and response rates from school sample. Rates are provided for 27 different levels of explicit public school strata plus 2 levelest for non-public schools, including three domain variables of interest (urbanicity, Orshansky percent, and schools size). Individual Orshansky-level eligibility and response rates are also given in Tables 3 and 4. Eligibility rates for both the public and Catholic schools (95.5 and 92.1) proved significantly higher than rates from non-Catholic private schools (71.0). Response rates range from a low of 46.7

Table 1. Frame Count with Number Selected, Eligibles, and Respondents by Strata for SSUPPS

Design Stratum (urbanicity, Orshansky[1], size[2])	Frame Count of Schools	Number of Selected Schools	Eligible Schools		Responding Schools	
Public Schools			Number	Percent	Number	Percent
1 - urban, low, small	57	10	8	80.0	6	75.0
2 - urban, low, medium	236	33	33	100.0	17	51.5
3 - urban, low, large	280	39	38	97.4	32	84.2
4 - urban, medium, small	262	39	33	84.6	17	51.5
5 - urban, medium, medium	1,007	134	127	94.8	89	70.1
6 - urban, medium, large	1,333	178	172	96.6	133	77.3
7 - urban, high, small	140	20	15	75.0	7	46.7
8 - urban, high, medium	615	83	78	94.0	60	76.9
9 - urban, high, large	886	119	114	95.8	82	71.9
10 - suburban, low, small	374	42	37	88.1	27	73.0
11 - suburban, low, medium	1,381	152	148	97.4	102	68.9
12 - suburban, low, large	1,219	138	135	97.8	99	73.3
13 - suburban, medium, small	776	70	60	85.7	46	76.7
14 - suburban, medium, medium	2,242	202	192	95.0	148	77.1
15 - suburban, medium, large	1,883	170	165	97.1	121	73.3
16 - suburban, high, small	383	35	32	91.4	22	68.8
17 - suburban, high, medium	830	75	70	93.3	51	72.9
18 - suburban, high, large	632	57	56	98.2	39	69.6
19 - rural, low, small	592	68	65	95.6	49	75.4
20 - rural, low, medium	505	57	56	98.2	41	73.2
21 - rural, low, large	120	14	14	100.0	11	78.6
22 - rural, medium, small	3,579	322	314	97.5	232	73.9
23 - rural, medium, medium	1,385	125	123	98.4	96	78.0
24 - rural, medium, large	233	21	20	95.2	14	70.0
25 - rural, high, small	1,440	130	124	95.4	83	66.9
26 - rural, high, medium	448	41	39	95.1	29	74.4
27 - rural, high, large	53	6	5	83.3	3	60.0
Urban	4,816	655	618	94.4	443	71.7
Suburban	9,720	941	895	95.1	655	73.2
Rural	8,355	784	760	96.9	558	73.4
Orshansky (low)	4,764	553	534	96.6	384	71.9
Orshansky (medium)	12,700	1261	1206	95.6	896	74.3
Orshansky (high)	5,427	566	533	94.2	376	70.5
Size (small)	7,603	736	688	93.5	489	71.1
Size (medium)	8,649	902	866	96.0	633	73.1
Size (large)	6,639	742	719	96.9	534	74.3
Public School	22,981	2,380	2,273	95.5	1,656	72.9
30 - Catholic	6,031	189	174	92.1	129	74.1
40 - Other Private	8,919	283	201	71.0	120	59.7
Non-Public School	14,950	472	375	79.4	249	66.4
All Schools	37,841	2,852	2,648	92.8	1,905	71.9

[1] The Orshansky scores are categorized as: Low - less than or equal to 10%, Medium - 11 to 40%, and High - 41% and higher.

[2] Eligible grade sizes are estimates of the total number o students enrolled at school for all middle school specific grades (5-8). The size distinctions are: Small - fewer than 200 students, Medium - 200 to fewer than 600 students, and Large - 600 and more students.

Table 2. Public School District Level Eligibility and Response Rates

Design Stratum (Orshansky)	Number of Districts Included	Eligible Districts		Responding Districts	
		Number	Percent	Number	Percent
Low	526	518	98.5	403	77.8
Medium	1,069	1,051	98.3	856	81.4
High	423	415	98.1	334	80.5
Total	2,018	1,984	98.3	1,593	80.3

Table 3. School Eligibility Percentage by School Type

School Type	Public	Catholic	Other Private	Total
Percentage	95.5	92.1	71.0	92.8

Table 4. School Responding Percentage by School Type

School Type	Public	Catholic	Other Private	Total
Percentage	72.9	74.1	60.2	72.0

percent (in the urban/high Orshansky/small stratum) to a high of 84.2 percent (in the urban/low Orshansky/large stratum) (see Table 1). Table 4 shows that response rates for public and Catholic schools (72.9 and 74.1%) are highly significant when compared to response rates from other private schools (60.2%). The combined response rate for all schools was 72.0 percent. The resulting overall unequal weighting effect (UWE) for n^r responding schools was 1.45, where

$$UWE = n^r \frac{\sum_h \sum_i (w_{hi}^a)^2}{(\sum_h \sum_i w_{hi}^a)^2}$$

Table 2 provides public school district-level eligibility and response rates for the three Orshansky levels. This table shows fairly equal eligibility rates across all levels with a slightly lower response rate (77.8%) for the lowest Orshansky level. The overall eligibility rate was 98.3% and the overall response rate was 80.3%. The overall non-response adjusted unequal weighting effect for school districts was 1.60.

5.4 Data Collection Experience

Table 5 shows some weighted questionnaire findings from the teacher data collection activities. Approximately 64 percent of all public schools have a coordinator or a

Table 5. School Weighted Percentage of Teachers Responsible for Substance Use Prevention Education

School Type	Public	Catholic	Other Private	Total
Percentage	64.1	42.2	40.0	56.3

Table 6. Weighted Percentage of Districts Responding Middle Schools with a Substance Use Prevention Education Coordinator

All Schools	Some Schools	No Schools
67.4	9.2	23.5

specific teacher who is primarily responsible for substance use prevention education at the school. This prevalence rate is significantly higher than that for Catholic and non-Catholic private schools (42.2 and 40.0 percent). Table 6 shows the weighted questionnaire findings for public school districts. Comparable to school findings, approximately 67 percent of public school districts indicated that a coordinator or a specific teacher was present in all schools within the district with the responsibility for substance use prevention activities in the school.

6 Conclusions

One of the most important sample design issues for this study relates to the identification of the sampling frame. When the SSUPPS study was designed, the data sources known were CCD, MDR, and QED. According to Hamann [5], all three sources are adequate for public schools, but QED is the only source containing both public and private schools. This feature led to our use of QED. However, since this survey was implemented, NCES has made available on the NCES website (http://nces.ed.gov) a detailed data source for private schools known as the Private School Universe Survey (PSS). If the surveys were implemented today, we would recommend a combination of CCD and PSS for the sampling frame.

Appendix: List of Acronyms Used

1. SAMHSA - Substance Abuse and Mental Health Services Administration
2. SSUPPS - School-Based Substance Use Prevention Programs Study
3. SDFSCA - Safe and Drug-Free Schools and Communities Act

4. CCD - Common Core of Data
5. NCES - National Center for Education Statistics
6. MDR - Market Data Retrieval
7. QED - Quality Education Data
8. PSS - Private School Universe Survey

Acknowledgments

The authors thank the supporter for this research: National Institute on Drug Abuse, Dr. Elizabeth Robertson (Project Officer), NIDA Grant 5R01 DA11492-02. Thanks to members of the School-Based Substance Use Prevention Programs Study project team for comments concerning this document: Allison Burns, Susan T. Ennett, Matthew C. Farrelly, Ruby E. Johnson, Kurt Ribisl, Christopher L. Ringwalt, Luanne Rohrbach, Eva W. Silber, Ashley P. Simons-Rudolph, Judy M. Thorne, Amy A. Vincus, and Dana L. Wenter. In addition, the authors thank James R. Chromy for his suggestions and technical review, David Kellogg for his editorial review, and Brenda E. Gurley for her clerical contributions.

References

1. W.G. Cochran, *Sampling Techniques, Third Edition* (John Wiley & Sons Inc., 1977).
2. Dun & Bradstreet Corporation, *Market Data Retrieval Database*(School Year 1994-1995).
3. L. Dusenbury and M. Falco, *Journal of School Health* **65**, 420 (1995).
4. L.D. Eigen and D.W. Rowden, *Making the Case for Prevention: A Discussion Paper on Preventing Alcohol, Tobacco, and Other Drug Problems* (Department of Health and Human Services, Public Health Service, Washington, DC, 1993).
5. T.A. Hamann, *Evaluating the Coverage of the U.S. National Center for Education Statistics Public Elementary/Secondary School Frame* (Presented at the International Conference on Establishment Surveys - II. Governments Division, U.S. Census Bureau, 2000).
6. Jones, S.M., and J.R. Chromy, In *Proceedings of the Section on Survey Research Methods* (American Statistical Association, 1982).
7. L. Kann, J.L. Collins, B.C. Pateman, M.L. Small, J.G. Russ and L.T. Kolbe, *Journal of School Health* **65**, 291 (1995).
8. L. Kish, *Survey Sampling* (John Wiley & Sons, New York, 1995).
9. Quality Education Data Inc., *QED National Education Database: Data Users Guide, Version 4.6* Quality Education Data, Inc., Denver, CO, 1998).

10. C.L. Ringwalt, J.M. Greene, S.T. Ennett, R. Iachan, R.R. Clayton and C.G. Leukefeld, *Past and Future Directions of the D.A.R.E. Program: An Evaluation Review* (Prepared for the National Institute of Justice, 1994).
11. Substance Abuse and Mental Health Services Administration, *Summary of Findings from the 1998 National Household Survey on Drug Abuse* (Office of Applied Studies, U.S. Department of Health and Human Services, 1999)..
12. SAS Institute Inc., *SAS Procedures Guide, Version 6*, Third Edition (1990).
13. B.V. Shah, B.G. Barnwell and G.S. Bieler, *SUDAAN Users Manual* (Research Triangle Institute, Research Triangle Park, NC, 1996).
14. M.G. Sirken, *Journal of the American Statistical Association* **67**, 224 (1972).
15. U.S. Department of Education, *Public Elementary/Secondary Universe Survey* (National Center for Education Statistics, School Year 1994-95).

ONE-STEP ESTIMATION FOR THE PARTIALLY LINEAR PROPORTIONAL HAZARDS MODEL

XUEWEN LU

Department of Mathematics and Statistics, University of Calgary, 2500 University Drive, Calgary, Alberta, T2N 1N4, Canada

R.S. SINGH

Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario, N1G 2W1, Canada

E-mail: rssingh@uoguelph.ca

The proportional hazards regression model usually assumes that the covariate has a log-linear effect on the hazard function. In this paper, we consider a semiparametric survival model with flexible covariate effects. We assume the baseline hazard function can be parameterized, while the risk function associated with covariates is modeled in a semiparametric way. A "one-step algorithm" is used to estimate the nonparametric components and the parametric components by maximizing the local and the global likelihood, respectively. Using the local linear method, the estimates of the unknown parameters and the unknown covariate function are determined, and their asymptotic distributions are obtained.

1 Introduction

The Cox regression model (see Cox [7]) is an important model in survival analysis, in which the conditional hazard of failure at time t given the covariate value z is modeled as

$$h(z|t) = h_0(t)\exp(z^T\beta).$$

Here, the baseline hazard $h_0(t)$ is nonparametric, while the dependency on the covariate z is parametric. The partial likelihood principle provides efficient estimates of β in the semiparametric model (see Andersen and Gill [1]). The Cox proportional hazards model has had profound impact on clinical trials for quantifying the effects of covariates on the survival time and for controlling confounding by means of a mathematical model which takes the outcome under consideration as the dependent variable, and includes both the postulated causal factor and confounding factors in the equation (see Fleming and Lin [13] and Elwood [10]).

Despite the advantages of the Cox model, many authors have considered nonparametric and semiparametric modeling of covariate effects on the censored failure time. For example, Beran [3], Dabrowska [8], McKeague and

Utikal [23] and Nielsen and Linton [25] study the fully nonparametric hazard model where $h(z,t)$ is unspecified. Fan, *et al.* [11] treat the multiplicative nonparametric model where $h(z,t) = h_0(t)\lambda(z)$ in which neither $h_0(\cdot)$ nor $\lambda(\cdot)$ is parametrically specified. Sasieni [29,30] examines a partially linear covariate effect in a model where $h(z,t) = h_0(t)\exp\{z_1^T\beta + \lambda(z_2)\}$. Dabrowska [9] discusses a generalized Cox regression model with baseline intensity dependent on a covariate in which $h(z,t) = h_0(t,z_1)\exp\{z_2^T\beta\}$. Hastie and Tibshirani [15,16] consider a fully nonparametric additive Cox model. Huang [17] studies a partly linear additive Cox model. Grambsch, Therneau and Fleming [14] and Fleming and Harrington [12] (Section 4.5, pages 163-168) propose to use smoothed martingale residuals to explore the functional form of the covariate effect in the Cox model. A survey of other regression models for censored survival data can be found in Andersen *et al.* [2].

In this paper, we study the following semiparametric survival model :

$$h(t|z,x) = h_0(t;\sigma)\exp\{z^T\beta + \lambda(x)\}, \tag{1}$$

where h_0 is a baseline hazard function with an unknown parameter vector σ. $(z^T,x)^T \in \mathbf{R}^p \times \mathbf{R}$ is a covariate vector. The variable x is continuous with values in a compact set $\mathcal{X} \in \mathbf{R}$, and z is discrete or continuous with values in $\mathbf{R}^p$. $\theta = (\beta^T,\sigma^T)^T$ is an unknown $q+p$ parameter vector, $\lambda(\cdot)$ is an unknown smooth univariate function. In this model, the baseline hazard is modeled parametrically; the covariate effects are modeled semiparametrically. In many applications the shape of the baseline hazard is well understood: for example, the Gompertz-Makeham hazard is used extensively in insurance problems (Jordan [19], page 21). A linear approximation to the logarithm of the baseline hazard is successful in a number of chronic diseases problems (Meshalkin and Kagan [24]). Also, the Weibull baseline is widely used in both biostatistical and reliability applications, *e.g.* Lawless [21]. The covariate effect, however, is rarely specified by a completely parameterized model. For instance, in a clinical trial study, when Z is a treatment covariate and X is a covariate describing characteristic of the patients, β can be interpreted as a measure of the treatment effect after adjusting for the effect of X. The effect of X can be modeled in a nonparametric way. Fleming and Lin [13] provide an overall review for the developments of survival analysis in clinical trials. Hence, this model allows flexible modeling of the covariate effect in many applications including clinical trials. We assume that the covariates are time-independent due to the fact that this type of data arises often in practice and the technical details involved in time-dependent covariate models are hard to tackle.

Nielsen *et al.* [26] study model (1) without the linear part $z^T\beta$ and they assume the covariate x is time-dependent. As they mention in their paper,

the partial likelihood principle fails to estimate the parametric part in model (1). A generally applicable approach for this type of semiparametric model is that of profile likelihood, which is discussed extensively in Severini and Wong [31] and Bickel *et al.* [4].

When $\lambda(x)$ has a parametric form, for example, $\lambda(x) = x^T\gamma$, it is well known that one can use maximum likelihood methods to analyze the model. Kalbfleisch and Prentice [20] and Lawless [21] give a detailed introduction to the use of maximum likelihood methods in the analysis of parametric regression models in the context of survival analysis. Most statistical software, for example, SAS and SPLUS, can analyze the Cox model, but they lack the ability to estimate the parameters in the baseline hazard function when the covariates are modeled in a nonparametric or semiparametric manner. That is, when the regression function $\lambda(x)$ is not parameterized as a linear form or is an unknown regression function. This paper discusses how to estimate both the parametric and nonparametric parts in model (1).

The paper is organized as follows. Section 2 of the paper introduces the likelihood for the survival model under right-censoring. Section 3 describes the local likelihood and local linear fit method. Section 4 proposes the one-step estimation method. Section 5 studies the asymptotic distributions for both parametric and nonparametric parts. Section 6 gives some conclusion remarks on the proposed method. Finally, the Appendix gives the technical proofs.

2 Likelihood Function for A Parametric Survival Model

Let $f(t|z,x)$ denote the conditional density function of lifetime T given $(Z,X) = (z,x)$, and let $S(t|z,x) = P\{T > t|(Z,X) = (z,x)\}$ be its conditional survival function. The conditional distribution function of a random censoring variable C given $(Z,X) = (z,x)$ is denoted by $G(t|z,x)$. Then under independent and *noninformative censoring* (*i.e.*, $G(t|z,x)$ does not involve the unknown parameters), the conditional likelihood function is given by

$$L = \prod_u f(Y_i|Z_i,X_i) \prod_c S(Y_i|Z_i,X_i), \tag{2}$$

where $\prod_u$ and $\prod_c$ denote, respectively, the products involving the uncensored and the censored observations.

Given that $H_0(t;\sigma)$ is the cumulative baseline hazard function and that $\lambda(x)$ is parameterized as $\lambda(x) = \lambda(x;\gamma)$, the conditional survivor function for T at $(Z,\ X) = (z,\ x)$, under the survival model (1), is of the form

$$S(t|z,x;\sigma,\beta,\gamma) = \exp[-H_0(t;\sigma)\exp\{z^T\beta + \lambda(x;\gamma)\}].$$

The hazard function and cumulative hazard functions are given by $h(t|z,x;\sigma,\beta,\gamma) = h_0(t;\sigma)\exp\{z^T\beta + \lambda(x;\gamma)\}$ and $H(t|z,x;\sigma,\beta,\gamma) = H_0(t;\sigma)\exp\{z^T\beta + \lambda(x;\gamma)\}$, respectively.

If we let δ_i represent the censoring indicator, *i.e.* $\delta_i = I[T_i \leq C_i]$, and Y_i represent a lifetime or a censoring time for i^{th} individual, *i.e.* $Y_i = \min(T_i, C_i)$, the likelihood function (2) for samples $\{(Z_i, X_i, Y_i, \delta_i),\ i = 1, \cdots, n\}$ can be written as

$$L(\theta,\gamma) = \prod_1^n \{h(Y_i|Z_i,X_i)\}^{\delta_i} S(Y_i|Z_i,X_i),$$

remembering that $\theta = (\beta^T, \sigma^T)^T$. Under the survival model (1), we have the log-likelihood for the sample

$$\begin{aligned} L_n(\theta,\gamma) &= \sum_i \{\delta_i \log f(Y_i|Z_i,X_i) + (1-\delta_i)\log S(Y_i|Z_i,X_i)\} \\ &= \sum_i [\delta_i\{\log h_0(Y_i;\sigma) + (Z_i^T\beta + \lambda(X_i;\gamma))\} \\ &\quad + \exp\{Z_i^T\beta + \lambda(X_i;\gamma)\}\log S_0(Y_i;\sigma)] \\ &= \sum_i [\delta_i\{\log h_0(Y_i;\sigma) + (Z_i^T\beta + \lambda(X_i;\gamma))\} \\ &\quad - \exp\{Z_i^T\beta + \lambda(X_i;\gamma)\}H_0(Y_i;\sigma)], \end{aligned} \tag{3}$$

using $H_0(t;\sigma) = -\log S_0(t;\sigma)$. Maximization of (3) leads to the maximum likelihood estimators of σ, β and γ.

3 Local Likelihood

Suppose that the form of $\lambda(x)$ is not specified, and that the first order derivative of $\lambda(x)$ at the point x exists. If we let $a_0 = \lambda(x), a_1 = \lambda'(x)$, then, by Taylor's expansion,

$$\lambda(X) \approx a_0 + a_1(X - x),$$

for X in a neighborhood of x.

Let b be the bandwidth parameter that controls the size of the local neighborhood and let W be a kernel function. Using this local model, one would find a_0, a_1 and $\theta = (\beta^T, \sigma^T)^T$ to maximize the local (log) likelihood

$$\ell_n(a_0,a_1,\beta,\sigma) = \sum_{i=1}^n l_i\{\sigma, Z_i^T\beta + a_0 + a_1(X_i - x), (Y_i,\delta_i)\}W_b(X_i - x), \tag{4}$$

where $l_i\{\sigma, Z_i^T\beta + a_0 + a_1(X_i - x), (Y_i, \delta_i)\} = \delta_i[\log h_0(Y_i;\sigma) + (Z_i^T\beta + a_0 + a_1(X_i - x))] - H_0(Y_i;\sigma)\exp[Z_i^T\beta + (Z_i^T\beta + a_0 + a_1(X_i - x))]$, $W_b(t) = b^{-1}W(t/b)$.

But the estimates of the true parameters $\theta_0 = (\beta_0^T, \sigma_0^T)^T$ from this local likelihood are not efficient since we use only a fraction of all data points. Lu *et al.* [22] consider model (1) using the generalized profile likelihood method with local constant fit and obtained an efficient estimate of (β_0, σ_0) which achieves the semiparametric information bound. But that approach requires a fully iterated algorithm and is not computationally convenient. In this paper, we provide the one-step algorithm with local linear fit. Application of the local linear fit enables us to reduce the bias of the estimator for the nonparametric component and to avoid boundary effects. In some important applications, we show that one-step algorithm achieves the same efficiency.

4 The One Step Estimation Method

Under model (1), the primary interest is to estimate the true parameters σ_0 and β_0 and the true function $\lambda_0(\cdot)$. Since $\lambda_0(\cdot)$ is modeled nonparametrically, it is natural to consider local likelihood. However, efficient estimation of the global parameters σ_0 and β_0 requires using all data points.

One approach is that we first find $\hat{a}_0$, $\hat{a}_1$ and $\hat{\theta}$ to maximize the local (log) likelihood (4). Denote $\hat{\lambda}(x) \equiv \hat{\lambda}(x;b) = \hat{a}_0$ and $\hat{\theta} \equiv \hat{\theta}(x;b)$. Having constructed the function $\hat{\lambda}(x;b)$, estimate the global parameters θ by maximizing the global likelihood. The following one-step algorithm makes it work.

Step 0: Fit a parametric linear model to obtain the initial estimates $\hat{\theta}$.

Step 1: Find $\hat{\lambda}(x;b)$ by maximizing the local likelihood (4). We take b to be an estimate of the bandwidth that is optimal for estimation of θ_0.

Step 2: With the estimated $\hat{\lambda}(x;b)$, find $\hat{\theta}$ by maximizing the global likelihood

$$\sum_{i=1}^{n} l_i\{\sigma, Z_i^T\beta + \hat{\lambda}(X_i;b), (Y_i, \delta_i)\}. \tag{5}$$

Step 3: Fix θ at its estimated value from Step 2. The final estimate of $\lambda_0(\cdot)$ is $\hat{\lambda}(x;b,\hat{\theta}) = \hat{a}_0$ where $(\hat{a}_0, \hat{b}_0)$ maximizes

$$\sum_{i=1}^{n} l_i\{\hat{\sigma}, Z_i^T\hat{\beta} + a_0 + a_1(X_i - x), (Y_i, \delta_i)\}W_b(X_i - x). \tag{6}$$

At this final step we take b to be an estimate of the bandwidth that is optimal for estimation of $\lambda_0(\cdot)$ when θ_0 is known.

To compare the one-step algorithm with the generalized profile likelihood method or the fully iterated algorithm, we also give the fully iterated algorithm as follows:

Step 0: Fit a parametric linear model to obtain the initial estimates $\hat{\theta}$.

Step 1: Find $\hat{\lambda}(x; b, \sigma, \beta)$ by maximizing the local likelihood (4). We take b to be an estimate of the bandwidth that is optimal for estimation of θ_0.

Step 2: Update $\hat{\theta}$ by maximizing

$$\sum_{i=1}^{n} l_i\{\sigma, Z_i^T\beta + \hat{\lambda}(x; b, \sigma, \beta), (Y_i, \delta_i)\}. \tag{7}$$

Step 3: Fix θ at its estimated value from Step 2. The final estimate of $\lambda_0(\cdot)$ is $\hat{\lambda}(x; b, \hat{\theta}) = \hat{a}_0$ where $(\hat{a}_0, \hat{a}_1)$ maximizes (6). At this final step we take b to be an estimate of the bandwidth that is optimal for estimation of $\lambda_0(\cdot)$ when θ_0 is known.

Instead of maximizing (7), an alternative approach is to update the current $\hat{\theta}$ by maximizing

$$\sum_{i=1}^{n} l_i\{\sigma, Z_i^T\beta + \hat{\lambda}(x; b, \hat{\sigma}, \hat{\beta}), (Y_i, \delta_i)\}.$$

We conjecture that the estimators resulting from this type of updating are equivalent to those obtained by maximizing (7); this would be interesting to verify.

5 Distribution Theory

5.1 One-Step Estimator of the Nonparametric Part

In this subsection, we investigate properties of the estimators of the nonparametric part $\lambda_0(\cdot)$ of (1) under two cases: (1) the one-step approach when θ_0 is estimated locally as in (4); and (2) the final estimate as in (6) when θ_0 is estimated globally. In case (1), we are using only $O(nh)$ data points to estimate both θ_0 and $\lambda_0(\cdot)$. In case (2), θ_0 is estimated at parametric rates, and thus $\lambda_0(\cdot)$ can be estimated asymptotically as well as if θ_0 were known.

We consider case (1) first. Let $f(\cdot)$ be the density of X. Define

$$\kappa_j = \int u^j W(u)\, du, \quad \nu_j = \int u^j W^2(u)\, du;$$

$$\Sigma(x) = E\left[\delta \begin{pmatrix} 1 \\ Z \\ \bar{h}_0(Y;\sigma_0) \end{pmatrix}^{\otimes 2} \mid X = x\right];$$

$$m_i = \lambda_0(X_i) + Z_i^T\beta;$$

$$W_i = \text{first element of the vector } \Sigma^{-1}(x) \begin{pmatrix} \{\delta_i - H_0(Y_i;\sigma_0)\exp(m_i)\}U_i \\ \{\delta_i - H_0(Y_i;\sigma_0)\exp(m_i)\}Z_i \\ \delta_i \bar{h}_0'(Y_i;\sigma_0) - H_0'(Y_i;\sigma_0)\exp(m_i) \end{pmatrix};$$

$$v(x) = \text{first diagonal element of the matrix } \Sigma^{-1}(x);$$

$$\bar{h}_0(t;\sigma) = \log\{h_0(t;\sigma)\} \text{ and } \bar{h}_0'(t;\sigma) = \partial \bar{h}_0(t;\sigma)/\partial\sigma.$$

Theorem 5.1. Under condition A given in the appendix, as $n \to \infty$, $b \to 0$ and $nb \to \infty$ and nb^5 is bounded, for the the maximizer of the local likelihood (4), we have

$$(nb)^{1/2} \left(\begin{bmatrix} \hat{\lambda}(x) - \lambda_0(x) \\ \hat{\beta} - \beta_0 \\ \hat{\sigma} - \sigma_0 \end{bmatrix} - \frac{\kappa_2}{2}\lambda_0''(x)b^2\Sigma^{-1}(x)E\left[\delta \begin{pmatrix} 1 \\ Z \\ \bar{h}_0(Y;\sigma_0) \end{pmatrix} | X = x\right] \right)$$
$$\xrightarrow{D} N\left[0, \frac{\nu_0}{f(x)}\Sigma^{-1}(x)\right].$$

Hence

$$(nb)^{1/2}\{\hat{\lambda}(x) - \lambda_0(x) - \frac{\kappa_2}{2}\lambda_0''(x)b^2\} \xrightarrow{D} N(0, \frac{\nu_0}{f(x)}v(x)). \tag{8}$$

5.2 Fully Iterated Estimator of the Nonparmetric Part

One-step estimation of the parameters σ_0 and β_0 requires a smaller bandwidth than the one that is optimal for estimating the nonparametric component $\lambda_0(\cdot)$. Thus, the estimate is undersmoothed. If one is interested in obtaining a good estimate of $\lambda_0(\cdot)$, a final step in the algorithm (6) should be carried out. In this case, (8) continues to hold if we replace $v(x)$ by $v_*(x) = \{E(\delta|X = x)\}^{-1}$. The result is given as follows:

Theorem 5.2.

$$(nb)^{1/2}\left\{\hat{\lambda}(x)-\lambda_0(x)-\frac{\kappa_0}{2}\lambda_0''b^2\right\}\xrightarrow{D}N\left[0,\frac{\nu_0}{f(x)E(\delta|X=x)}\right].$$

The asymptotic variance in this result coincides with the univariate result given in Nielsen *et al.* [26] and Lu *et al.* [22] when covariates are time-independent, but the bias here is different, because we use the local linear fit. This result suggests bandwidth estimators in the spirit of that of Ruppert, Sheather and Wand [28]. Following the discussions made in Carroll *et al.* [5,6], for example, consider estimation of $\lambda_0(\cdot)$ at the final step, for a given function $w(\cdot)$ with compact support, minimizing the asymptotic weighted mean squared error with weight $f(\cdot)w(\cdot)$ yields the optimal global bandwidth

$$b_{opt}=C(W)n^{-1/5}\left\{\frac{\int v_*(x)w(x)\,dx}{\int\lambda_0''(x)^2f(x)w(x)\,dx}\right\}^{1/5},$$

where $C(W)=(\nu_0\kappa_2^{-2})^{1/5}$.

For the one-step algorithm in Step 1, a relatively *ad hoc* choice of b is

$$\hat{b}_{opt}\times n^{1/5}\times n^{-1/3}=\hat{b}_{opt}\times n^{-2/15}.$$

On the other hand, for the fully iterated algorithm (7), we have shown that ordinary bandwidth rates are permissible (Lu *et al.*, [22]). Severini and Staniswallis [32], Hunsberger [18], Severini and Wong [31] show the same thing for generalized partially linear model.

5.3 One-Step Estimator for the Parametric Part

We now study the global estimators for the parametric components (σ_0,β_0).

Theorem 5.3. Let $\hat{\beta}$ and $\hat{\sigma}$ be the one-step estimates which maximizes the likelihood (5). Under conditions A given in the appendix, as $n\to\infty$, $nb^4\to 0$ and $nb^2/\log(1/b)\to\infty$,

$$n^{1/2}\begin{pmatrix}\hat{\beta}-\beta_0\\ \hat{\sigma}-\sigma_0\end{pmatrix}\xrightarrow{D}N(0,B^{-1}\Sigma_1B^{-1}),$$

where

$$B=E(\delta VV^T),$$

$$\Sigma_1=B+E\{\gamma(X)\gamma^T(X)e_1^T\Sigma^{-1}(X)e_1^T\},$$

$\gamma(x)=E(\delta V|X=x)$, $V=(Z^T,\bar{h}_0'(Y;\sigma_0)^T)^T$, e_1 is the unit vector with 1 in the first position.

Therefore, one iteration leads already to a root-n consistent estimator of $\theta_0 = (\beta_0^T, \sigma_0^T)^T$.

We have derived the results using local constant fit for the fully iterated estimator defined by (7) (see Theorem 5.2, Lu *et al.*, [22]). The results also hold for the local linear fit in this paper. For the sake of completeness, we cite the result as follows:

Theorem 5.4. Let $\hat{\theta} = (\hat{\beta}^T, \hat{\sigma}^T)^T$ be the fully iterated estimates which maximizes the likelihood (7). Under Condition I, S and A given in Lu *et al.* [22], assume i_{θ_0} is positive definite and finite, then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} N(0, i_{\theta_0}^{-1}),$$

where i_{θ_0} is the marginal Fisher information matrix for θ_0 given by

$$i_{\theta_0} = E_0 \left(\frac{\partial l}{\partial \theta}(\theta_0, \lambda_0) + \frac{\partial l}{\partial \lambda}(\theta_0, \lambda_0)(v^*) \right)^{\otimes 2},$$

$v^* = \lambda'_{\theta_0}$ is the least favorable direction. i_{θ_0} can be consistently estimated by $\hat{i}_{\theta_0} = -\frac{1}{n} \frac{d^2}{d\theta^2} L_n(\theta, \hat{\lambda}_\theta) \big|_{\theta=\hat{\theta}}$.

i_{θ_0} has the following simple expression:

$$\begin{aligned} i_{\theta_0} &= E_0 \int \frac{\partial \mu_{\theta_0}}{\partial \theta} \frac{\partial \mu_{\theta_0}}{\partial \theta^T}(t, Z, X) h_0(t; \sigma_0) \exp[Z^T \beta_0 + \lambda_0(X)] I(t)\, dt \\ &= E_0 \int \frac{\partial \mu_{\theta_0}}{\partial \theta} \frac{\partial \mu_{\theta_0}}{\partial \theta^T}(t, Z, X)\, dN(t) \\ &= E_0[\delta \{ \frac{\partial \mu_{\theta_0}}{\partial \theta} \frac{\partial \mu_{\theta_0}}{\partial \theta^T}(Y, Z, X) \}] \\ &= E_0(\delta V V^T) - E_0 \left\{ \frac{E_0(\delta V | X) E_0(\delta V^T | X)}{E_0(\delta | X)} \right\}, \end{aligned}$$

where $\mu_\theta(t, z, x) = \log[h_0(t; \sigma) \exp\{z^T \beta + \lambda_\theta(x)\}]$, $\lambda_\theta(x) = \lambda_0(x) + r_{\theta_0}(x) - r_\theta(x)$ and $r_\theta(x) = \log E\{\exp(Z^T \beta) H_0(Y; \sigma) | X = x\}$. We have proved i_{θ_0} is the semiparametric information bound.

Remark 5.1. The fully iterated estimator is uniformly as or more efficient than the one-step estimator. However, when X and Z are independent and $\delta \equiv 1$ (there is no censoring), the one-step estimator is as efficient as the fully iterated estimator. Therefore, iteration is not necessary when X and Z are weakly correlated and the censoring is not heavy. In fact, if we set $E(Z) = 0$ without loss of generality, then, when X and Z are independent and $\delta \equiv 1$, we have $E(\delta Z | X) = E(Z) = 0$, and the asymptotic variance for the one-step estimator of $\hat{\beta}$ equal to $\{E(ZZ^T)\}^{-1}$, which is also the asymptotic variance for the fully iterated estimator given in (7).

6 Conclusions

For the generalized partially linear single-index models, the asymptotic distribution theory for the one-step and the fully iterated estimator, based on the local linear fit, has been established by Carroll *et al.* [5]. In this paper, we formulate our approach in the frame work of proportional hazards regression models. We can draw some similar conclusions from our study:

(i) estimation of the nonparametric part $\lambda_0(\cdot)$ of model (1) can be done just as well as if the parameters (σ_0, β_0) were known;

(ii) the parametric parts can be estimated at the usual parametric rate of convergence;

(iii) the fully iterative estimator can have a smaller variance-covariance matrix than the one-step estimator, but for the case without censoring and with weakly correlated X and Z both estimators share the same variance-covariance matrix;

(iv) estimation of the nonparametric part in the model has an effect on the distribution of the estimates for the parametric parts;

(v) best estimation of the parametric parts requires undersmoothing of the nonparametric part.

Our model (1) is restricted on univariate in the nonparametric part. When there are several covariates in the nonparametric part, dimensionality reducing strategies such as additive and index models can be used. For instance, one could fit the following partially linear single-index model:

$$h((t; z, x) = h_0(t, \sigma) \exp\{z^T\beta + \lambda(x^T\gamma)\},$$

where $\lambda(\cdot)$ is of unknown form and, σ, β and γ are unknown parameters. Further development of this issue needs to be investigated.

Appendix: Technical Proofs

The following conditions will be needed:
Condition A:

(i). $\theta_0 = (\sigma_0^T, \beta_0^T)^T$ is an interior point of the parameter space Θ.

(ii). Z is bounded. X takes values in $\mathcal{X}$, a compact set in $\mathbf{R}$.

For any $x \in \mathcal{X}$,

(iii). There exists $\xi > 0$ such that

$$E[\{H_0(Y;\sigma_0)\exp(Z^T\beta_0)\}^{2+\xi}|X], \quad E\{|H_0(Y;\sigma_0)\exp(Z^T\beta_0)Z|^{2+\xi}|X\}$$

and

$$E\{|\bar{h}_0'(Y;\sigma_0)|^{2+\xi}|X\}, \quad E\{|H_0'(Y;\sigma_0)\exp(Z^T\beta_0)|^{2+\xi}|X\}$$

are finite and continuous at the point $X = x$, where $\bar{h}_0(Y;\sigma_0) = \log h_0(Y;\sigma_0)$.

(iv). The functions $E\{H_0(Y;\sigma_0)\exp(Z^T\beta_0)|X\}$, $E\{H_0'(Y;\sigma_0)\exp(Z^T\beta_0)|X\}$, $E\{H_0'(Y;\sigma_0)\exp(Z^T\beta_0)Z|X\}$, $E\{H_0(Y;\sigma_0)\exp(Z^T\beta_0)ZZ^T|X\}$, $E\{H_0''(Y;\sigma_0)\exp(Z^T\beta_0)|X\}$, $E\{\delta\bar{h}_0'(Y;\sigma_0)|X\}$ and $E\{\delta\bar{h}_0''(Y;\sigma_0)|X\}$ are all continuous at the point $X = x$. The function $E(\delta|X)$ is positive and has a continuous $2nd$ derivative around the point x .

(v). There exists a function $M(y)$, with $E\{M(Y)\} < \infty$, such that

$$\left|\frac{\partial^3}{\partial\sigma_j\partial\sigma_k\partial\sigma_l}\bar{h}_0(y;\sigma)\right| < M(y), \quad \left|\frac{\partial^3}{\partial\sigma_j\partial\sigma_k\partial\sigma_l}H_0(y;\sigma)\right| < M(y),$$

for all y, and for all σ in a neighborhood of σ_0.

(vi). The kernel function $W \geq 0$ is a bounded density supported on $[-1,1]$ symmetric about zero.

(vii). The function $\lambda_0(x)$ has a continuous $2nd$ derivative around the point x.

(viii). The density $f(\cdot)$ of X has a continuous $2nd$ derivative around the point x and $f(x) > 0$.

Let $g_0 = \bar{h}_0(Y;\sigma_0)$, define

$$S_0(x) = E\left\{\delta\begin{pmatrix} 1 & 0 & Z^T & g_0^T \\ 0 & \kappa_2 & 0 & 0 \\ Z & 0 & ZZ^T & Zg_0^T \\ g_0 & 0 & g_0Z^T & g_0^{\otimes 2} \end{pmatrix} \mid X = x\right\}$$

and

$$S_1(x) = E\left\{\delta\begin{pmatrix} \nu_0 & 0 & \nu_0Z^T & \nu_0g_0^T \\ 0 & \nu_2 & 0 & \nu_00 \\ \nu_0Z & 0 & \nu_0ZZ^T & \nu_0Zg_0^T \\ g_0 & 0 & g_0Z^T & \nu_0g_0^{\otimes 2} \end{pmatrix} \mid X = x\right\}.$$

Proof of Theorem 5.1: Let $c_n = (nh)^{-1/2}$,

$$X_i^* = \begin{pmatrix} 1 \\ (X_i - x)/b \\ Z_i \end{pmatrix}, \quad \hat{\beta}^* = \begin{pmatrix} c_n^{-1}\{\hat{a} - \lambda_0(x)\} \\ c_n^{-1} h\{\hat{b} - \lambda_0'(x)\} \\ c_n^{-1}(\hat{\beta} - \beta_0) \end{pmatrix}, \quad U_i = \begin{pmatrix} 1 \\ (X_i - x)/b \end{pmatrix},$$

and let

$$\hat{\sigma}^* = c_n^{-1}(\hat{\sigma} - \sigma_0), \quad \hat{\theta}^* = \begin{pmatrix} \hat{\beta}^* \\ \hat{\sigma}^* \end{pmatrix}.$$

Denote $\bar{\lambda}_i = \bar{\lambda}_i(x) = \lambda_0(x) + Z_i^T \beta_0 + \lambda_0'(x)(X_i - x)$. If $(\hat{a}_0, \hat{a}_1, \hat{\theta})^T$ maximizes (4) then $\hat{\theta}^*$ maximizes

$$\sum_{i=1}^{n} l_i(c_n \sigma^* + \sigma_0, c_n \beta^{*T} X_i^* + \bar{\lambda}_i, (Y_i, \delta_i)) W_b(X_i - x)$$

with respect to θ^*. Consider the normalized function

$$l_n(\theta^*) = h \sum_{i=1}^{n} [l_i(c_n \sigma^* + \sigma_0, c_n \beta^{*T} X_i^* + \bar{\lambda}_i, (Y_i, \delta_i)) - l_i(\sigma_0, \bar{\lambda}_i, (Y_i, \delta_i))] W_b(X_i - x),$$

which is maximized by $\hat{\theta}^*$. $l_n(\theta^*)$ is a concave function, we can apply the Convexity Lemma (see Pollard [27]) to it. By a Taylor expansion of the function $l_i(\cdot, \cdot, (Y_i, \delta_i))$ we obtain that

$$l_n(\theta^*) = W_n^T \theta^* - \frac{1}{2} \theta^{*T} A_n \theta^* \{1 + o_p(1)\}, \tag{9}$$

where

$$W_n = h c_n \sum_{i=1}^{n} \begin{pmatrix} \{\delta_i - H_0(Y_i; \sigma_0) \exp(\bar{\lambda}_i)\} U_i \\ \{\delta_i - H_0(Y_i; \sigma_0) \exp(\bar{\lambda}_i)\} Z_i \\ \delta_i \bar{h}_0'(Y_i; \sigma_0) - H_0'(Y_i; \sigma_0) \exp(\bar{\lambda}_i) \end{pmatrix} W_b(X_i - x)$$

and

$$A_n = hc_n^2 \sum_{i=1}^{n} \begin{pmatrix} e^{\bar{\lambda}_i} H_0(Y_i; \sigma_0) U_i U_i^T & e^{\bar{\lambda}_i} H_0(Y_i; \sigma_0) U_i Z_i^T & e^{\bar{\lambda}_i} U_i H_0'(Y_i; \sigma_0)^T \\ e^{\bar{\lambda}_i} H_0(Y_i; \sigma_0) Z_i U_i^T & e^{\bar{\lambda}_i} H_0(Y_i; \sigma_0) Z_i Z_i^T & e^{\bar{\lambda}_i} Z_i H_0'(Y_i; \sigma_0)^T \\ e^{\bar{\lambda}_i} H_0'(Y_i; \sigma_0) U_i^T & e^{\bar{\lambda}_i} H_0'(Y_i; \sigma_0) Z_i^T & -\delta_i \bar{h}_0''(Y_i; \sigma_0) + H_0''(Y_i; \sigma_0) e^{\bar{\lambda}_i} \end{pmatrix} W_b(X_i - x).$$

It is shown that

$$A_n = f(x) S_0(x) + o_p(1) \equiv A + o_p(1).$$

Therefore, by (9),

$$l_n(\theta^*) = W_n^T \theta^* - \frac{1}{2} \theta^{*T} A \theta^* + o_p(1). \tag{10}$$

By applying the Convexity Lemma, we obtain that

$$\hat{\theta}^* = A^{-1}W_n + o_p(1).$$

Hence, the asymptotic normality of $\hat{\theta}^*$ will follow from that of W_n. Since W_n is a sum of i.i.d. random vectors, we only need to compute the first two moments and check Liapounov's conditions. It can be shown that

$$E(W_n) = \sqrt{nb}\{f(x)\frac{\kappa_2}{2}\lambda_0''(x)b^2 + o(b^2)\},$$

$$Var(W_n) = f(x)S_1(x) + o(1).$$

Therefore, we have

$$W_n - EW_n \xrightarrow{D} N\{0, f(x)S_1(x)\}.$$

Proof of Theorem 5.3: We need two lemmas, Lemma A.1 and Lemma A.2, given by Carroll *et al.* [5] in order to prove Theorem 5.2.

By Lemma A.2, each element in A_n converges uniformly to its corresponding element in A. Hence, expression (9) holds uniformly in $x \in \mathcal{X}$. By the Convexity Lemma, it also holds uniformly in $\theta^* \in C$ and $x \in \mathcal{X}$ for any compact set C. Lemma A.1 then yields

$$\sup_{x\in\mathcal{X}} |\hat{\theta}^*(x) - A^{-1}W_n(x)| \xrightarrow{P} 0, \tag{11}$$

where $\hat{\theta}^*(x)$ and $W_n(x)$ are defined in the proof of Theorem 5.1, both depend on x. By considering the first element of the vectors in (11), we have

$$\sup_{x\in\mathcal{X}} \left| \hat{\lambda}(x) - \lambda_0(x) - \frac{1}{nf(x)} \sum_{i=1}^{n} W_i W_b(X_i - x) \right| = o_P(c_n),$$

where W_i is the first element of the vector

$$\Sigma^{-1}(x) \begin{pmatrix} \{\delta_i - H_0(Y_i;\sigma_0)\exp(\bar{\lambda}_i)\}U_i \\ \{\delta_i - H_0(Y_i;\sigma_0)\exp(\bar{\lambda}_i)\}Z_i \\ \delta_i \bar{h}_0'(Y_i;\sigma_0) - H_0'(Y_i;\sigma_0)\exp(\bar{\lambda}_i) \end{pmatrix}.$$

By a result of Carroll *et al.* [5], the following stronger result holds:

$$\sup_{x\in\mathcal{X}} \left| \hat{\lambda}(x) - \lambda_0(x) - \frac{1}{nf(x)} \sum_{i=1}^{n} W_i W_b(X_i - x) \right| = O_P\{b^2 c_n + c_n^2 \log^{1/2}(1/b)\}.$$

Let $\hat{\theta}_1 = n^{1/2}(\hat{\beta} - \beta_0)$, $\hat{\theta}_2 = n^{1/2}(\hat{\sigma} - \sigma_0)$, and $\hat{\theta} = (\hat{\theta}_1^T, \hat{\theta}_2^T)^T$. Denote by

$$\hat{m}_i = \hat{\lambda}(X_i) + Z_i^T\beta_0, \quad \text{and} \quad m_i = \lambda_0(X_i) + Z_i^T\beta_0.$$

Then, $\hat{\tilde{\theta}}$ maximizes

$$l_n(\tilde{\theta}) = \sum_{i=1}^{n} [l_i(n^{-1/2}\theta_1 + \sigma_0, n^{-1/2}Z_i^T\theta_2 + \hat{m}_i, (Y_i, \delta_i)) - l_i(\sigma_0, \hat{m}_i, (Y_i, \delta_i))],$$

By Taylor's expansion, we have

$$l_n(\tilde{\theta}) = n^{-1/2} \sum_{i=1}^{n} \phi_i^T(\hat{m}_i, (Y_i, \delta_i))\tilde{\theta} - \frac{1}{2}\tilde{\theta}^T B_n \tilde{\theta}, \tag{12}$$

where

$$\phi_i(\hat{m}_i, (Y_i, \delta_i)) = \begin{pmatrix} \{\delta_i - H_0(Y_i; \sigma_0)\exp(\hat{m}_i)\}Z_i \\ \delta_i \bar{h}_0'(Y_i; \sigma_0) - H_0'(Y_i; \sigma_0)\exp(\hat{m}_i) \end{pmatrix},$$

$$B_n = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} b_{11}^i & b_{12}^i \\ b_{21}^i & b_{22}^i \end{pmatrix},$$

where $b_{11}^i = \exp(\hat{m}_i + n^{-1/2}Z_i\theta_{1i})H_0(Y_i; \sigma_0 + n^{-1/2}\theta_{2i})Z_iZ_i^T$, $b_{12}^i = \exp(\hat{m}_i + n^{-1/2}Z_i\theta_{1i})Z_iH_0'(Y_i; \sigma_0 + n^{-1/2}\theta_{2i})^T$, $b_{21}^i = \exp(\hat{m}_i + n^{-1/2}Z_i\theta_{1i})H_0'(Y_i; \sigma_0 + n^{-1/2}\theta_{2i})Z_i^T$ and $b_{22}^i = -\delta_i\bar{h}_0''(Y_i; \sigma_0 + n^{-1/2}\theta_{2i}) + H_0''(Y_i; \sigma_0 + n^{-1/2}\theta_{2i})\exp(\hat{m}_i + n^{-1/2}Z_i\theta_{1i})$ with θ_{1i} is between 0 and θ_1 and θ_{2i} is between 0 an θ_2.

It can be shown that

$$B_n = \begin{pmatrix} E\{\delta ZZ^T\} & E\{\delta Z\bar{h}_0'(Y; \sigma_0)^T\} \\ E\{\delta\bar{h}_0'(Y; \sigma_0)Z^T\} & E\{\delta(\bar{h}_0'(Y; \sigma_0))^{\otimes 2}\} \end{pmatrix} + o_P(1) \equiv -B + o_P(1).$$

As for the first term in (12), we have

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n} \phi_i(\hat{m}_i, (Y_i, \delta_i)) &= n^{-1/2}\sum_{i=1}^{n} \phi_i^T(m_i, (Y_i, \delta_i)) \\ &\quad + n^{-1/2}\sum_{i=1}^{n} \psi_i^T(m_i, (Y_i, \delta_i))\{\hat{\lambda}(X_i) - \lambda_0(X_i)\} \\ &\quad + O_P(n^{1/2}||\hat{\lambda} - \lambda_0||_\infty^2), \end{aligned}$$

where

$$\psi_i(m_i, (Y_i, \delta_i)) = \frac{\partial\phi_i}{\partial m_i}(m_i, (Y_i, \delta_i)) = -\begin{pmatrix} H_0(Y_i; \sigma_0)\exp(m_i)\}Z_i \\ H_0'(Y_i; \sigma_0)\exp(m_i) \end{pmatrix}.$$

The second term in the above expression can be expressed as

$$n^{-3/2}\sum_{i=1}^{n}\psi_i^T(m_i,(Y_i,\delta_i))f(X_i)^{-1}\sum_{i=1}^{n}W_jW_b(X_j-X_i)$$
$$+O_P\{n^{1/2}c_n^2\log^{1/2}(1/h)\}\equiv T_{n1}+O_P\{n^{1/2}c_n^2\log^{1/2}(1/h)\}.$$

Define $\nu_j=\nu(X_j,Z_j,(Y_j,\delta_j))$ as the first element of

$$\Sigma^{-1}(X_j)\begin{pmatrix}\delta_j-H_0(Y_j;\sigma_0)\exp(m_j)\\ \{\delta_j-H_0(Y_j;\sigma_0)\exp(m_j)\}Z_j\\ \delta_j\bar{h}_0'(Y_j;\sigma_0)-H_0'(Y_j;\sigma_0)\exp(m_j)\end{pmatrix}$$
$$=\Sigma^{-1}(X_j)\int\begin{pmatrix}1\\ Z_j\\ \bar{h}_0'(t;\sigma_0)\end{pmatrix}dM(t).$$

Using the definition of $\bar{\lambda}_j(X_i)$, we obtain $\hat{\lambda}_j-m_j=O((X_j-X_i)^2)$ and therefore

$$T_{n1}=n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi_i^T(m_i,(Y_i,\delta_i))f(X_i)^{-1}\nu_jW_b(X_j-X_i)+O_P(n^{1/2}h^2)$$
$$\equiv T_{n2}+O_P(n^{1/2}h^2).$$

By a similar procedure used in Carroll *et al.* [5], it can be shown that

$$T_{n2}-T_{n3}\xrightarrow{P}0, \tag{13}$$

where

$$T_{n3}=-n^{-1/2}\sum_{j=1}^{n}\gamma(X_j)\nu_j$$

with

$$\gamma(x)=E\left\{\begin{pmatrix}H_0(Y_i;\sigma_0)\exp(m_i)Z_i\\ H_0'(Y_i;\sigma_0)\exp(m_i)\end{pmatrix}|X=x\right\}.$$

It is seen that

$$\phi_i(m_i,(Y_i,\delta_i))=\int\begin{pmatrix}Z_i\\ \bar{h}_0'(t;\sigma_0)^T)\end{pmatrix}dM_i(t)$$

Combining (12)-(13) we obtain that

$$l_n(\tilde{\theta})=n^{-1/2}\tilde{\theta}^T\sum_{i=1}^{n}\Omega(X_i,(Y_i,\delta_i),Z_i)-\frac{1}{2}\tilde{\theta}^TB\tilde{\theta}+o_P(1),$$

where

$$\Omega(X_i, (Y_i, \delta_i), Z_i) = \phi_i(m_i, (Y_i, \delta_i)) - \gamma(X_i)\nu_i.$$

By the Convexity Lemma we have

$$\hat{\tilde{\theta}} = B^{-1}n^{-1/2}\sum_{i=1}^{n}\Omega(X_i, (Y_i, \delta_i), Z_i) + o_P(1),$$

from which it follows that

$$\hat{\tilde{\theta}} \xrightarrow{D} N(0, B^{-1}\Sigma_1 B^{-1}).$$

This establishes the result in Theorem 5.3.

Acknowledgments

Research of R.S. Singh is supported in part by the National Science and Engineering Research Council Grant # A4631.

References

1. P.K. Andersen and R.D. Gill, *Ann. Statist.* **10**, 1100 (1982).
2. P.K. Andersen, O. Borgan, R.D. Gill and N. Keiding, *Statistical Models Based on Counting Processes* (Springer-Verlag, New York, 1993).
3. R.J. Beran, *Nonparametric regression with randomly censored survival data* (Technical report, Univ. California, Berkeley, 1981).
4. P.J. Bickel, C.A.J. Klaassen, Y. Ritov and J.A. Wellner, *Efficient and Adaptive Inference in Semiparametric Models* (Johns Hopkins Univ. Press, 1993).
5. R.J. Carroll, J. Fan, I. Gijbels, and M.P. Wand, *Discussion Paper #9506* (Institute of Statistics, Catholic University of Louvain, Louvain-la-Neuve, Belgium, 1995).
6. R.J. Carroll, J. Fan, I. Gijbels, and M.P. Wand, *J. Amer. Statist. Assoc.* **92**, 477 (1997).
7. D.R. Cox, *J. Royal Statist. Soc.* **34**, 187 (1972).
8. D.M. Dabrowska, *Scand. J. Statist.* **14**, 181 (1987).
9. D.M. Dabrowska, *Ann. Statist.* **25**, 1510 (1997).
10. J.M. Elwood, *Critical appraisal of epidemiological studies and clinical trials*, Second Edition (Oxford University Press, London, 1998).
11. J. Fan, I. Gijbels and M. King, *Ann. Statist.* **25**, 1661 (1997).
12. T.R. Fleming and D.R. Harrington,*Counting Processes and Survival Analysis* (John Wiley & Sons, New York, 1991).

13. T.R. Fleming and D.Y. Lin, *Biometrics* **56**, 971 (2000).
14. P.M. Grambsch, T.M. Therneau and T.R. Fleming, *Biometrika* **77**, 147 (1990).
15. T. Hastie and R. Tibshirani, *Generalized Additive Models* (Chapman and Hall, London, 1990).
16. T. Hastie and R. Tibshirani, *Biometrics* **46**, 1005 (1990).
17. J. Huang, *Ann. Statist.* **27**, 1536 (1999).
18. S. Hunsberger, *J. Amer. Statist. Assoc.* **89**, 1354 (1994).
19. C.W. Jordan, *Life Contingencies* (The Society of Actuaries, Chicago, 1975).
20. J.D. Kalbfleisch and R.L. Prentice, *The Statistical Analysis of Failure Time Data* (John Wiley and Sons, New York, 1980).
21. J.F. Lawless, *Statistical Models and Methods for Lifetime Data* (John Wiley & Sons, New York,1982).
22. X. Lu, R.S. Singh and A.F. Desmond, *Journal of Statistical Planning and Inference* **98**, 119 (2001).
23. I.W. McKeague and K.J. Utikal, *Ann. Statist.* **18**, 1172 (1991).
24. L.D. Meshalkin and A.R. Kagan, *Discussion of "Regression models and life tables"* by D.R. Cox. *J. Roy. Statist. Soc. Ser. B* **34**, 213 (1972).
25. J.P. Nielsen, O.B. Linton, *Ann. Statist.* **23**, 1735 (1995).
26. J.P. Nielsen, O.B. Linton and P.J. Bickel, *Ann. Statist.* **26**, 215 (1998).
27. D. Pollard, *Econometric Theory* **7**, 186 (1991).
28. D. Ruppert, S.J. Sheather and M.P. Wand, *J. Amer. Statist. Assoc.* **90**, 1257 (1995).
29. P. Sasieni, *Scand. J. Statist.* **19**, 215 (1992).
30. P. Sasieni, *J. Roy. Statist. Soc. Ser. B* **54**, 617 (1992).
31. T.A. Severini and W.H. Wong, *Ann. Statist.* **20**, 1768 (1992).
32. T.A. Severini and J.G. Staniswalis, *J. Amer. Statist. Assoc.* **89**, 501 (1994).

A NESTED FRAILTY SURVIVAL MODEL FOR RECURRENT EVENTS

RENJUN MA

Department of Mathematics and Statistics, and Canadian Research Institute for Social Policy, University of New Brunswick, Fredericton, Canada, E3B 5A3
E-mail: renjun@unb.ca

J. DOUGLAS WILLMS

Canadian Research Institute for Social Policy, University of New Brunswick, Fredericton, Canada, E3B 5A3
E-mail: willms@unb.ca

RICHARD T. BURNETT

Environmental Health Directorate, Health Canada, Ottawa, Canada, K1A 0L2
E-mail:rick_burnett@hc-sc.gc.ca

We consider a nested frailty Cox proportional hazards model for recurrent events. We adopt a Poisson modelling approach and the principal results depend only on the first and second moments of the unobserved frailties. The use of the proposed methods is illustrated through the reanalysis of chronic granulotomous disease data previously reported by Fleming and Harrington [5].

1 Introduction

Recurrent events often arise naturally in practice. Examples of such events include disease onsets and emergency room visits in medical studies, process stoppages in reliability studies, as well as criminal offences and unemployment in social studies. Such recurrent events are often clustered by subjects and their residential areas or other groupings. Until recently, the recurrent events are usually modelled by incorporating a single level of subject frailties, or random effects, into the survival models, while ignoring other level of clustering effects (Yau [16]). Furthermore, the existing approaches to frailty survival models generally rely on specific distributional assumption of frailties such as gamma, inverse Gaussian and log-normal (Sastry [13], Yau [16]).

In this paper, we consider a nested frailty Cox proportional hazards model for recurrent events where the nested frailties are specified only up to the first and second moments. The treatment of ties and stratification is explicitly incorporated in our approach in the same way as in the standard Cox model. Our estimation procedure is based on a characterisation of this frailty Cox model as an auxiliary random effects Poisson regression model (Ma *et al.* [11]).

This approach gives optimal and consistent parameter estimators based on the orthodox best linear unbiased predictor approach to the auxiliary random effects Poisson models. For a single level of gamma frailty, given the frailty parameter, our approach coincides with the hierarchical likelihood approach (Ha *et al.* [6]) and the EM algorithm approach (Klein[10], Nielsen *et al.* [12] and Andersen *et al.* [2]).

We introduce the nested frailty Cox proportional hazards model and its auxiliary random effects Poisson models in sections 2 and 3, respectively. In section 4, we discuss the estimation of the nested frailty Cox models, using the orthodox best linear unbiased predictor approach to the auxiliary random effects Poisson models. An application of our approach to the chronic granulotomous disease data is illustrated in section 5. Some alternative frailty models are discussed in section 6. Some comparisons of our approach with others in the literature are discussed in section 7.

2 Nested Frailty Cox Model

In this section, we consider a Cox model with two levels of nested frailties. Suppose that the cohort of interest is composed of m independent clusters indexed by i. Within the ith cluster, there are J_i correlated subjects indexed by (i,j). Specifically, we assume that the cluster-level frailties $U_1,\ldots,U_m$ are positive, independent and identically distributed with

$$E(U_i) = 1 \text{ and } var(U_i) = \sigma^2,\ i = 1,...,m. \tag{1}$$

We also assume that there is a subject frailty U_{ij} associated with the subject j from the ith cluster, $j = 1,...,J_i; i = 1,\ldots,m$. We further assume that, given the cluster-level frailties $U_* = u_* = (u_1,\ldots,u_m)$, the subject-level frailties $\{U_{ij}\}, j = 1,...,J_i; i = 1,\ldots,m$ are positive and conditionally independent, and that the conditional distribution of U_{ij}, given $U_* = u_*$, depends on u_i only, with

$$E(U_{ij}|U_* = u_*) = u_i \text{ and } var(U_{ij}|U_* = u_*) = \nu^2 u_i, \tag{2}$$

$j = 1,\ldots,J_i$ and $i = 1,\ldots,m$. Furthermore, assume that there are n_{ij} recurrent events within the subject (i,j) where the kth recurrent event time of the $j$the subject in the ith cluster is given by

$$t_{ijk} = \begin{cases} \text{time to infection/censoring} & \text{first episode } (k=1) \\ \text{time to infection/censoring} & \text{since latest event } (k>1) \end{cases}$$

Let the hazard function for the kth recurrent event time of the jth subject in the ith cluster at time t be denoted by $\lambda_{ijk}(t)$. Given both the cluster

and subject level frailties $U = u$, we assume that the hazard functions are conditionally independent, with

$$\lambda_{ijk}(t) = \lambda_0^{(k)}(t) u_{ij} \exp\{\beta^\top x_{ijk}(t)\}, \tag{3}$$

where $k = 1, 2, \ldots, a$ with a being the maximum number of observed recurrences of the event, where $\lambda_0^{(k)}(t)$ denotes the unspecified baseline hazard function (Cox and Oakes [4]) and $x_{ijk}(t)$ denotes the time-dependent covariate vector for the kth recurrent event time of subject i, j at time t. The components of this time-dependent covariate vector $x_{ijk}(t)$ will be approximated by piecewise constants described in the next section. The distribution of frailties is assumed not to depend on the regression parameter β. Clearly, the recurrent event times, either observed or censored, within the same cluster are correlated. An implication of this model is that the subject n_{ij} is not at risk of $(k+1)$th event until this subject has experienced the kth event. A Cox proportional hazards model with a single level of frailties is obtained as a special case of the Cox model with two levels of frailties by setting $\nu^2 = 0$ and $J_i = 1$ for all i.

Our model only requires the specification of the first two moments of frailties since it is generally impractical to assume that the random mechanism by which the unobserved frailties were generated is entirely known. On the other hand, our assumptions (1) and (2) on random effects do cover a wide range of frailty distributions including gamma, inverse Gaussian and log-normal.

3 Auxiliary Random Effects Poisson Models

Let S_k be the set such that $(i,j) \in S_k$ if and only if the subject (i,j) has experienced his $(k-1)$th recurrent event time. Let $\tau_{k1} < \ldots < \tau_{kq_k}$ denote the distinct kth recurrent event times in S_k, with m_{kh} indicating the multiplicity of kth recurrent events occurring at time τ_{kh} $(k = 1, \ldots, a)$. The risk set at time τ_{kh} is a subset of S_k such that $\mathcal{R}(\tau_{kh}) = \{(i,j) : t_{ijk} \geq \tau_{kh}\}$, where t_{ijk} is the kth observed recurrent event time for the jth subject in the ith cluster. In addition, for any $(i,j) \in S_k$, let $Y_{ijk,h}$ be 1 if the kth recurrent event occurs for subject (i,j) at time τ_{kh} and 0 otherwise. Let Y and U denote the vectors of the $Y_{ijk,h}$ and the random effects (frailties) U_{ij}, respectively. Given the random effects $U = u$, Peto's version of the conditional partial likelihood (Cox & Oakes, 1984, p.103) is

$$\ell_p(\beta; Y|u) = \prod_{k=1}^{a} \prod_{h=1}^{q_k} \frac{\prod_{(i,j)\in\mathcal{R}(\tau_{kh})} u_{ij}^{Y_{ijk,h}} \{\exp(\beta^\top x_{ijk}(\tau_{kh}))\}^{Y_{ijk,h}} (m_{kh}!)}{\{\sum_{(i,j)\in\mathcal{R}(\tau_{kh})} u_{ij} \exp(\beta^\top x_{ijk}(\tau_{kh}))\}^{m_{kh}}}, \tag{4}$$

where the time-dependent covariate $x_{ijk}(t)$ is approximated by a piecewise constant function $x_{ijk}(t) = x_{ijk}(\tau_{kh})$ $t \in (\tau_{k(h-1)}, \tau_{kh}]$ for $h = 1, \ldots, q_k$ with $\tau_{k0} = 0$. The value of $x_{ijk}(\tau_{kh})$ can be taken as any reasonable quantity such as the median of $x_{ijk}(t)$ over the interval $(\tau_{k(h-1)}, \tau_{kh}]$.

We now define an auxiliary random effects Poisson regression model. Assume that the components of Y are conditionally independent, given random effects $U = u$, with

$$Y_{ijk,h} \sim \text{Poisson}\{u_{ij} \exp(\alpha_{kh} + \beta^\top x_{ijk}(\tau_{kh}))\} \ (i,j) \in \mathcal{R}(\tau_{kh}). \tag{5}$$

This auxiliary random effects Poisson model extends that of Whitehead (1980) for standard Cox models to frailty Cox models. Given the random effects, the conditional likelihood for the random effects Poisson model is

$$\ell(\alpha, \beta; Y|u) = \prod_{k=1}^{a} \prod_{h=1}^{q_k} \frac{\prod_{(i,j)\in\mathcal{R}(\tau_{kh})} u_{ij}^{Y_{ijk,h}} \{\exp(\alpha_{kh} + \beta^\top x_{ijk})\}^{Y_{ijk,h}}}{\exp\{\sum_{(i,j)\in\mathcal{R}(\tau_{kh})} u_{ij} \exp(\alpha_{kh} + \beta^\top x_{ijk}(\tau_{kh}))\}}. \tag{6}$$

It has been shown by Ma *et al.* [11] that

$$\ell(\hat{\alpha}, \hat{\beta}; Y, U) = \prod_{k=1}^{a} \left\{ \prod_{h=1}^{q_k} \frac{m_{kh}^{m_{kh}} \exp(-m_{kh})}{m_{kh}!} \right\} \ell_p(\hat{\beta}; Y, U),$$

where the term in parentheses on the right-hand side does not depend on the parameters of interest. This demonstrates that the maximum joint Poisson likelihood estimators for the regression parameter vector β from (6) are the maximum joint partial likelihood estimators for the regression parameter vector β from (4). We may therefore make inferences on the frailty Cox models by fitting random effects Poisson models.

In the remainder of this paper, we focus on the nested frailty Cox proportional hazards models specified by (1), (2) and (3) by way of fitting the auxiliary nested random effects Poisson models specified by (1), (2) and (5).

4 Orthodox Best Linear Unbiased Predictor Approach

4.1 Prediction of random effects

We predict the random effects by the orthodox best linear unbiased predictor of U given Y. If we let U and Y be random vectors with finite second moments, the orthodox best linear unbiased predictor of U given Y is

$$\widehat{U} = E(U) + \text{Cov}(U, Y)(\text{Cov}(Y))^{-1} \{Y - E(Y)\}.$$

This is the linear unbiased predictor of U given Y which minimises the mean squared distance between the random effects U and their predictor within the class of linear functions of Y.

The cluster random effects predictor can be expressed as

$$\widehat{U}_i = \frac{1+\sigma^2 \sum_{k=1}^{a} \sum_{h=1}^{q_k} \sum_{(i,j)\in\mathcal{R}(\tau_{kh})} w_{ij} Y_{ijk,h}}{1+\sigma^2 \sum_{k=1}^{a} \sum_{h=1}^{q_k} \sum_{(i,j)\in\mathcal{R}(\tau_{kh})} w_{ij} \mu_{ijk,h}}, \tag{7}$$

where (i,j) runs over the risk set $\mathcal{R}(\tau_{kh})$ for any given i. Here,

$$\begin{aligned}\mu_{ijk,h} &= \exp\left(\alpha_{kh} + \beta^\top x_{ijk}\right) \\ &= \exp\left\{(\alpha^\top, \beta^\top) x_{ijk,h}\right\},\end{aligned}$$

and, for fixed (i,j),

$$w_{ij} = \left(1+\nu^2 \sum_{k=1}^{a} \sum_{h=1}^{q_k} \sum_{(i,j)\in\mathcal{R}(\tau_{kh})} \mu_{ijk,h}\right)^{-1}.$$

The right-hand side of the equation (7) clearly involves unknown quantities γ, σ^2 and ν^2. These quantities together with random effects predictors will be iteratively estimated. At each iteration, the $\widehat{U}_i$ will be updated by evaluating those unknown quantities on the right-hand sides at their current values. The detailed discussion on the iterative algorithm will be presented in section 4.4. The unknown quantities involved in the following equations will be understood similarly. The subject random effects predictors are

$$\widehat{U}_{ij} = w_{ij}\hat{U}_i + \nu^2 w_{ij} \sum_{k=1}^{a} \sum_{h=1}^{q_k} \sum_{(i,j)\in\mathcal{R}(\tau_{kh})} Y_{ijk,h}. \tag{8}$$

4.2 Estimation of regression parameters

Consider first estimation of the regression parameters in the case of known dispersion parameters. Estimation of unknown dispersion parameters will be discussed in section 4.3. Differentiating the joint log-likelihood of the auxiliary model for the data and random effects yields the joint score function. Replacing the random effects with their predictors, we have an unbiased estimating function for the regression parameters $\gamma = (\alpha^\top, \beta^\top)^\top$:

$$\begin{aligned}\psi(\gamma) &= \sum_{k=1}^{a} \sum_{h=1}^{q_k} \sum_{(i,j)\in\mathcal{R}(\tau_{kh})} x_{ijk,h} \left\{Y_{ijk,h} - \widehat{U}_{ij}(\gamma)\mu_{ijk,h}(\gamma)\right\} \\ &= \sum_{i=1}^{m} \psi_i(\gamma),\end{aligned}$$

This estimating function $\psi(\gamma)$ is then a vector function of the same dimension as that of γ. The second equality is simply a rearrangement of terms under summation by clusters where

$$\psi_i(\gamma) = \sum_{k=1}^{a} \sum_{h=1}^{q_k} \sum_{(i,j)\in\mathcal{R}(\tau_{kh})} x_{ijk,h} \left\{ Y_{ijk,h} - \widehat{U}_{ij}(\gamma)\mu_{ijk,h}(\gamma) \right\}$$

corresponds to the ith cluster. For each i, it is clear from (7) and (8) that the random effect predictor $\widehat{U}_{ij}(\gamma)$ is unbiased and involves only the induced observations, $y_{ijk,h}$, within the ith cluster, and so is the estimating function $\psi_i(\gamma)$. Therefore the unbiased estimating functions $\psi_1(\gamma), ..., \psi_m(\gamma)$ are independent. The solutions of the estimating equation $\sum_{i=1}^{m} \psi_i(\gamma) = 0$ provide estimators of the regression parameters. Noting that the dimension of parameter γ increases with the number of clusters, the standard estimating equations theory based on the fixed number of parameters would apply if the dimension of parameter γ is bounded. Otherwise, it can be shown that, under some regularity conditions, the component-wise asymptotically normality of parameter estimator, $\hat{\gamma}$, remains valid if the dimension of parameter γ grows slowly (He and Shao [7]). Specifically we have, for any scalar vector b of dimension of γ such that $b^\top b = 1$, $b^\top\hat{\gamma}$ is asymptotically normal with asymptotic mean $b^\top\gamma$ and asymptotic variance given by $b^\top S(\gamma)^{-1} V(\gamma) S(\gamma)^{-\top} b$. Here, the sensitivity matrix $S(\gamma)$ and the variability matrix $V(\gamma)$ are given by

$$S(\gamma) = \sum_{i=1}^{m} S_i(\gamma) = \sum_{i=1}^{m} E_\gamma \left\{ \frac{\partial \psi_i(\gamma)}{\partial \gamma^\top} \right\},$$

$$V(\gamma) = \sum_{i=1}^{m} V_i(\gamma) = \sum_{i=1}^{m} E_\gamma \left\{ \psi_i(\gamma)\psi_i^\top(\gamma) \right\}.$$

It has been verified that $S(\gamma) = -V(\gamma)$ for the nested random effects Poisson model (Ma *et al.* [11]); therefore the asymptotic variance of $b^\top\hat{\gamma}$ is simply $-b^\top S(\gamma)^{-1} b$. If the dimension of parameter γ is bounded, the regression parameter estimator, $\hat{\gamma}$, itself can be shown to be asymptotically normal with asymptotic mean γ and asymptotic variance given by $-S(\gamma)^{-1}$ under mild regularity conditions (Artes and Jørgensen [3]). The estimating function $\sum_{i=1}^{m} \psi_i(\gamma)$ can easily be shown to be optimal in the sense that it attains the minimum asymptotic covariance for the estimator $b^\top\hat{\gamma}$ among a certain class of linear functions of Y.

This estimating equation $\sum_{i=1}^{m} \psi_i(\gamma) = 0$ can be solved by the Newton scoring algorithm introduced by Jørgensen *et al.* [9]:

$$\gamma^* = \gamma - S^{-1}(\gamma)\psi(\gamma),$$

where γ^* denotes the updated value for γ. Here negative sensitivity matrix plays a role similar to that of the Fisher information matrix in the Fisher scoring algorithm. In fact, the negative sensitivity matrix here corresponds to the so-called Godambe information matrix. The exact expression of the sensitivity matrix is given by

$$S(\gamma) = \sum_{i=1}^{m} c_i e_i e_i^\top + \sum_{i=1}^{m}\sum_{j=1}^{J_i} \nu^2 w_{ij} f_{ij} f_{ij}^\top - \sum_{k=1}^{a}\sum_{h=1}^{q_k}\sum_{(i,j)\in\mathcal{R}(\tau_{kh})} \mu_{ijk,h} x_{ijk,h} (x_{ijk,h})^\top, \tag{9}$$

where

$$e_i = \left(\sum_{k=1}^{a}\sum_{h=1}^{q_k}\sum_{(i,j)\in\mathcal{R}(\tau_{kh})} w_{ij}\mu_{ijk,h} x_{ijk,h} \right), \tag{10}$$

$$f_{ij} = \left(\sum_{k=1}^{a}\sum_{h=1}^{q_k}\sum_{(i,j)\in\mathcal{R}(\tau_{kh})} \mu_{ijk,h} x_{ijk,h} \right).$$

Here, the index (i,j) in (10) runs over the risk set $\mathcal{R}(\tau_{kh})$ for fixed i, and (i,j) runs freely over the risk set $\mathcal{R}(\tau_{kh})$ in the last term of (9). In addition, $c(i)$ denotes the mean squared distances between the random effect U_i and its predictor, specifically

$$c_i = E(\widehat{U}_i - U_i)^2 = \frac{\sigma^2}{1 + \sigma^2 \sum_{k=1}^{a}\sum_{h=1}^{q_k}\sum_{(i,j)\in\mathcal{R}(\tau_{kh})} w_{ij}\mu_{ijk,h}},$$

where (i,j) runs over the risk set $\mathcal{R}(\tau_{kh})$ for fixed i.

An analog of Wald's test is available for testing the hypothesis $H_0 : \beta_{(1)} = 0$, where $\beta_{(1)}$ is a subvector of β. The test statistic is

$$W = \widehat{\beta}_{(1)}^\top \left\{ \mathrm{J}^{11}(\widehat{\gamma}) \right\}^{-1} \widehat{\beta}_{(1)},$$

where $J^{11}(\widehat{\gamma})$ is the block of the asymptotic covariance matrix of $\widehat{\gamma}$ corresponding to $\beta_{(1)}$. Asymptotically, W follows a $\chi^2(k)$ distribution, where k is the size of the subvector $\widehat{\beta}_{(1)}$.

4.3 Estimation of dispersion parameters

We now suppose that the dispersion parameters are unknown. By analogy with generalized linear models, we adopt the following adjusted Pearson estimator for the dispersion parameter σ^2:

$$\hat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{m}\left\{(\hat{U}_i - 1)^2 + c_i\right\},$$

the first term being the Pearson estimator and the second term being a bias correction term.

The corresponding adjusted Pearson estimator for ν^2 is

$$\hat{\nu}^2 = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{J_i}\sum_{j=1}^{J_i}\left\{(\widehat{U}_{ij} - \widehat{U}_i)^2 + c_{ij} + c_i - 2c_i w_{ij}\right\}.$$

Again, the first term is the Pearson estimator and the remaining terms are bias correction terms where c_{ij} denotes the mean squared distance between the random effect U_{ij} and its predictor, specifically

$$\begin{aligned} c_{ij} &= E(\widehat{U}_{ij} - U_{ij})^2 \\ &= w_{ij}\left(\nu^2 + c_i w_{ij}\right). \end{aligned}$$

These dispersion parameter estimators can also be shown to be consistent as $m \to \infty$. Unlike previous approaches in the literature, the asymptotic variance of our regression parameter estimator is not affected by variability in the dispersion parameter estimators.

4.4 Computational procedures

Due to the lack of closed-form solutions for those unknown parameters, the computational algorithms for random effects modeling are inevitably iterative. Initial values for the regression parameters are taken as the regression parameter estimates obtained from standard Poisson regression techniques assuming independent responses, $Y_{ijk,h}$. Initial random effects predictions $\hat{U}_i$ and $\hat{U}_{ij}$ are given by the average of the responses within cluster i divided by the average of all responses, and the average of the responses within subject (i,j) divided by the average of all responses, respectively. The initial dispersion parameter estimates are calculated from the adjusted Pearson estimators, omitting the bias-correction terms.

After initial values are given, the algorithm then iterates between updating the regression parameter estimates via the Newton scoring algorithm,

updating random effect predictors via the orthodox best linear unbiased predictor, and updating dispersion parameter estimates via the adjusted Pearson estimators until the algorithm converges. At each iteration, the left-hand sides in the above equations are updated by evaluating those unknown quantities on the right-hand sides at their current values.

5 An example

Fleming and Harrington [5] presented the chronic granulotomous disease (CGD) data of 128 patients from 14 medical clusters. These patients were randomized to either interferon gamma (rIFN-g, trt=1) or placebo (trt=0). Other variables such as age and sex were also recorded. The disorders were characterized by recurrent infections and the maximum number of infections observed was 7. Fourteen of the 63 patients in the treatment group had at least one infection and the total number of infections in this group was 20. On the other hand, 30 of the 65 patients in the placebo group had at least one infection and the total number of infections in the placebo group was 56. The question is if the treatment is effective in reducing the frequency of serious infections in patients. This problem is complicated by the possible intra-subject correlation and clustering effects of medical centres on these patients.

To account for possible centre and subject effects, we consider four versions of the Cox models with both centre and subject level frailties specified by (1), (2) and (3). The treatment effect was taken as the only covariate. These four models are the standard Cox model without frailties ($\sigma^2 = \nu^2 = 0$), centre frailty Cox model ($\nu^2 = 0$), subject frailty Cox model ($\sigma^2 = 0$) and the Cox model with both centre and subject frailties, respectively. The parameter estimates corresponding to these four models are displayed in the first, second, third and last row of table 1.

This table shows that the dispersion parameter estimates for both centre and subject frailties in the fourth model are essentially zero. That is, there is little centre or subject effect. This conclusion has also been confirmed by the results based on the Cox models with a single level of frailties at either centre or subject level. This result is not surprising. In the literature, Therneau and Grambsch [14] found that there is little subject effect based on the Cox model with a single level of gamma frailties and Yau [16] found that there is little centre effect, but some subject effect based on the Cox model with both centre and subject level lognormal frailties.

Table 1. Parameter estimates for the chronic granulotomous disease data.

Cox model	Treatment		Dispersion parameter	
	β	s.e.	Center	Subject
Standard	−0.8759	0.2782		
Center level	−0.8759	0.2782	0	
Subject level	−0.8759	0.2782		0
Two levels	*−0.8771*	*0.2785*	*0.0013*	*0.0009*

6 Alternative Models

We analyzed the CGD data based on our model specified by (1), (2) and (3) where recurrent events times are stratified on the basis of event type, that is, the number of recurrences. More specifically, all the first event times form the first stratum, all the second event times form the second stratum, and so on. It implies that a patient's risk of the next infection might have been changed after he experienced his latest infection. If the occurrence of each infection has permanently compromised the immune system, the risk of the subsequent infections will be increased. However, some clinical scientists believe that the risk of recurrent CGD infection remains unchanged regardless of the number of previous infections (Therneau and Grambsch [14]). The data can then be modeled by assuming that all the event times share the same baseline hazard as follows (Andersen and Gill [1]):

$$\lambda_{ijk}(t) = \lambda_0(t) u_{ij} \exp\{\beta x_{ijk}(t)\}, \tag{11}$$

where the centre and subject frailties are still specified by (1) and (2). We fit the CGD data with the above model specified by (1), (2) and (11) with treatment as the sole covariate. The centre and subject level dispersion parameter estimates are 0.097 and 0.545, respectively. This result indicates some subject effects and relatively small centre effects. The regression parameter estimates standard error for treatment effect are now −1.1183 and 0.30465 in comparison with −1.0829 and 0.26763 for the standard Cox model without frailty. The standard error has increased slightly.

The difference between (3) and (11) is that the recurrent event times are stratified on the basis of the number of recurrences in the former model, but not in the latter one. So the latter model can be mathematically viewed as a special case of (3) by setting $a = 1$. However, there is some conceptual difference between these two models in terms of risk sets. Suppose a patient has experienced his $(k+1)$th recurrent event at time τ, then every patient who is under observation at time τ is at risk at that time in the latter model. But in the former model, among all subjects who are under observation at

time τ, only those who have experienced their kth recurrent event, but have not yet experienced their $(k+1)$th recurrent event are at risk at time τ.

The extension of both models to include extra stratification based on certain external variables is straightforward. Suppose these models are extended to include extra stratification based on, say gender; therefore the former model is still in the form of (3), but with $2a$ strata based on both gender and event type being indexed by $k = 1, \ldots, 2a$. On the other hand, the latter model is also in the form of (3), but with only 2 strata based on gender being indexed by $k = 1, 2$.

7 Discussion

Yau [16] has recently proposed a best linear unbiased predictor approach to the Cox model with nested log-normal frailties. His approach was a generalization of the linear mixed model equations of Henderson [8] where the estimating equations are obtained by differentiating the joint partial likelihood with respect to the regression parameters and the frailties. It is an analytically tractable analog of best linear unbiased predictors; therefore the corresponding frailties predictors are still called best linear unbiased predictors (Yau [16]). However, these frailty predictors actually correspond to the conditional mode of the frailty given the data, and are thus neither linear nor unbiased in general for non-normal distributions. In fact, these pseudo best linear unbiased predictors often lead to biased estimating equations and inconsistent parameter estimators. This approach remains controversial, especially for cases where the number of centres increases with the sample size (Ha *et al.* [6]).

On the other hand, our approach to the nested frailty Cox model for recurrent events is based on a truly best linear unbiased predictor approach to the random effects Poisson models. Our approach gives unbiased estimating equations and leads to optimal and consistent parameter estimators. The extension of our approach to the Cox models with more levels of nesting effects is also straightforward.

Acknowledgments

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada and in part by the Geomatics for Informed Decisions network of Canada. The authors gratefully acknowledge the computing support of Dr. Edward Hughes. The authors are also very grateful to Drs. Xuming He and Qiman Shao for their further clarification on their asymptotic results on parameters of increasing dimension.

References

1. P.K. Andersen and R.D. Gill, *Ann. Statist.* **10**, 1100 (1982).
2. P.K. Andersen, J.P. Klein, K.M. Knudsen and P.R. Tabaneray, *Biometrics* **53**, 1475 (1997).
3. R. Artes and B. Jørgensen, *Scand. J. Statist.* **27**, 321 (2000).
4. D.R. Cox and D. Oakes, *Analysis of Survival Data* (Chapman and Hall, New York, 1984).
5. T.R. Fleming and D.P. Harrington, *Counting Processes and Survival Analysis* (Wiley, New York, 1991).
6. I.D. Ha, Y. Lee and J.K. Song, *Biometrika* **88**, 233 (2001).
7. X. He and Q. Shao, *J. Multi. Anal.* **73**, 120 (2000).
8. C.R. Henderson, *Biometrics* **31**, 423 (1975).
9. B. Jørgensen, S. Lundbye-Christensen, X.-K. Song and L. Sun, *Statist. Med.* **15**, 823 (1996).
10. J.P. Klein, *Biometrics* **48**, 795 (1992).
11. R. Ma, D. Krewski and R. Burnett, *Random effects Cox model: a Poisson modelling approach* (Tech. Rep. No 338. Laboratory for Research in Statistics and Probability, Carleton University, Canada, 2000).
12. G.G. Nielsen, R.D. Gill, P.K. Andersen and T.I.A. Sørensen, *Scand. J. Statist.* **19**, 25 (1992).
13. N. Sastry, *J. Am. Statist. Assoc.* **92**, 426 (1997).
14. T.M. Therneau, and P.M. Grambsch, *Modeling Survival Data: Extending the Cox Model* (Springer-Verlag, New York, 2000).
15. J. Whitehead, *Appl. Statist.* **29**, 268 (1980).
16. K.K.W. Yau, *Biometrics* **57**, 96 (2001).

MULTIPLE COMPARISON PROCEDURES FOR LINEAR MODELS UNDER ORDER RESTRICTIONS

HOSSEIN MANSOURI AND ROBERT PAIGE
Department of Mathematics and Statistics, Texas Tech University
Lubbock, TX 79409-1042
E-mail: rpaige@koch.math.ttu.edu

Multiple comparison procedures are some of the most frequently use statistical methods. Although, there exists an extensive amount of literature on multiple comparison procedures, few articles have addressed the problem of multiple comparisons under order restrictions. Currently available procedures are mainly developed for the one-way layouts and hence are not applicable to more complex designs such as unbalanced factorial designs or analysis of covariance models. To achieve these extensions, the focus of the present investigation is to develop simultaneous multiple comparison procedures for linear models under order restrictions of the parameters.

1 Introduction

Consider the linear model

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{Y}$ is an $n \times 1$ observation vector, X is an $n \times p$ known matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of independent and identically distributed random variables having a normal distribution with mean 0 and variance $\sigma^2 > 0$.

A common problem of interest is to make simultaneous inferences about components of a linear function $C\boldsymbol{\beta}$, where C is a $r \times p$ matrix with known elements. Focusing on simultaneous testing of the constrained parameters, suppose that we are interested in testing the hypothesis

$$H_0 : C\boldsymbol{\beta} = \boldsymbol{d} \tag{2}$$

against the one-sided alternatives $H_1 - H_0$, where

$$H_1 : C\boldsymbol{\beta} \geq \boldsymbol{d} \tag{3}$$

Without loss of generality it can be assumed that $\boldsymbol{d} = \mathbf{0}$.

Depending on dimensionality of C, we are dealing with a different class of tests. For instance, for testing subhypotheses against ordered alternatives in regression models or testing for main effects in factorial analysis of variance models, then $r \leq p$, and for simultaneous one-sided multiple comparison

procedures $r > p$. A third class of tests which is the main focus of this investigation, is the problem of simultaneous multiple comparisons under the assumption that the parameter space is constrained a priori. Although the first two problems are well developed, at least theoretically, not much is available on the third problem.

Simultaneous tests for the order restricted (one-sided) hypotheses are well developed. Well-known tests are the likelihood ratio test, Wald's test, and Rao's efficient scores test. These are discussed in detail in Robertson *et al.* [10]. Shapiro [12] presented a unified approach by assuming the parameter space to be a closed convex cone and the observations follow a multivariate normal distribution with a known dispersion matrix. Mansouri [4] proposed a test based on the aligned ranking approach. One-sided simultaneous multiple comparison procedures for linear models, although not discussed explicitly can be derived from the results in Hochberg and Tamhane [3] and is outlined in Westfall *et al.* [13].

Simultaneous multiple comparison procedures under order restrictions of the parameters has not received much attention in the literature. The existing techniques are mainly developed for the one-way layouts under the assumption of simple order restriction of the means. Marcus and Peritz [8] proposed one-sided Scheffé-type simultaneous confidence intervals for monotone contrasts in the means of normal distributions, Williams [14] proposed a Tukey-type range test for monotone contrasts for the balanced one-way layouts, Marcus [7] proposed a two-sided Scheffé-type simultaneous confidence intervals for monotone contrasts for the means of normal distributions. Hayter [2] argued that simultaneous inference for monotone contrasts does not extend to simultaneous pairwise comparisons and developed a one-sided studentized range technique under the assumption of simple order without the restriction on monotonicity of the contrasts. Recently, Mansouri and Shaw [6] proposed distribution-free simultaneous multiple comparison tests under the assumption of simple order for the balanced complete and incomplete randomized blocks. Mansouri [5] extended the latter results for repeated measures designs with incomplete observations.

The main focus of this article is to develop a technique for simultaneous multiple comparisons in linear models. This technique provides a procedure that is applicable to balanced as well as unbalanced designs in analysis of variance and analysis of covariance models. In addition, we consider both the one-sided inferences without assuming monotonicity of the contrasts.

In Section2, a brief summary of simultaneous one-sided inference in linear models is presented. In Section 3, simultaneous multiple comparison procedures are proposed and their distributions are discussed.

2 One-Sided Tests

In this section, a brief discussion of one-sided tests for subhypotheses, ($q = r \leq p$) and simultaneous one-sided multiple comparisons, ($r > q, q \leq p$) will be presented.

2.1 One-Sided Test for Subhypotheses

Let $X = (X_1, X_2)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, where X_1 is $n \times q, X_2$ is $n \times p - q, \boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $q \times 1$ and $(p - q) \times 1$, respectively. Then the linear model in (1), can be written as

$$Y = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{4}$$

Without loss of generality, the linear hypotheses in (2) and (3) can be written as $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ and $H_1 : \boldsymbol{\beta}_1 \geq \mathbf{0}$.

Let

$$\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2')'$$

be the unrestricted maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$.

Let

$$S = (\mathbf{Y} - X\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - X\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'X'X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

Then the likelihood ratio statistic is equivalent to

$$\begin{aligned}
&\min_{\boldsymbol{\beta} \in H_0} S - \min_{\boldsymbol{\beta} \in H_1} S \\
&\quad = \min_{\boldsymbol{\beta} \in H_0} (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)'\Delta_{11}^{-1}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \\
&\qquad - \min_{\boldsymbol{\beta} \in H_1} (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)'\Delta_{11}^{-1}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \\
&\quad = \widehat{\boldsymbol{\beta}}_1'\Delta_{11}^{-1}\widehat{\boldsymbol{\beta}}_1 - (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)'\Delta_{11}^{-1}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) \\
&\quad = \boldsymbol{\beta}_1^{*'}\Delta_{11}^{-1}\boldsymbol{\beta}_1^*
\end{aligned}$$

where $\boldsymbol{\beta}_1^*$ is the MLE of $\boldsymbol{\beta}_1$ under H_1 and

$$\Delta_{11} = (X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1)^{-1}.$$

Under H_0 and known σ^2, the test statistic

$$\bar{\chi}^2 = \sigma^{-2}\boldsymbol{\beta}_1^{*'}\Delta_{11}^{-1}\boldsymbol{\beta}_1^*$$

has a $\bar{\chi}^2$-distribution given by

$$P\{\bar{\chi}^2 \geq t\} = \sum_{j=0}^{q} \omega_j P\{\chi^2 \geq t\}, \quad t > 0$$

where the weights ω_j are nonnegative satisfying $\sum_{j=0}^{q} \omega_j = 1$ and $\chi_0^2 = 0$. Computation of these weights depend on q, Δ_{11} and the alternative hypothesis H_1. More details on ω_j's are provided in Shapiro [12] sec. 5. For the one-way layouts, ω_j's are called level probabilities and are discussed in detail in Barlow *et al.* [1].

2.2 One-Sided Multiple Comparisons

Again consider the linear model (4), and assume that we are interested in testing H_{i0} and H_{i1} where

$$H_{i0} : \boldsymbol{c}_i'\boldsymbol{\beta}_1 = \mathbf{0} \text{ and } H_{i1} : \boldsymbol{c}_i'\boldsymbol{\beta}_1 \geq \mathbf{0}, \quad i = 1, \cdots, r > q$$

Let $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2')'$ be the unrestricted MLE of $\boldsymbol{\beta}$ and let S^2 be an unbiased estimator of σ^2. Define the test statistics

$$T_i = \boldsymbol{c}_i'\widehat{\boldsymbol{\beta}}_1 / S(\boldsymbol{c}_i'\Delta_{11}\boldsymbol{c}_i)^{1/2}, \quad i = 1, \cdots, r.$$

Then we reject H_{i0} in favor of $H_{i1} - H_{i0}$ if

$$T_i > c_\alpha, \quad i = 1, \cdots, r$$

where c_α is upper α-th quantile of the distribution of $\max_i T_i$.

We note that $T_1, \cdots, T_r$ have a joint r-dimensional t-distribution with dispersion matrix $R = D^{-1/2}C'\Delta_{11}CD^{-1/2}$, where $D = diag(\boldsymbol{c}_1'\Delta_{11}\boldsymbol{c}_1, \cdots, \boldsymbol{c}_r'\Delta_{11}\boldsymbol{c}_r)$, see Westfall *et al.* (1999), chapter 5.

3 Multiple Comparisons Under Order Restrictions

Consider the linear model in (4) and assume further that the parameter space is subject to the constraint that $\boldsymbol{\beta}_1 \in \mathcal{K}$, where $\mathcal{K}$ is a closed convex cone. The simple order restriction, considered in the cited literature in Section 1 is a member of $\mathcal{K}$. Suppose that we are interested in testing H_{i0} against H_{i1} where

$$H_{i0} : \mathbf{c}_i'\boldsymbol{\beta}_1 = \mathbf{0} \text{ and } H_{i1} : \mathbf{c}_i'\boldsymbol{\beta}_1 \geq \mathbf{0}, \quad i = 1, \ldots, r > q$$

Let $\boldsymbol{\beta}_1^*$ denote the orthogonal projection of $\widehat{\boldsymbol{\beta}}_1$ onto $\mathcal{K}$, i.e. $\widehat{\boldsymbol{\beta}}_1^*$ is the solution of

$$\min_{\boldsymbol{\beta}_1 \in \mathcal{K}} \left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right)' \Delta_{11}^{-1} \left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right).$$

Define the test statistics

$$T_i^* = \mathbf{c}_i' \widehat{\boldsymbol{\beta}}_1^* / \sigma \left(\mathbf{c}_i' \Delta_{11} \mathbf{c}_i\right)^{1/2}, \quad i = 1, \ldots, r.$$

Using the Union-Intersection principle of Roy (1953), we reject H_{i0} in favor of $H_{i1} - H_{i0}$ if

$$T_i^* > c_\alpha, \quad i = 1, \cdots, r$$

where c_α is upper α-th quantile of the distribution of $\max_i T_i^*$. By converting the simultaneous testing problem to simultaneous confidence region, we obtain simultaneous upper confidence limits with family confidence coefficient $1 - \alpha$ given by

$$\boldsymbol{c}_i' \boldsymbol{\beta}_1 \leq \boldsymbol{c}_i' \widehat{\boldsymbol{\beta}}_1^* - c_\alpha \sigma (\boldsymbol{c}_i' \Delta_{11} \boldsymbol{c}_i)^{1/2}, \quad i = 1, \cdots, r$$

To find the critical values of the sampling distribution of $\max_i T_i^*$, in most cases one has to resort to simulation techniques. This is a common practice in testing under order restrictions as well as situations where simultaneous multiple comparison techniques are used. This is the case because in both inferential approaches, except for the balanced analysis of variance designs, the sampling distribution of the resulting statistic is intractable, hence there is no choice but to resort to simulation. Similarly, if we are testing H_{i0} against the alternative

$$H_{i1}' : \boldsymbol{c}_i' \boldsymbol{\beta}_1 \neq \mathbf{0}, \quad i = 1, \cdots, r > q$$

we reject H_{i0} if $|T_i^*| > d_{\alpha/2}$, where d_α is the upper α-th quantile of the distribution of $\max_i |T_i^*|$. Finally, a simultaneous confidence region with a family confidence coefficient $1 - \alpha$ is given by

$$\boldsymbol{c}_i' \boldsymbol{\beta}_1 \in \{\boldsymbol{c}_i' \widehat{\boldsymbol{\beta}}_1^* \pm d_{\alpha/2} \sigma (\boldsymbol{c}_i' \Delta_{11} \boldsymbol{c}_i)^{1/2}\}, \quad i = 1, \cdots, q.$$

To simulate the null distribution of the test statistics, the following algorithm may be used.

1. Generate $\varepsilon_1, \cdots, \varepsilon_n$ from a standard normal distribution.

2. Generate $Y_1, \cdots, Y_n$ using the model $\boldsymbol{Y} = X_R\boldsymbol{\beta}_R + \boldsymbol{\varepsilon}$ which is the reduced model under the $H_0 = \bigcap_i H_{i0}$.

3. Calculate $T_1^*, \cdots, T_r^*$ $(|T_1^*|, \cdots, |T_r^*|)$.

4. Calculate $\max_{1\le i\le r} \max_{1\le i\le r} |T_i^*|T_i^*$ for the one-sided ($\max_{1\le i\le r} |T_i^*|$ for the two-sided) simultaneous confidence interval.

5. Repeat the process M times.

6. Find $c_\alpha(d_{\alpha/2})$ the upper α-th ($\alpha/2$-th)quantile of $\max_{1\le i\le r} T_i^*$ $\max_{1\le i\le r} |T_i^*|$ based on M simulated values.

4 Example

Our dataset comes from Montgomery [9](pp. 170). Here, an engineer is studying the effect of cutting speed on the rate of metal removal in a machining operation. However, the rate of metal removal is also related to the hardness of the test specimen. Five observations are taken at each cutting speed. The amount of metal removed (y) and the hardness of the specimen (x) are shown in the following table:

Cutting Speed (rpm)					
1000		1200		1400	
x	y	x	y	x	y
68	120	112	165	118	175
90	140	94	140	82	132
98	150	65	120	73	124
77	125	74	125	92	141
88	136	85	133	80	130

We assumed analysis of covariance model

$$y_{ij} = \mu + \tau_i + \gamma\,(x_{ij} - \bar{x}..) + \varepsilon_{ij} \quad \text{for } i = 1, 2, 3 \text{ and } j = 1, \ldots, 5$$

where $\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$. Since the rate of metal removal can be expected to be monotonically increasing in cutting speed it is conceivable that

$$\tau_1 \le \tau_2 \le \tau_3.$$

We then considered the following sets of hypotheses:

$$\begin{array}{ll} H_{10} : \tau_2 - \tau_1 = 0 & H_{11} : \tau_2 - \tau_1 \ge 0 \\ H_{20} : \tau_3 - \tau_1 = 0 & H_{21} : \tau_3 - \tau_1 \ge 0 \\ H_{30} : \tau_3 - \tau_2 = 0 & H_{31} : \tau_3 - \tau_2 \ge 0 \end{array} \tag{5}$$

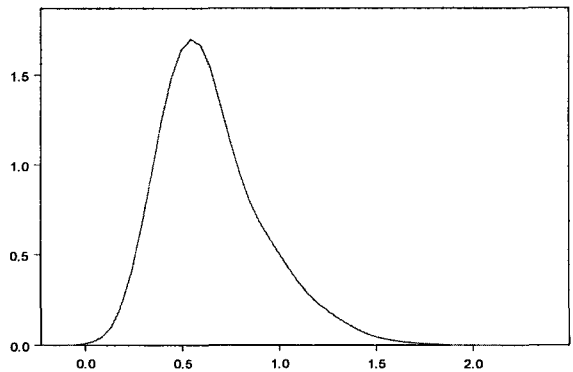


Figure 1: Density Plot of $\max_i T_i^*$

Next, identifiability constraint $\sum_i \tau_i = 0$ was imposed. Under this, constraint our omnibus null hypothesis was

$$H_0 : \tau_1 = \tau_2 = \tau_3 = 0.$$

For the reduced model, under the omnibus null, we obtained the following parameter estimates:

$$\begin{aligned} \hat{\gamma} &= 0.9299 \\ \hat{\mu} &= \bar{y} = 86.40 \\ \hat{\sigma} &= 2.744 \end{aligned}$$

We simulated from the estimated reduced model. Statistics T_1^*, T_2^* and T_3^*, associated with following hypotheses sets $(H_{10}, H_{11}), (H_{20}, H_{21})$ and (H_{30}, H_{31}) respectively, were computed and the quantity $\max_i T_i^*$ was obtained. This process was repeated 9,999 times to yield a total of 10,000 simulated $\max_i T_i^*$ values.

Next, $\hat{c}_\alpha$, an estimate for c_α, the upper α-th quantile of the $\max_i T_i^*$, was obtained by first ordering the $\max_i T_i^*$ values from smallest to largest and taking our estimate to be the realization of the $10,000\,(1-\alpha)$-th order statistic. Note that, for our example α was set at 0.05 which guaranteed an integral $10,000\,(1-\alpha)$ value. Our simulated 0.05-th quantile of the $\max_i T_i^*$ turned out to be 1.1652.

Figure 1 shows the S-Plus density plot of the simulated $\max_i T_i^*$ values.

For our example, the observed values of the T_i^*s were as follows:

i	T_i^*
1	0.4157
2	0.4237
3	0.2638

resulting in failure to reject H_{i0} for $i = 1, 2, 3$.

5 Conclusion

In this article, we have proposed a method for making simultaneous one-sided and two-sided multiple comparisons for linear models, when the parameters of interest belong to a closed convex cone. The exact distributions for the resulting statistics are derived. However, because of complexity of the distributions for all except the balanced analysis of variance designs, one must obtain the critical values by simulation. An algorithm for generating the approximate values of these quantiles is proposed. Lastly, we applied this procedure to a data set.

References

1. R.E. Barlow, D.J. Bartholomew, J.M. Bremner and H. Brunk, *Statistical inference under order restrictions* (John Wiley, New York, 1972).
2. A.J. Hayter, *Journal of the American Statistical Association* **85**, 778 (1990).
3. Y. Hochberg and A.C. Tamhane, *Multiple comparison procedures* (John Wiley, New York, 1987).
4. H. Mansouri, *Journal of Statistical Planning and Inference* **24**, 107 (1990).
5. H. Mansouri, *Distribution-free multiple comparisons for incomplete data* (Submitted for publication, 2000).
6. H. Mansouri and C. Shaw, *Nonparametric multiple comparison procedures for ordered parameters in balanced incomplete blocks* (Submitted for publication, 1999).
7. R. Marcus, *Communications in Statistics–Theory and Methods* **11**, 615 (1982).
8. R. Marcus and E. Peritz, *Journal of the Royal Statistical Society, Ser. B* **38**, 157 (1976).
9. D. Montgomery, *Design and Analysis of Experiments*, 4th edition (John Wiley, New York, 1997).

10. T. Robertson, F.T. Wright and R.L. Dykstra, *Order restricted statistical inference* (John Wiley, New York, 1988).
11. S.N. Roy, *Annals Of Mathematical Statistics* **24**, 220 (1953).
12. A. Shapiro, *International Statistical Review* **56**, 49 (1988).
13. P.H. Westfall, R.D. Tobias, D. Rom, R.D. Wolfinger and Y. Hochberg, *Multiple comparisons and multiple tests using the SAS system* (SAS Institute, Cary, NC, USA 1999).
14. D.A. Williams, *Biometrika* **64**, 9 (1977).

TESTING GOODNESS-OF-FIT OF THE GAMMA MODELS

CAROL E. MARCHETTI
Department of Mathematics and Statistics, Rochester Institute of Technology, Rochester, NY 14623, USA
E-mail: cemsma@rit.edu

GOVIND S. MUDHOLKAR AND GREGORY E. WILDING
Department of Biostatistics, University of Rochester, Rochester, NY 14642, USA
E-mail: govind@bst.rochester.edu; wilding@bst.rochester.edu

Gamma models are extensively employed in statistical analyses, especially for reliability and lifetime data. Assessing the appropriateness of these models is of obvious importance. While characterization theorems in probability and statistics are widely appreciated for their role in clarifying the structures of families of probability distributions, they are not as well recognized for being natural, logical and effective starting points for constructing tests for goodness-of-fit problems. It is well known that two independent identically distributed random variables (X_1, X_2) have a gamma distribution if, and only if, $X_1 + X_2$ and X_1/X_2 are independent. This characterization is used by Locke [18] to propose a method for testing the composite gamma hypothesis. It is based upon the random creation of n pairs from a sample of size $2n$, and using tests of bivariate independence, such as those based on the rank correlation coefficient or Kendall's tau, for testing independence of the sums and the ratios. However, the tests of bivariate independence he considered are consistent against only some dependence alternatives. Instead, we propose employing the classical rank test due to Hoeffding [9] or its asymptotic equivalent due to Blum, Kiefer and Rosenblatt [2], which are known to be consistent against all dependence alternatives. The resulting goodness-of-fit tests are consistent against all non-gamma alternatives and, in moderate size samples, offer substantial power advantages. Additionally, an alternative approach, which does not suffer from the caprice of chance implicit in the random pairing, is also considered.

1 Introduction and Summary

Characterization theorems in probability and statistics are generally well appreciated for their aesthetic appeal, mathematical completeness and the light they shed on the structures of the probability distributions. Although logically self evident, but not well recognized, is the fact that they can be natural and effective bases for constructing goodness-of-fit (GOF) tests needed for assessing the validity of models based on parametric families, such as normal, exponential, and inverse Gaussian (IG), commonly used in statistical practice.

The earliest explicit use of a characterization theorem for constructing a goodness-of-fit test is by Vasicek [37], who used Shannon's maximum entropy characterization to construct a test for the composite hypothesis of normality.

Now there exists a substantial body of literature on the goodness-of-fit tests based on characterization theorems. Maximum entropy characterizations have been used by Gokhale [7] and Mudholkar and Lin [27] to construct goodness-of-fit tests for exponentiality, and more recently by Mudholkar and Tian [30] to construct a GOF test for the inverse Gaussian model. Characterization results based on the statistical independence of sample statistics have been used by Lin and Mudholkar [15] and Mudholkar, Marchetti and Lin [28] to construct GOF tests for normality, and by Mudholkar, Natarajan and Chaubey [29] to test inverse Gaussian goodness-of-fit. For an overview of characterizations and goodness-of-fit, see Marchetti and Mudholkar [20].

The gamma family of unimodal, right-skewed distributions with nonnegative support, has been used to model a wide variety of applications in diverse fields such as geology (McCullagh and Lang [25]), ecology (Matis, Rubink, and Makela [23]), inventory control and queuing (Yeh [40]), economics (McDonald and Jensen [26]), meteorology (Bougeault [3]), reliability (Reiser and Rocke [35]), biomedical studies (Tan [36]) and genetics (Yang [39]). Testing the appropriateness of the gamma models for use in these applications is vital.

Empirical distribution function (EDF) tests for the composite gamma hypothesis have been developed by Pettitt and Stephens [34] and Lockhart and Stephens [16]. For an account of these, see D'Agostino and Stephens [4]. It is well known that the distribution of two IID random variables (X_1, X_2) is gamma if and only if $X_1 + X_2$ and X_1/X_2 are independent (see Lukacs, [19]). Locke [17],Locke76 proposed testing the composite gamma hypothesis by creating n pairs from a sample of size $2n$ and testing the independence using procedures such as the one based on Kendall's tau. Since the rank tests of bivariate independence he considered are consistent against only some dependence alternatives, the resulting gamma GOF tests lack consistency against all non-gamma alternatives.

The existing goodness-of-fit tests of the composite gamma hypotheses are reviewed in Section 2. A modification of one of these tests, together with a comparative study of its power function is presented in Section 3. The modified test, which is consistent against all non-gamma alternatives, offers a power advantage with moderate sample sizes. In Section 4, an alternative approach is considered. Section 5 is given to conclusions.

2 The Gamma Goodness-of-Fit Tests

The literature on testing goodness-of-fit for the gamma models is rather scant, and most of the proposed solutions lack simplicity for use in practice. The tests available, which may be grouped as the EDF tests and the tests based on characterization results, are now reviewed.

2.1 *The Empirical Distribution Function Tests*

The empirical distribution function for a sample $X_1, X_2, ..., X_n$ is defined by $F_n(x) = \#(X_i \leq x)/n$. To test that the sample is from a population with a distribution function $F(x) = F_o(x)$, where $F_o(x)$ is fully specified, one may measure the discrepancy between the $F_n(x)$ and $F_o(x)$.

Kolmogorov [13] introduced the first EDF statistic, D, by defining

$$D_+ = \sup_x \{F_n(x) - F_o(x)\}, \tag{1}$$

$$D_- = \sup_x \{F_o(x) - F_n(x)\}, \quad \text{and} \tag{2}$$

$$D = \sup_x |F_n(x) - F_o(x)| = \max(D_+, D_-). \tag{3}$$

The Cramér-von Mises family is a class of statistics of the form

$$Q = \int_{-\infty}^{\infty} \{F_n(x) - F_o(x)\}^2 \psi(x) dF(x). \tag{4}$$

When $\psi(x) = 1$, the Cramér-von Mises W^2 statistic is obtained; when $\psi(x) = [F(x)(1 - F(x))]^{-1}$, one obtains the Anderson-Darling [1] statistic A^2.

The EDF tests have been adapted for the composite goodness-of-fit hypothesis $F(x) = F_o(x, \boldsymbol{\theta})$, where F_o is known but $\boldsymbol{\theta}$ is not, by plugging in estimated $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$. Pettitt and Stephens [34] considered the problem of testing the composite gamma hypothesis in which the shape parameter α is known and scale parameter β is unknown. They provide asymptotic percentiles for three EDF statistics for the problem. Lockhart and Stephens [16] examined the situation of a gamma distribution with unknown shape parameter and scale parameter either known or unknown. They suggest using a maximum likelihood estimator of α and give asymptotic percentiles for the resulting EDF statistics. D'Agostino and Stephens [4] provide an excellent overview of goodness-of-fit for the gamma and other distributions.

2.2 *Characterizations and Gamma GOF Tests*

A nice summary of characterizations of the gamma family, from the earliest results due to Nabeya [33], Goodman [8], Mauldon [24] and Laha [14], appears in Johnson, Kotz and Balakrishnan [12]. The simplest of these, the following result due to Lukacs [19], is the basis of Locke's [18] test.

Proposition 2.1: Two independent random variables (X_1, X_2) have a common gamma distribution iff $X_1 + X_2$ and X_1/X_2 are independent.

Extensions, applications and reviews of the above characterization may be found in Findeisen [6], Marsaglia [21,22] and Wang [38]. Since this characterization is asymmetric in X_1 and X_2, Locke [17] considers the dependence in terms of $U = X_1 + X_2$ and $V = \max(X_1/X_2, X_2/X_1)$.

Before discussing the specifics of Locke's test we note a general maximum entropy result, from which characterizations of several members of the exponential family have been fruitfully used to develop consistent goodness-of-fit tests. The examples include the maximum entropy tests of the composite hypotheses of normality (Vasicek [37]), uniformity (Dudewicz and van der Meulen [5]), exponentiality (Gokhale [7], Mudholkar and Lin [27]) and the inverse Gaussian distribution (Mudholkar and Tian [30,31]). Gokhale [7] presents a general discussion of construction and suggests a GOF test for the gamma model, but offers no details. The following is a relatively recent characterization due to Hwang and Hu [11] which will be considered in Section 4.

Proposition 2.2: The mean $\bar{x}$ and coefficient of variation $s/\bar{x}$ of a random sample are independent iff the population is gamma.

2.3 Locke's Tests

Specifically, Locke [18] proposed testing the composite gamma hypothesis by randomly pairing observations from a sample of size $2n$, to obtain n pairs (Y_i, Z_i), where

$$Y_i = X_{2i} + X_{2i-1}, \qquad Z_i = \max\{X_{2i}/X_{2i-1}, X_{2i-1}/X_{2i}\}, \tag{5}$$

$i = 1, 2, ..., n$. Then testing the composite gamma hypothesis is equivalent to testing the independence of Y and Z. Locke employs the quadrant tests and tests based on the rank correlation coefficient and Kendall's tau for this purpose.

An obvious drawback to this method is that with random pairing, the same data set can produce different results. However, more importantly the rank tests that Locke considers are consistent against only some dependence alternatives. In the next section, we propose a modification of Locke's procedure that addresses the consistency issue.

3 A Consistent Modification of Locke's Procedure

Locke's [18] approach of randomly pairing the observations in the sample can be combined with use of the consistent tests of bivariate independence due to Hoeffding [9] and Blum, Kiefer and Rosenblatt [2] to construct gamma GOF

tests which would be consistent against all non-gamma alternatives. These modifications would be considered improvements over Locke's test provided they show power advantages with moderate size samples. In this section, we first describe the consistent tests of bivariate independence, the modification of Locke's procedure, and then present the results of a study of the power functions.

3.1 Two Consistent Tests of Bivariate Independence

Hoeffding's [9] notes that

$$D(x,y) = F_{X,Y}(x,y) - F_X(x)\,F_Y(y) = 0, \tag{6}$$

if and only if X and Y are independent. Nonparameteric estimation of the quantity $\int D^2(x,y)dF(x,y)$ results in the statistic,

$$D_n = \frac{Q - 2(n-2)R + (n-2)(n-3)S}{n(n-1)(n-2)(n-3)(n-4)}, \tag{7}$$

in which

$$Q = \sum_{i=1}^{n}(R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2), \tag{8}$$

$$R = \sum_{i=1}^{n}(R_i - 2)(S_i - 2)c_i, \tag{9}$$

$$S = \sum_{i=1}^{n}(c_i - 1)c_i, \tag{10}$$

where R_i and S_i are the respective ranks of X_i among the X's and Y_i among the Y's, and c_i is the number of bivariate observations (X_j, Y_j) for which $X_j \leq X_i$ and $Y_j \leq Y_i$. Hoeffding [9] provides the exact null distribution of nD_n for sample sizes $n = 5, 6, 7$ which is extended to $n = 8, 9$ by Hollander and Wolfe [10]. However, Hoeffding [9] was unable to obtain the asymptotic distribution of nD_n.

Blum, Kiefer and Rosenblatt [2] consider a computationally simpler statistic,

$$B_n = n^{-5}\sum_{i=1}^{n}[N_1(i)N_4(i) - N_2(i)N_3(i)]^2, \tag{11}$$

where $N_1(i)$, $N_2(i)$, $N_3(i)$, and $N_4(i)$ are the number of points lying in the four quadrants determined by the vertical and horizontal lines through the

bivariate point (X_i, Y_i). Independence is rejected for large values of the test statistic.

Blum, Kiefer and Rosenblatt [2] show that nB_n is asymptotically equivalent to Hoeffding's [9] statistic and provide the asymptotic percentiles. Mudholkar and Wilding [32] obtain the empirical distribution of the two statistics for small and moderate size samples. A selection of the 10%, 5% and 1% points of the nD_n and nB_n statistics appears in Table 1.

A Gaussian approximation for the nB_n statistic, which can be used to obtain both the associated percentiles and probabilities, is currently under development by Mudholkar and Wilding.

3.2 The Modified Test

The tests of bivariate independence due to Hoeffding [9] and due to Blum, Kiefer and Rosenblatt [2] may be used for testing the independence of Y and Z. Since the two rank tests are consistent against all dependence alternatives, it is obvious that the modified versions of Locke's procedure using these tests will produce goodness-of-fit tests which are consistent against all non-gamma alternatives.

Let $X_1, X_2, ..., X_{2n}$ be a random sample from a population with d.f. F and consider the problem of testing the composite hypothesis that $F(\cdot)$ is the d.f. of a gamma random variable. If the sample size is odd, as in Locke [18], delete an observation at random. We propose testing the composite hypothesis by randomly pairing observations from the sample, to obtain n pairs (Y_i, Z_i), as given in (5). Testing the composite gamma hypothesis is equivalent to testing the independence of Y and Z. The rank tests due to Hoeffding [9] and to Blum, Kiefer and Rosenblatt [2] are employed for this purpose.

3.3 A Monte Carlo Experiment

A Monte Carlo experiment was conducted using 100,000 simulated samples of size $n = 20, 30, 50, 100$ from a variety of non-negative populations. Each sample was subjected to the above tests of the composite gamma hypothesis which are based on Locke's premise and use tests of bivariate independence due to Kendall, Spearman, Hoeffding and Blum *et al.*

A selection of the results of the simulation experiment are shown in Table 2. In the population column, GTL stands for Generalized Tukey-Lambda family. Where parameters are unspecified, the standard distribution is used. The Spearman, Hoeffding and Blum *et al.* tests have excellent Type I Error control, but the Kendall test is not so accurate in this respect for small or even moderate sample sizes. For most of the populations, but not all, the tests

Table 1. Percentiles* (y) of statistics nD_n and nB_n.

	$P(nD_n \leq y)$			$P(nB_n \leq y)$		
n	0.90	0.95	0.99	0.90	0.95	0.99
5	0.0000	0.1667	0.1667	0.0720	0.0976	0.1408
6	0.0333	0.1000	0.2000	0.0710	0.0872	0.1435
7	0.0333	0.0667	0.1333	0.0675	0.0875	0.1312
8	0.0381	0.0619	0.1143	0.0649	0.0842	0.1274
9	0.0357	0.0571	0.1095	0.0639	0.0820	0.1250
10	0.0331	0.0529	0.1005	0.0626	0.0802	0.1218
11	0.0317	0.0500	0.0960	0.0613	0.0783	0.1195
12	0.0308	0.0485	0.0929	0.0604	0.0771	0.1172
13	0.0296	0.0466	0.0894	0.0596	0.0762	0.1163
14	0.0289	0.0455	0.0869	0.0589	0.0750	0.1137
15	0.0281	0.0443	0.0852	0.0582	0.0740	0.1123
16	0.0275	0.0433	0.0835	0.0575	0.0731	0.1110
17	0.0271	0.0424	0.0818	0.0571	0.0724	0.1097
18	0.0265	0.0418	0.0801	0.0564	0.0716	0.1087
19	0.0260	0.0411	0.0788	0.0560	0.0711	0.1080
20	0.0257	0.0405	0.0775	0.0557	0.0704	0.1072
21	0.0254	0.0401	0.0771	0.0553	0.0701	0.1064
22	0.0251	0.0396	0.0759	0.0550	0.0696	0.1057
23	0.0249	0.0393	0.0754	0.0545	0.0689	0.1044
24	0.0246	0.0389	0.0746	0.0543	0.0687	0.1042
25	0.0243	0.0385	0.0739	0.0540	0.0683	0.1033
30	0.0234	0.0371	0.0715	0.0528	0.0665	0.1006
35	0.0228	0.0360	0.0695	0.0522	0.0656	0.0995
40	0.0223	0.0354	0.0681	0.0516	0.0649	0.0979
45	0.0221	0.0349	0.0669	0.0512	0.0643	0.0968
50	0.0217	0.0345	0.0661	0.0507	0.0637	0.0958
60	0.0212	0.0337	0.0648	0.0501	0.0627	0.0940
70	0.0210	0.0334	0.0644	0.0496	0.0621	0.0932
80	0.0207	0.0330	0.0634	0.0493	0.0617	0.0923
90	0.0206	0.0326	0.0633	0.0490	0.0613	0.0918
100	0.0204	0.0326	0.0625	0.0489	0.0612	0.0916
∞	0.0192	0.0306	0.0591	0.0469	0.0584	0.0869

*$n = 5, 6, 7$, exact percentiles of nD_n adapted from Hoeffding; [9]

$n = 8, 9$, exact percentiles of nD_n adapted from Hollander and Wolfe; [10]

$n = \infty$, asymptotic percentiles interpolated from Blum, Kiefer and Rosenblatt [2] (1961); remaining percentiles empirically generated via 1,000,000 iterations

based on the consistent methods of Hoeffding and Blum *et al.* have higher power than those based on the methods of Kendall and Spearman.

Table 2. Empirical* power function estimates of the tests of the composite gamma hypothesis; $\alpha = 0.05$.

Population	n	Kendall	Spearman	Hoeffding	Blum *et al.*
Gamma	30	0.059	0.049	0.050	0.050
	50	0.047	0.049	0.050	0.050
Beta(.5,.5)	30	0.659	0.588	0.653	0.422
	50	0.863	0.833	0.916	0.838
Extreme Value	30	0.209	0.186	0.282	0.331
	50	0.309	0.310	0.549	0.573
F(1,,1)	30	0.360	0.325	0.289	0.385
	50	0.525	0.529	0.493	0.580
F(30,6)	30	0.267	0.237	0.202	0.287
	50	0.388	0.395	0.348	0.438
GTL(-2,1)	30	0.466	0.376	0.459	0.181
	50	0.678	0.600	0.767	0.515
GTL(-.5,-2)	30	0.181	0.149	0.207	0.245
	50	0.245	0.229	0.405	0.425
GTL(2.5,1)	30	0.217	0.190	0.365	0.409
	50	0.324	0.313	0.691	0.697
GTL(3,-2)	30	0.615	0.544	0.508	0.663
	50	0.819	0.789	0.785	0.874
IG(1,1)	30	0.226	0.200	0.172	0.253
	50	0.325	0.333	0.294	0.383
Johnson SB	30	0.511	0.453	0.492	0.260
	50	0.719	0.693	0.789	0.646
Lognormal	30	0.186	0.164	0.141	0.210
	50	0.260	0.270	0.233	0.311
Pareto(5)	30	0.849	0.797	0.780	0.871
	50	0.972	0.962	0.969	0.985
Pearson V (3)	30	0.412	0.374	0.321	0.428
	50	0.598	0.603	0.549	0.641
Uniform	30	0.454	0.407	0.397	0.174
	50	0.664	0.648	0.680	0.499
Weibull(.3)	30	0.116	0.101	0.093	0.142
	50	0.142	0.149	0.134	0.193

*Based on a Monte Carlo experiment of 100,000 replications each.

4 An Alternative Approach

To remove the caprice inherent in the random pairing of observations, another approach to testing goodness-of-fit for the gamma distribution is in progress. It is based upon the following recent characterization of the gamma distribution due to Hwang and Hu [11], given in Proposition 2.2.

For developing the test we use the technique introduced by Lin and Mudholkar [15] in their test for normality. Given a random sample $x_1, x_2, ..., x_n$, we create n pairs $(\bar{x}_{-i}, c_{-i})$ by removing one observation at a time from the sample. The product moment correlation between the pairs is used to measure dependence between $\bar{x}$ and c. We therefore consider the statistic

$$r = \frac{\sum_{i=1}^{n}(\bar{x}_{-i} - \bar{x})(c_{-i} - \tilde{c})}{\sqrt{\sum_{i=1}^{n}(\bar{x}_{-i} - \bar{x})^2 \sum_{i=1}^{n}(c_{-i} - \tilde{c})^2}} \tag{12}$$

where $\tilde{c} = n^{-1}\sum_{i=1}^{n} c_{-i}$. The asymptotic null distribution of r, however, depends on the shape parameter of the gamma distribution α.

$$\sqrt{n}r \to N\left(0, 3 + \frac{10}{\alpha}\right). \tag{13}$$

One can compute the maximum likelihood estimate $\hat{\alpha}$ and substitute to obtain a measure of the standard error, but such estimation requires solving a nonlinear equation involving the di-gamma function. There is also the question of use with small samples. Investigation of these issues is in progress.

5 Conclusion

In this paper we have examined the use of Hoeffding's [9] test of bivariate independence and its asymptotic equivalent due to Blum, Kiefer and Rosenblatt [2], to obtain consistent modifications of Locke's test of the composite gamma hypothesis. It is seen that the modified tests have substantial power advantages over the original.

References

1. T.W. Anderson and D.A. Darling, *Journal of the American Statistical Association* **49**, 765 (1954).
2. J. R. Blum, J. Kiefer and M. Rosenblatt, *The Annals of Mathematical Statistics* **32**, 485 (1961).
3. P. Bougeault, *Journal of the Atmospheric Sciences* **39**, 2691 (1982).

4. R.B. D'Agostino and M.A. Stephens, *Goodness-of-fit Techniques* (Marcel Dekker, New York, 1986).
5. E. Dudewicz and E.C. van der Meulen, *Journal of the American Statistical Association* **76**, 967 (1981).
6. P. Findeisen, *Annals of Statistics* **6**, 1165 (1978).
7. D.V. Gokhale, *Computational Statistics and Data Analysis* **96**, 157 (1983).
8. L.A. Goodman, *Annals of the Institute of Statistical Mathematics* **3**, 123 (1952).
9. W. Hoeffding, *The Annals of Mathematical Statistics* **19**, 546 (1948).
10. M. Hollander and D.A. Wolfe, *Nonparametric Statistical Methods* (John Wiley, New York, 1999).
11. T.Y. Hwang and C.Y. Hu, *Annals of the Institute of Statistical Mathematics* **51**, 749 (1999).
12. N.L. Johnson, S. Kotz and N. Balakrishnan, *Continuous Univariate Distributions* (John Wiley, New York, 1994).
13. A. Kolmogorov, *Gior. Ist. Ital. Attuari* **4**, 83 (1933).
14. R.G. Laha, *Annals of Mathematical Statistics* **25**, 784 (1954).
15. C.T. Lin and G.S. Mudholkar, *Biometrika* **67**, 455 (1980).
16. R.A. Lockhart, and M.A. Stephens, *Goodness-of-fit tests for the gamma distribution.* (Technical Report, Department of Mathematics and Statistics, Simon Fraser University, 1985).
17. C. Locke, *Communications in Statistics - Theory and Methods* **4**, 357 (1975).
18. C. Locke, *Communications in Statistics - Theory and Methods* **5**, 351 (1976).
19. E. Lukacs, *The Annals of Mathematical Statistics* **26**, 319 (1955).
20. C.E. Marchetti and G.S. Mudholkar, *Goodness-of-Fit Tests and Model Validity* (Birkhäuser, Boston, 2001).
21. G. Marsaglia, *Proceedings of the Symposium on Statistics and Related Topics* (Carleton Univ., Ottawa, Ont., 1974).
22. G. Marsaglia, In *Contributions to Probability and Statistics. Essays in Honor of Ingram Olkin* (Springer-Verlag, New York, 1989).
23. J. H. Matis, W. L. Rubink and M. Makela, *Environmental Entomology* **21**, 436 (1992).
24. J.G. Mauldon, *Quarterly Journal of Mathematics* **27**, 155 (1956).
25. P. McCullagh and P. Lang, *Journal of the Royal Statistical Society B* **46**, 344 (1984).
26. J.B. McDonald and B.C. Jensen, *Journal of the American Statistical Association* **74**, 856 (1979).

27. G.S. Mudholkar and C.T. Lin, *Colloquia Mathematica Societatis*, Janos Bolyai **45**, 395 (1984).
28. G.S. Mudholkar, C.E. Marchetti and C.T. Lin, *Jounal of Statistical Planning and Inference* (To appear, 2002).
29. G.S. Mudholkar, R. Natarajan and Y. P. Chaubey, *Sankhyā Ser. B* **63**, 362 (2002).
30. G.S. Mudholkar and L. Tian, *in Jounal of Statistical Planning and Inference* (To appear, 2002).
31. G.S. Mudholkar and L. Tian, *Communications in Statistics - Theory and Methods* (To appear, 2002).
32. G.S. Mudholkar and G.E. Wilding, *On the conventional wisdom regarding two consistent tests of bivariate independence* (Technical Report, Department of Biostatistics, University of Rochester, 2001).
33. S. Nabeya, *Annals of the Institute of Statistical Mathematics* **2**, 13 (1950).
34. A.N. Pettitt and M.A.Stephens, *EDF Statistics for Testing the Gamma Distribution* (Technical Report, Department of Statistics, Stanford University, 1983).
35. B. Reiser, and D.M. Rocke, *IAPQR Transactions, Journal of the Indian Association for Productivity, Quality & Reliability* **18**, 1 (1993).
36. W. Y. Tan, *Biometrical Journal* **37**, 319 (1995).
37. O. Vasicek, *Journal of the Royal Statistical Society B* **38**, 54 (1976).
38. Y.H. Wang, *Analytical Methods in Probability Theory, Lecture Notes in Math.* **861**, 166 (Springer, New York, 1981).
39. Z. Yang, *Journal of Molecular Evolution* **39**, 105 (1994).
40. Q. Yeh, *Production and Inventory Management* **38**, 51 (1997).

ON FRAILTY MODELS AND COPULAS

DAVID OAKES
Department of Biostatistics, University of Rochester Medical Center
601 Elmwood Avenue Box 630, Rochester NY 14642, USA
E-mail: oakes@bst.rochester.edu

We review the connection between archimedean copula models and frailty models for bivariate failure-time data, emphasizing applications in biostatistics, reliability and extreme value theory. A new method for semiparametric estimation of the dependence parameter in an archimedean copula model is proposed.

1 Bivariate Frailty Models

Many models have been proposed for multivariate data failure-time data (T_1, T_2) arising in biostatistics, reliability and other applications. The classical model of Marshall and Olkin [16] has a discontinuity along the diagonal $T_1 = T_2$ which may be appropriate if the T_j represent times to failure of a machine component and failure of one component results in increased stress on the other component. However this model would not be appropriate for the data reported in Table 1 of Oakes [21] from Lawless [12] (p. 477) in which T_1 is the time to appearance of a fracture in a component and T_2 is the subsequent time to failure. A plot of the data, shown in Oakes [21] (Figure 2) suggests that T_1 and T_2 are not independent so that a model is needed which allows the bivariate survivor function $\text{pr}(T_1 > t_1, T_2 > t_2) = S(t_1, t_2)$ to be continuous over the whole positive quadrant while still allowing dependence between T_1 and T_2. Oakes [22] analyzed an example of Gumbel and Mustafi [8] from extreme value theory. Here T_1 is the height in inches of the flood of the Fox River at Berlin, WI (upstream) and T_2 the height at Wrightstown WI (downstream).

Of the many possible models we focus on one class, bivariate frailty models, which allow for very flexible modeling of the marginal distributions of T_1 and T_2 and, separately, of the dependence between them. We assume that there is an unobserved random variable W, called here a frailty, common to T_1 and T_2, and such that T_1 and T_2 are conditionally independent given W. In the reliability example above, W might represent the susceptibility or inherent weakness of the component. In studies of human populations W might represent the influence on mortality of unmeasured genetic or environmental factors common to two related individuals. Each of T_1 and T_2 follows a proportional hazards model in W, so that

$$\Pr(T_j > t_j | W = w) = B_j(t_j)^w,$$

where the $B_j(t_j)$ are baseline survivor functions, corresponding to a component with unit frailty $W = 1$.

The law of iterated expectations gives

$$\begin{aligned} S(t_1, t_2) &= E\{\text{pr}(T_1 > t_1, T_2 > t_2 | W)\} = E\{B_1(t_1)^W B_2(t_2)^W\} \\ &= p[-\log\{B_1(t_1)\} - \log\{B_2(t_2)\}]. \end{aligned}$$

where $p(\cdot) = E\{\exp(- \cdot W)\}$ is the Laplace transform of the distribution of W.

Often the $B_j(t_j)$ will not be observable. However if we write $q(\cdot)$ for the inverse function to $p(\cdot)$ we obtain the expression

$$S(t_1, t_2) = p[q\{S_1(t_1)\} + q\{S_2(t_2)\}]$$

for $S(t_1, t_2)$ in terms of its marginal survivor functions $S_1(t_1) = S(t_1, 0)$ and $S_2(t_2) = S(0, t_2)$. Genest and MacKay [5] described these as "archimedean copula models".

When W follows a gamma distribution with unit scale parameter and index κ, we have $p(s) = (1 + s)^{-\kappa}$ and

$$\begin{aligned} S(t_1, t_2) &= \left[\frac{1}{1 - \log\{B_1(t_1)\} - \log\{B_2(t_2)\}}\right]^\kappa \\ &= \left[\frac{1}{S_1(t_1)^{(-1/\kappa)} + S_2(t_2)^{(-1/\kappa)} - 1}\right]^\kappa. \end{aligned}$$

When $B_j(t_j) = \exp(-\rho_j t_j)$ we obtain the bivariate Pareto distribution $S(t_1, t_2) = (1 + \rho_1 t_1 + \rho_2 t_2)^{-\kappa}$. The form of $S(t_1, t_2)$ with exponential marginals is mentioned in Johnson and Kotz [11] (p. 288), but the general form appears to have been first proposed by Clayton [2], who also gave an alternative derivation via local odds ratios that is described below. The model was later given independently by Lindley and Singpurwalla [14] who viewed the W as the equivalent of a "random environment".

By allowing W to have the so-called "positive stable distribution" (Feller, [3] Chapter XII) with Laplace transform $p(s) = \exp(-s^\alpha)$, for some $0 < \alpha < 1$, we obtain a family considered by Gumbel, [7] with

$$S(t_1, t_2) = \exp\{-([-\log\{S_1(t_1)\}]^{(1/\alpha)} + [-\log\{S_2(t_2)\}]^{1/\alpha})^\alpha\}.$$

Hougaard [9] gave the derivation via frailties. See also Oakes and Manatunga [23].

2 Local Odds ratios

The question arises as to whether the distribution of W is determined by $S(t_1, t_2)$. Oakes [22] answered this question in the affirmative, by considering the local odds ratio

$$\theta^*(t_1, t_2) = \frac{S^{(11)}(t_1, t_2)S^{(00)}(t_1, t_2)}{S^{(01)}(t_1, t_2)S^{(10)}(t_1, t_2)}.$$

Here the superscripts denote order of differentiation in t_1 and t_2 so that the terms in the numerator are the joint density and survivor functions of (T_1, T_2) and those in the denominator are the two mixed survivor-density functions. The function $\theta^*(t_1, t_2)$ may be thought of as the odds ratio of the 2×2 table formed by dichotomizing each side of the region above and to the right of the point (t_1, t_2) at $t_1 + \Delta_1$ and $t_2 + \Delta_2$ respectively, where the Δ_j are small. It also equals the ratio of the conditional hazard functions at t_2 of T_2 given $T_1 = t_1$ and of T_2 given $T_1 \geq t_1$.

It is easily seen that

$$\theta^*(t_1, t_2) = \frac{p'(s)p(s)}{p'(s)^2},$$

where $s = q\{S_1(t_1)\} + q\{S_2(t_2)\} = q\{S(t_1, t_2)\}$, and so depends on (t_1, t_2) only through the value v say of $S(t_1, t_2)$. That is $\theta^*(t_1, t_2) = \theta\{S(t_1, t_2)\}$ for some function $\theta(v)$. Moreover, since $s = q(v)$, the inverse function theorem gives

$$\theta(v) = -\frac{vq''(v)}{q'(v)},$$

which can be integrated twice, showing that $\theta(v)$ determines $q(v)$ up to an arbitrary multiplicative factor. Hence $p(s)$ is determined up to a scale factor by $\theta(v)$.

For Clayton's model, with a gamma frailty distribution, we find that $\theta(v) \equiv 1/\kappa + 1$, i.e. free of (t_1, t_2), while for Hougaard's model,

$$\theta(v) = 1 + \frac{1 - \alpha}{-\alpha \log(v)},$$

which declines from infinity to unity as v declines from $v = 1$ to $v = 0$. Hougaard's model seems more realistic in most applications.

3 The Kendall Distribution

An alternative approach to characterizing archimedean copula models, which include frailty models, was proposed by Genest and Rivest [6]. They considered the (univariate) distribution function $K(v)$ and density $k(v)$ of the random variable $V = S(T_1, T_2)$, analogous to the usual univariate probability integral transform. Here the distribution of V is not uniform, but has

$$K(v) = v - \frac{q(v)}{q'(v)}, \quad k(v) = \frac{q''(v)q(v)}{q'(v)^2}.$$

The similarity of the functional forms of $\theta(s)$ in terms of $p(s)$ and $k(v)$ in terms of $q(v)$ is purely coincidental. It is clear however that either $K(v)$ or $k(v)$ determines $q(v)$ up to a constant factor. For Clayton's model we find that

$$K(v) = (\kappa + 1)v - \kappa v^{1+1/\kappa},$$

and for Hougaard's model

$$K(v) = v(1 - \alpha \log v).$$

For Hougaard's model it turns out that $Z = -\log(V)$ has the simple mixed gamma density $f(z) = (1 - \alpha + \alpha z)\exp(-z)$, a result originally given by Lee [13]. An immediate corollary of Lee's result is that if Z has the mixed gamma density $f(z)$ and U is uniform $(0, 1)$, with Z and U independent, then $T = U^{\alpha}Z$ has a unit exponential distribution. This amusing result can be shown directly from the facts that

$$E(T^n) = E(U^{n\alpha})E(Z^n) = \frac{1}{n\alpha + 1}\{(1 - \alpha)n! + \alpha(n + 1)!\} = n!,$$

and that the exponential distribution is determined by its moments (Feller, [3] p. 234).

To see the reason for the name Kendall distribution, note that the population analog τ of Kendall's coefficient of concordance $\hat{\tau}$ can be written $\tau = 4\text{pr}(T_1' > T_1, T_2' > T_2) - 1$ where the pairs (T_1, T_2) and (T_1', T_2') are drawn independently from $S(t_1, t_2)$. We have

$$\tau = 4E\{\text{pr}(T_1' > T_1, T_2' > T_2)|T_1, T_2\} - 1 = 4ES(T_1, T_2) - 1 = 4E(V) - 1.$$

For bivariate frailty models $\tau = 4\int sp''(s)p(s) - 1$. Clayton's model has $\tau = 1/(1 + 2\kappa)$, while Hougaard's model has $\tau = 1 - \alpha$.

Given a random sample $\{(T_1^{(i)}, T_2^{(i)}),\ i = 1, \ldots, n\}$ from the bivariate distribution, a simple analog estimator $\hat{K}(v)$ of $K(v)$ can be calculated

as the empirical distribution function $\hat{K}(v) = n^{-1}\#\{i : \hat{V}_i \leq v\}$, where $\hat{V}_i = (n+1)^{-1}\#\{j : T_1^{(j)} \geq T_1^{(i)}, T_2^{(j)} \geq T_2^{(i)}\}$. Plots of $\hat{K}(v)$ against v or possibly $\hat{K}(v)/v - 1$ against v may be useful diagnostics. Barbe *et al.* [1] prove weak convergence of a suitably normalized version of Kendall's process $\{\hat{K}(v),\ 0 \leq v \leq 1\}$ to a Gaussian process and give an expression for the covariance function. Note that the limiting covariance exhibits long-range dependence.

A smoother plot may be obtained from risk sets formed from the componentwise minima of two pairs of observations, i.e. from the $n(n-1)/2$ pairs $(T_1^{(i,j)}, T_2^{(i,j)})$ where $T_1^{(i,j)} = \min(T_1^{(i)}, T_1^{(j)})$ and $T_2^{(i,j)} = \min(T_2^{(i)}, T_2^{(j)})$. Let R_{ij} denote the size of the corresponding bivariate risk set, i.e.

$$R_{ij} = \#\{k : T_1^{(k)} \geq T_1^{(i,j)}, T_2^{(k)} \geq T_2^{(i,j)}\}.$$

and

$$\Delta_{ij} = \text{sign}\{(T_1^{(i)} - T_1^{(j)})(T_2^{(i)} - T_2^{(j)})\},$$

the indicator of concordance or discordance for the pair (i,j). It is easily seen that

$$\text{pr}(\Delta_{ij} = 1 | T_1^{(i,j)} = t_1, T_2^{(i,j)} = t_2, R_{ij} = r) = \frac{\theta^*(t_1,t_2)}{r - 1 + \theta^*(t_1,t_2)}.$$

This approach relates closely to the original proposal of Clayton (1978) for estimation of θ in his model: here the conditional probability that a pair is concordant given the size of the bivariate risk set is simply

$$\text{pr}(\Delta_{ij} = 1 | R_{ij} = r) = \frac{\theta}{r - 1 + \theta}.$$

For general archimedean copula models

$$\text{pr}(\Delta_{ij} = 1 | R_{ij} = r) \approx \frac{\theta(r/n)}{r - 1 + \theta(r/n)}.$$

Oakes [22] gave the exact probability for the positive stable frailty model.

4 Estimation of a Dependence Parameter

Suppose that a copula model is specified up to a single parameter α indexing the dependence and the marginal distributions are either parameterized fully or are kept arbitrary. Then the following strategies are possible for estimating α.

(a) (Semi)parametric maximum likelihood or an equivalent technique. Results of Murphy [17,18] and Parner [24] for special frailty models and of Murphy and van der Vaart [19] for general semiparametric profile likelihood functions suggest that this approach can be expected to be consistent, asymptotically normal, and asymptotically fully efficient. However there are some questions about its performance in small samples. The EM algorithm can sometimes be used to assist computation.

(b) Marginal profile ("pseudo-")likelihood estimation. Fit each marginal separately, either by parametric or nonparametric maximum likelihood, and then assume that each marginal is fixed at its estimated value and maximize the likelihood in α. See, for example, Shih and Louis, [25] Genest, Ghoudi and Rivest [4] and Hougaard [10]). This approach appears to work well in practice, has a relatively simple asymptotic theory, but is not fully efficient.

(c) Use an analog estimator based on Kendall's τ or another similar summary measure of correlation (Oakes, [20] Manatunga and Oakes [15] and Genest and Rivest [6]).

(d) Derive estimating equations based on bivariate risk sets, i.e. consider $\mathrm{pr}(\Delta_{ij} = 1 | R_{ij})$ - see the discussion at the end of the previous section.

(e) Fit the Kendall density $k(v)$ to the distribution of the empirical estimates $\hat{V}_i$ defined in Section 3. Asymptotic theory needs to be modified to account for the correlations among the $\hat{V}_i$. Properties of this estimator are being studied by A. Wang and D.O. It seems to work quite well. Details will be given elsewhere. We used this approach to compute estimates $\tilde{\kappa}$ and $\tilde{\alpha}$ and the corresponding values $\tilde{\tau}$ of Kendall's tau from fitting the gamma and positive stable frailty models to the data for the two examples mentioned in Section 1.

For the cable insulation data (with $\hat{\tau} = 0.46$) the estimates from the gamma model and positive stable model are respectively

Gamma Model: $\tilde{\kappa} = 0.46$ $\tilde{\tau} = 0.52$, *Stable Model*: $\tilde{\alpha} = 0.45$ $\tilde{\tau} = 0.55$.

For the Fox river data (with $\hat{\tau} = 0.52$) we follow Gumbel and Mustafi [8] and Oakes [22] in applying our models to the inverted ranks, which gives,

Gamma Model: $\tilde{\kappa} = 0.38$ $\tilde{\tau} = 0.57$, *Stable Model*: $\tilde{\alpha} = 0.43$ $\tilde{\tau} = 0.57$.

Acknowledgments

This work was supported by grant R01 52572 from the (US) National Cancer Institute.

References

1. P. Barbe, C. Genest, K. Ghoudi and B. Remillard, *J. Mult. Anal.* **58**, 197 (1996).
2. D. G. Clayton, *Biometrika* **65**, 141 (1978).
3. W. Feller, *An Introduction to Probability Theory and its Applications, Vol. II*, 2nd Ed. (John Wiley, New York, 1971).
4. C. Genest, K. Ghoudi and L.-P. Rivest, *Biometrika* **82**, 543 (1995).
5. C. Genest and R. J. MacKay, *Can. J. Statist.* **14**, 145 (1986).
6. C. Genest and L.-P. Rivest, *J. Am. Statist. Assoc.* **88**, 1034 (1993).
7. E. J. Gumbel, *J. Am. Statist. Assoc.* **55**, 698 (1960).
8. E. J. Gumbel and C.K. Mustafi, *J. Am. Statist. Assoc.* **62**, 569 (1967).
9. P. Hougaard, *Biometrika* **73**, 671 (1986).
10. P. Hougaard,*Analysis of Multivariate Survival Data* (Springer-Verlag, New York, 2000).
11. N. L. Johnson and S. Kotz, *Continuous Univariate Distributions* (John Wiley, New York, 1972).
12. J. Lawless, *Statistical Models and Methods for Lifetime Data* (John Wiley, New York, 1982).
13. L. Lee, *J. Mult. Anal.* **9**, 266 (1979).
14. D.V. Lindley and N.D. Singpurwalla, *J. Appl. Prob.* **23**, 418 (1986).
15. A.K. Manatunga and D. Oakes, *J. Mult. Anal.* **56**, 60 (1996).
16. A. W. Marshall and I. Olkin, *J. Am. Statist. Assoc.* **62**, 30 (1967).
17. S. A. Murphy, *Ann. Statist.* **22**, 712 (1994).
18. S. A. Murphy, *Ann. Statist.* **23**, 182 (1995).
19. S. A. Murphy and A. W. van der Vaart, *J. Am. Statist. Assoc.* **95**, 449 (2000).
20. D. Oakes, *J. R. Statist. Soc. B* **44**, 414 (1982).
21. D. Oakes, In *Modern Statistical Methods in Chronic Disease Epidemiology*, eds. R. L. Prentice and S.H. Moolgavkar (John Wiley, New York, 1986).
22. D. Oakes, *J. Am. Statist. Assoc.* **84**, 487 (1989).
23. D. Oakes and A. K. Manatunga, *Biometrika* **79**, 827 (1992).
24. E. Parner, E. *Ann. Statist.* **26**, 183 (1998).
25. J. Shih and T.A. Louis, *Biometrics* **51**, 1384 (1995).

SURROGATE DATA AND FRACTIONAL BROWNIAN MOTION

PETER RABINOVITCH
Alcatel Research and Innovation
600 March Rd. Ottawa ON K2K 2E6, Canada
Email:Peter.Rabinovitch@Alcatel.com

Statistical analysis of fractional Brownian motion is difficult because it exhibits long range dependence. In this paper, we describe experiments that show that the method of surrogate data can help with this problem.

1 Introduction

Fractional Brownian motion (Mandelbrot and Van Ness [7]) has become one of the standard models of telecommunications traffic due to three facts. When appropriate, it is a parsimonious model, as it is parameterized by three parameters; it has Gaussian marginal distributions; and it exhibits long range dependence, as many real traffic streams do.

The long range dependence, while necessary for accurate modeling, presents statistical difficulties. For example, the standard way of proving the Central Limit Theorem (CLT) for dependent random variables (see Resnick [11]) is to find an integer m such that random variables further apart than m are independent. Then, in essence, one applies the regular, independent CLT and estimates the error term introduced by leaving out blocks of m random variables, and shows that this error term can be made as small as desired. However, in the long range dependent setting, finding such an m may not be possible.

One could also appeal to the generalizations of the bootstrap to dependent data, such as the moving blocks bootstrap. However, it is a known result (Lahiri [6]) that the moving blocks bootstrap does not necessarily work with long range dependent data.

The solution to this problem comes from an unlikely source, nonlinear dynamics. Surrogate data (also known as "phase scrambling", "Fourier bootstrap methods", *etc.*) was developed for hypothesis testing in nonlinear dynamics in the early 1990's. At that time, many papers were published claiming that some observational set of data had a particular fractional correlation dimension, and therefore that the data was chaotic. The surrogate data method was developed (Theiler *et al.* [13]) to be able to test this claim, by generating many linear, Gaussian synthetic data sets that had the same first and second

order statistics as the original data. One could then compare a statistic, such as correlation dimension, of the original data set to a histogram of the same statistic of the synthetic data sets, and if they were significantly different, one could reject the null hypothesis that the original data could be generated by a linear, Gaussian process.

We will present evidence, both by referring to the research literature, and by numerical experiments, that the surrogate data method captures the relevant characteristics of fractional Brownian motion, including long range dependence. Pointer to S+ source code is provided so that the readers can perform their own experiments.

2 Fractional Brownian Motion

Definition 2.1. A stochastic process $\{X(t), t \geq 0\}$ is said to be a fractional Brownian motion with Hurst parameter H if

1. $X(t)$ has stationary increments for $t > 0$;
2. $X(t)$ is normally distributed with mean 0;
3. $X(0) = 0$ almost surely; and
4. The increments of $X(t), Z(j) := X(j+1) - X(j), j = 0, 1, ...$ satisfy,

$$\rho_z(k) = \frac{1}{2}\left\{|k+1|^{2H} + |k-1|^{2H} - 2k^{2h}\right\}$$

where ρ_z denotes the autocorrelation function of Z.

Remark 2.1. Let $\{X(t), t \geq 0\}$ be a fractional Brownian motion with Hurst parameter $H > 1/2$. Then the increments of $X(t)$ are long range dependent.

There are many ways to simulate fractional Brownian motion, such as random mid-point displacement; superposition of heavy tailed on/off processes, wavelet methods, and Fourier methods. In this paper we use the last method, as described in Paxson [8].

A common model for long range dependent network traffic is

$$W(t) = \lambda t + \sigma X_H(t),$$

where $W(t)$ represents the total amount of traffic that has arrived by time t, λ and σ are parameters representing the mean and standard deviation of the arrivals, and $X_H(t)$ is a fractional Brownian motion with Hurst parameter H.

3 Surrogate Data

Theiler *et al.* [13] proposed a method of generating synthetic data that preserves the mean and the autocorrelation of the original series. The mean is preserved by subtracting it from the original series, and then adding it back to the synthetic series. This is standard, as it is frequently also a step in moving blocks bootstrap, or even classical time series analysis ($ARMA$ modeling, *etc.*). For the autocorrelation, they proposed taking the Fourier transform of the series, and then scrambling the phases of the transformed data, while ensuring that the phases remain symmetric so that when the data is transformed back it is real valued. The phases can be either sampled with replacement from the original phases (*i.e.* bootstrapped), or they can be sampled uniformly on the interval. Here, we use the latter choice for its simplicity.

Now, because the autocorrelation function of a time series is determined by the power of its Fourier transform (*i.e.* the square of the coefficients), and that the power of the Fourier transform does not depend on the phase of the data, the output of this procedure has the same autocorrelation as the input does. Furthermore, this procedure requires only that the data be a realization of a linear Gaussian process, and a particular such process does not have to be specified.

Here is how we generate a single Surrogate Data Sample.

Algorithm : To Generate a Surrogate Data Sample with Gaussian Marginals

- 1. Input a time series $y[t], t = 1...N$.
- 2. Compute the discrete Fourier transform of the data: $z[t] := DFT(y[t])$. Note that $z[t]$ has both real and imaginary parts.
- 3. Randomize the phases: $z'[t] := z[t]\mathrm{e}^{i\phi[t]}$ where $\phi[t]$ where is uniformly distributed on $(0, 2\pi]$.
- 4. Symmetrize the phases to ensure that $z'[t] := z''[t]$. The mechanics of this will depend on the way in which your software stores the components of the FFT.
- 5. Invert the discrete Fourier transform: $y'[t] := DFT^{-1}z''[t]$. . Note that because of the symmetry of the phases, the resulting time series $y'[t]$ is real.
- 6. Output $y'[t]$.

Suppose we want to generate a surrogate data set for the time series $(0.707, 1.0, 0.707, 0.0, -0.707, -1.0, -0.707, 0.0, 0.707, 1.0, 0.707, 0.0, -0.707, -1.0, -0.707, 0.0)$. The DFT of this series is approximately $(0, 0, 1.414+1.414i, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1.414-1.414i, 0)$. Random, symmetric phases are $(0.0, 1.433, 4.543, 1.996, 0.878, 4.724, 2.792, 0.736, 0.0, -0.736, -2.792, -4.724, -0.878, -1.996, -4.543, -1.433)$. Our data to inverse transform is then $(0.0, 0.0, 1.155 - 1.632i, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.155 + 1.632i, 0.0)$. Finally, our inverse DFT'd data, *i.e.* our surrogate data set, is $(0.577, 0.985, 0.816, 0.169, -0.577, -0.985, -0.816, -0.169, 0.577, 0.985, 0.816, 0.169, -0.577, -0.985, -0.816, -0.169)$. In order to calculate a statistic, one would normally generate many surrogate data sets of the original, calculate the statistic of each, assemble them into a histogram, and proceed as in the bootstrap methods.

This method is frequently used in nonlinear time series analysis to perform hypothesis tests, however little theoretical work has been done to assess it's limits. The following two results are notable exceptions.

Theorem 3.1. (Braun and Kulperger [1]) Suppose the autocovariance function of the original process is absolutely summable. Let

$$f_{j,Y^*} = \sum_{t=1} nY^* e^{\frac{2\pi ijt}{n}},$$

then

$$\left| \frac{1}{2\pi n} E\left[|f_{j,Y^*}|^2 \right] - f\left(\frac{2\pi j}{n} \right) \right| \to 0$$

as $n \to \infty$, where f is the spectral density of the original process (which determines the autocorrelation of the process), and Y^* is a surrogate data version of the original process.

In other words, the surrogate data has the same Fourier spectrum as the original data, asymptotically. The next result says that (roughly) the marginal distribution of the surrogate data converges in distribution to a Gaussian distribution.

Theorem 3.2. (Braun and Kulperger [1]) If $\{Y_h, h \geq 1\}$ is a stationary ergodic sequence such that $\{Y_h^2, h \geq 1\}$ is ergodic with $E[Y_1^2] = \sigma^2$, then then

$$Y_j^* \xrightarrow{D} N(0, \sigma^2) \text{almost surely for } j = 1, 2, ...$$

Remark 3.1. The surrogate data method generates synthetic data that has the same sample autocorrelation structure as the original series, not the same autocorrelation as the population, *i.e.* we may be reproducing artifacts of the sample, and not the true distribution we are seeking.

Remark 3.2. The autocorrelation we are talking about is the circular autocorrelation, defined by which in the limit of infinite sample size approaches the population autocorrelation function, but in the case of finite sample size is only approximately equal to the population autocorrelation function, defined by

$$\rho^c(T) := \frac{1}{N}\left(\sum_{t=1}^{N-t} x_t x_{t+T} + \sum_{t=N-T+1}^{N} x_t x_{t+T-N}\right)$$

Remark 3.3. The documentation for the source code that we use for the surrogate data method (Davison and Hinkley [4]) says "The types of statistic for which this method produces reasonable results is very limited and the other methods seem to do better in most situations." Although this may be true, surrogate data seems to work well in this case.

In Percival, Sardi and Davison, [9] the following two claims about this method are made.

Remark 3.4. "Unfortunately this resampling scheme and its variants apply to a very limited range of statistics, because they mimic only second-order properties of the original data." This is true, but fractional Brownian motion is determined by second-order properties, thus in this case, it is not a limitation.

Remark 3.5. "Moreover variability is underestimated because this resampling scheme fixes the periodogram, unlike for the original series whose periodogram is random, and statistics that can be computed from the periodigram such as , display no variation across samples." This fixed periodigram problem may be able to be overcome by i) resampling the phases of the data, or ii) applying more sophisticated methods, such as those of Schreiber and Schmitz [12].

Also, it is the purpose of this note to encourage research into this method, and thus the above criticisms of the method should be seen as challenges and opportunities for further research.

4 Fractional Brownian Motion and Surrogate Data

The fact that the surrogate data method generates synthetic data sets that are linear and have Gaussian marginals is not as restrictive as it seems, as "a long-memory process can always be approximated by an $ARMA(p, q)$ process" (Brockwell and Davis, [2] p 520), *i.e.* a sufficiently complicated linear Gaussian process. Also, a linear, Gaussian time series is determined by its autocorrelation function, so we have the following chain of reasoning.

- A fractional Brownian motion (long memory process) can be approximated arbitrarily well by an $ARMA(p,q)$ process with sufficiently large p and q. It is determined by its mean and autocovariance function.
- The approximating $ARMA(p,q)$ process is a linear process with Gaussian marginals, and as such it is determined by its mean and autocovariance function.
- The surrogate data method outputs a series with the same autocovariance as its input, and approximately Gaussian marginals.
- Therefore, the output is a fractional Brownian motion with the same long range dependence properties as the input.

The only obstacle to proving this is that the proof of Theorem 1 requires that the autocovariance function of the series must be absolutely summable, which is not the case for long range dependent data. In fact it is sometimes used as the definition of long range dependence!

However, in the next section, we present experimental results that suggest that the surrogate data method does, in fact, work for fractional Brownian motion.

5 Experimental Results

For the following experiments, all trace lengths were 215, Hurst parameters were estimated with the periodigram method, and for the Moving Blocks Bootstrap, a block length of 32 was used (trace length 1/3, as suggested in Künsch [5]).

In the graph that follows for each target Hurst parameter in the set 0.5,0.525,,0.975, we generated one fractional Brownian motion. Then 100 surrogate versions of the data were generated, and a 95% confidence interval for the Hurst parameter of the surrogate data sets is presented. Similarly, 100 moving block bootstrap versions of the data were generated, and a 95% confidence interval for the Hurst parameter of these moving block bootstrap data sets is presented.

Note that the surrogate data confidence intervals are right on line with the actual Hurst parameters, those of the moving blocks bootstrap are significantly low. We see that the moving blocks bootstrap becomes more biased as the Hurst parameter increases, which is to be expected, as the moving blocks bootstrap works for independent data ($H = 0.5$) and does not for long range dependent data ($H > 0.5$). The surrogate data method has only small fluctuations in bias as the Hurst parameter increases.

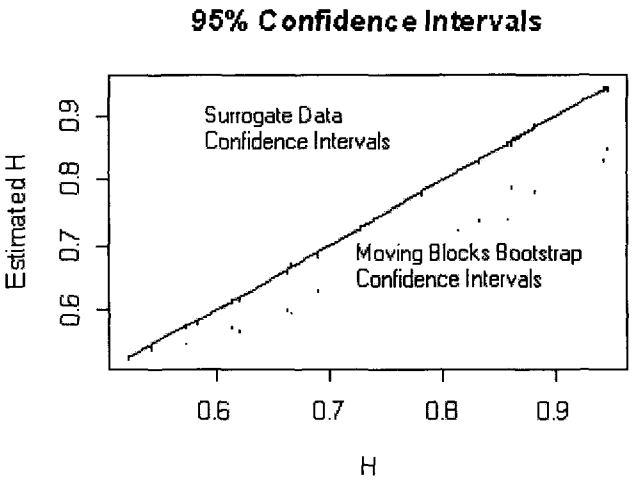


Figure 1: 95% Confidence Intervals of Hurst Parameter

6 Conclusion

It seems that the absolute summability hypothesis of Theorem 1 may be stronger than is needed to achieve the result, and that the method of surrogate data may work in the long range dependent case too. In all cases, the Surrogate Data method has less bias than the Moving Blocks Bootstrap.

In this paper we used, as our point estimator, the periodogram estimator of the Hurst parameter. However, it is expected that the surrogate data method provides a method to calculate confidence intervals for a much larger variety of statistics for long range dependent processes. An example of this is the Dembo estimator of the effective bandwidth of a fractional Brownian motion traffic stream, discussed in Rabinovitch [10]. Also, note that the same experiments were run with the Veitch-Abry wavelet estimator of the Hurst parameter, and Mandelbrot's method of generating a fractional Brownian motion with essentially the same results.

7 Source Code

Code to calculate the Hurst parameter of a data set by using the periodogram method, (and others) is available from Murad Taqqu's web site at

http://math.bu.edu/INDIVIDUAL/murad/methods/index.html. Vern Paxson's source code to generate a fractional Brownian motion was obtained from http://ita.ee.lbl.gov/html/contrib/fft-fgn.html, and is described in Paxson [8]. Code for the surrogate data method, as well as the moving blocks bootstrap was written by Angelo Canty and is included and described in Davison and Hinkley [4]. It is also available at http://lib.stat.cmu.edu/S/DH/bootlib.sh.Z. Code for Mandelbrot's method of generating a fractional Brownian motion and for the Veitch-Abry wavelet estimator of the Hurst parameter are available in Coeurjolly [3].

Acknowledgements

The author would like to thank Amit Bose as well as the anonymous reviewer for many valuable comments.

References

1. W. J. Braun and R. J. Kulperger, *Communications in Statistics - Theory and Methods* **26(6)**, 1329 (1997).
2. P.J. Brockwell and R.A. Davis, *Time Series: Theory and Methods, 2nd Edition* (Springer-Verlag, New York, 1991).
3. J.-F. Coeurjolly, *Journal of Statistical Software* **5**, (2000) (http://www.jstatsoft.org/v05/i07/).
4. A.C. Davison and D.V. Hinkley, *Bootstrap Methods and Their Application* (Cambridge University Press, UK, 1997).
5. H.R. Künsch, *The Annals of Statistics* **17**, 1217 (1989).
6. S.N. Lahiri, *Statist. Probab. Lett.* **18**, 405 (1993).
7. B.B. Mandelbrot and J.W. Van Ness, *SIAM Review* **10**, 422 (1968).
8. V. Paxson, *Computer Communications Review* **27**, 5 (1997).
9. D.B. Percival, S. Sardy and A.C. Davison, *Wavestrapping Time Series: Adaptive Wavelet-Based Bootstrapping* (Unpublished Manuscript, 2000).
10. P. Rabinovitch, *Statistical Estimation of Effective Bandwidth* (M.Sc. Thesis, Carleton University, 2000).
11. S.I. Resnick, *A Probability Path* (Birkhaüser, Boston, 1998).
12. T. Schreiber and A. Schmitz, *Surrogate Time Series* (Unpublished Manuscript, 1999, http://xxx.lanl.gov/ps/chao-dyn/9909037).
13. J. Theiler, S. Eubank, A. Longtin, B. Galdrikan and J. D. Farmer, *Physica D* **58**, 77 (1992).

ANALYSIS OF OCCULT TUMOUR TRIAL DATA WITH VARYING LETHALITY

S. N. RAI

Department of Biostatistics, St. Jude Children's Research Hospital, 332 North Lauderdale St., Memphis, TN 38105-2794
E-mail: Shesh.Rai@stjude.org

J. SUN

Department of Statistics, University of Missouri, Columbia, MO

D. HUNT

Department of Biostatistics,St. Jude Children's Research Hospital,332 N. Lauderdale St, Memphis TN 38105

The primary motivation for this work is drawn from problems arising in the analysis of incomplete data. Although data can be incomplete in many ways, we are most interested in those situations where an intermediate event, which is of prime importance, cannot be observed. For example, in the analysis of data from occult tumour trials, the time of tumour onset is not known. Due to incompleteness of the data, analyses become very complex and many assumptions are often required to develop a basis for inference concerning the tumour incidence. We briefly discuss the assumptions made that lead to many different analyses of occult tumour trial data in the past two decades. From the prospect of reducing the sample size in occult tumour trials, some models have been proposed. These semi-parametric models assume relationships in tumor-bearing and tumor-free animals and have impact on tumor onset rates and potency measures of carcinogenic substances that we have explored further.

1 Introduction

Rodent tumorigenicity experiments are commonly used to screen chemicals, drugs, and food additives for carcinogenic effects. Experiments of this type have three different, though related, purposes; these are:

(a) to estimate the rate of tumour development, which is assumed to be irreversible,

(b) to estimate the effect of tumour presence on the death rate, and

(c) to estimate carcinogenic potency, i.e., the magnitude of the dose effect of the substance of interest.

In addition, with respect to (b) one is interested in estimating how tumour presence alters the rate of death in tumour-bearing subjects; in other words,

how lethal is the tumour? Generally, the lethality of a tumour can be classified as incidental, lethal or intermediate. An incidental tumour does not alter the death rate in tumour-bearing subjects, i.e, the death rates in tumour-bearing and tumour-free animals are the same. On the other hand, if an experimental animal dies almost immediately after tumour onset, the corresponding tumour is known to be one of the lethal type. Tumours which are neither lethal nor incidental are known as tumours of intermediate lethality.

Another interesting aspect of this type of experiment is the problem of estimating the rate at which tumour develops in a specific environment on specific sets of subjects. However, the most important focus of this type of experiment is to compare a potentially carcinogenic agent to its absence in relation to the rate of tumor onset.

We now turn our attention to those trials which involve occult tumours, i.e., the presence of a tumour is determined only at the time of postmortem. A typical experiment involves about 600 experimental animals of both sexes in each of two strains randomized to a control group or one of two or three exposure groups. In most of these experiments, the animals, which are usually mice or rats, are maintained in a controlled environment and dosed with the potential carcinogen according to the experimental protocol. During the experiment, animals may be selected for interim sacrifice according to the protocol in order to determine their tumour status. At the conclusion of the study, all surviving animals are killed for humane reasons and to discover their tumour status as well; this is known as the terminal sacrifice. In experiments which involve smaller sample sizes, often only a terminal sacrifice is performed.

Occult tumour studies represent an important source of information concerning the possible carcinogenic effect of potentially hazardous substances such as chemicals, drugs and food additives. Animal survival/sacrifice experiments are commonly used in such studies, and furnish data that typically include the administered dose of the suspected carcinogen, the age of the animal at death and indicators of the presence or absence, at death, of various tumours. These data are frequently both grouped and incomplete, and are invariably difficult to analyze.

Hoel and Walburg [11] were the first investigators to consider the analysis of data from carcinogenicity studies involving occult tumours in living animals, and made a useful distinction between rapidly lethal tumours and incidental tumours. In the former, the time to death following tumour onset is short, and therefore time to death is a good proxy for the time of tumour onset. Accordingly, an analysis based on time to death with tumour is indicated. In the incidental tumour case, the tumour has no effect on the death rate and the proportion of deaths with tumour provides an estimate of the tumour preva-

lence at that time. Most tumours, however, are neither lethal nor incidental; consequently, neither of these methods is appropriate. Under the assumption that the cause of death can be identified, i.e., whether death is principally due to tumour or due to competing risks, Kodell and Nelson [14] consider a semi-Markov model with Weibull transition intensity functions. They estimate the parameters by maximizing the likelihood function for the model, and illustrate their results by analyzing a carcinogenicity trial involving the toxic substance benzidine dihydrochloride. Kalbfleisch, Krewski and Van Ryzin [12] provide a thorough review of the field, and describe the construction of the full likelihood function in detail. Other related articles include Kodell, Shaw and Johnson [15], Dinse and Lagakos [7], Turnbull and Mitchell [26] and Portier [20]; in these papers, the authors are interested in estimating the tumour onset distribution based on a multiple decrement analysis.

Another key assumption used in the development of methods for the analysis of occult tumour trials is cause of death. That is the deaths can be classified as due to tumour or due to competing causes. Although the context of observation is a fairly common assumption on which the analysis of carcinogenicity experiments is based, the cause of death is not always known, or it may be uncertain (see Finkelstein and Ryan [10], Lagakos and Ryan [16] or Kodell, Shaw and Johnson [15]). For example, in an empirical investigation of the ED_{01} data, Lagakos and Ryan [16] found the cause of death information to be inadequate for several tumour types occurring in that experiment. Consequently, it seems unwise to assume that reliable information of this type is likely to be available. Alternative to the assumption of the cause of death information, there are some procedures that are based on estimating the number of deaths due to fatal tumours and non-fatal tumours first and then estimating the tumour onset rates and testing the effects of the potential carcinogen on the tumour onset rates (Ahn *et al.* [1]).

McKnight and Crowley [18] and Dewanji and Kalbfleisch [5] both provide an extensive survey of nonparametric methods of estimation in occult tumour studies. McKnight and Crowley [18] argue that the tumour incidence rate should be the principal quantity of interest in carcinogenicity experiments, and propose a nonparametric estimator of this quantity. Dewanji and Kalbfleisch [5] derive a nonparametric estimate of this rate using the EM algorithm. In both papers, information from numerous interim sacrifices is essential in estimating this key rate. Some other results which also require numerous interim sacrifices may be found in the papers by Williams and Portier [27,28].

From the perspective of an experimentalist, interim sacrifices represent an undesirable aspect of the nonparametric approach, because they frequently inflate the size and cost of a proposed trial. One approach which reduces

the necessity for interim sacrifices involves the use of parametric models for the death rates experienced by tumour-free and tumour-bearing animals. Although many different parametric models might be possible, two particular forms were considered quite naturally in this occult tumour context.

In the first parametric form, which is referred as the constant risk ratio (CRR) model (Dinse [7,8], and Lindsey and Ryan [17]) and multiplicative failure rate (MFR) model (Rai and Matthews [23]), are based on the Cox proportional hazards model (Cox [3]). In the second, which corresponds to a cause-separable hazards model, the rates of death for tumour-bearing and tumour-free animals differ by a constant, referred as constant risk difference (CRD) model (Dinse [7,8]) and additive failure rate (AFR) model by Rai and Matthews [23]; when the death process is a discrete random variable, there is not any preferred version of the AFR model proposed in the occult tumour trials. In all these semi-parametric models, the lethality is assumed to be constant over time. When time to tumour onset is considered a continuous random variable and time to death is considered a discrete random variable, all these semi-parametric approaches provide estimates of tumour onset rates in experiments that have as few as only a terminal sacrifice. But in some experiments these rates are heavily dependent on the form of the lethality parameter (Rai *et al.* [24]), suggesting that the relationship is too restrictive and model dependent. Therefore when there is at least one more interim sacrifice, a more general model for lethality can be accommodated, that we consider in this manuscript.

In the remainder of this manuscript, we organize the remaining sections as follows. In Section 2 we define quantities of interest in occult tumour trials and construction of the likelihood. In Section 3 we define some additional notations for nonparametric estimation and also suggest some parametric relationship for describing lethality. The estimation is discussed very briefly in Section 4. In the penultimate section, we reanalyzed a subset of a data set which was previously analyzed by Dewanji and Kalbfleisch [5] and Rai and Matthews [23]. Some discussions are presented in Section 6.

2 Preliminary Considerations

Consider a carcinogenicity trial in which the presence of tumour is not clinically observable, i.e., tumour status can be determined only at necropsy. At the time of observation, an experimental unit can occupy any of the three states illustrated in Figure 1. Let the stochastic process $\{X(t)\}$ identify the state occupied by an animal at time t.

For simplicity, we suppose that n animals in state 1 at time $t = 0$ are randomly selected as experimental units and are observed for the duration

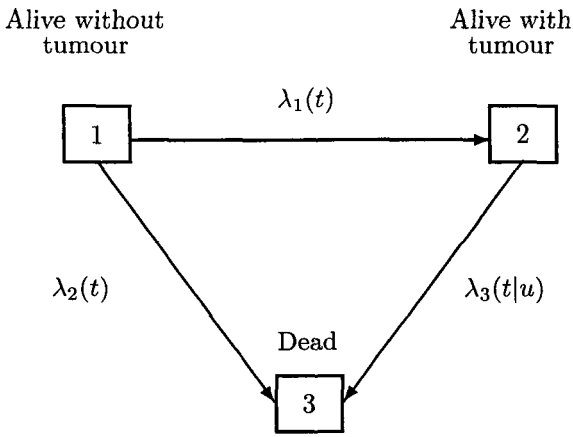


Figure 1. An illness-death model involving three states. State 1 corresponds to animals which are alive without tumour. Tumour-bearing animals which are alive are in state 2, which is frequently unobservable in carcinogenicity experiments involving occult tumours. State 3 is an absorbing state and corresponds to death.

of the trial. Let the random variable T denote the time of death and U the time of tumour onset. We also assume that the development of a tumour is an irreversible event, and therefore transitions from state 2 to state 1 do not occur, as illustrated in Figure 1.

From time to time, experimental units are sacrificed to determine their status. These animals are chosen for sacrifice independent of their health status, etc., to ensure that sacrifices can be regarded as independent of the times of the events of interest.

The intensities shown in Figure 1 are defined as the limits

$$\lambda_1(t) = \lim_{\Delta t \to 0} Pr\{X(t+\Delta t) = 2 | X(t) = 1\}/\Delta t, \tag{1}$$

$$\lambda_2(t) = \lim_{\Delta t \to 0} Pr\{T \in [t, t+\Delta t) | X(t) = 1\}/\Delta t, \tag{2}$$

and

$$\lambda_3(t|u) = \lim_{\Delta t \to 0} Pr\{T \in [t, t+\Delta t) | X(t) = 2, U = u\}/\Delta t, \tag{3}$$

for $u \leq t$; otherwise $\lambda_3(t|u) = 0$. We now define various quantities of interest, as functions of the three hazard rates $\lambda_1(t), \lambda_2(t)$ and $\lambda_3(t|u)$. The marginal distribution of T, the time to failure, can be determined from the survivor

function

$$S(t) = E\left[\exp\left\{-\int_u^t \lambda_3(v|u)dv\right\}\right]$$

where the expectation E is taken with respect to the distribution of U, the onset time. The pseudo-survival functions corresponding to the intensities $\lambda_1(.)$, $\lambda_2(.)$ and $\lambda_3(t|u)$ are

$$Q_i(t) = \exp\left\{-\int_0^t \lambda_i(v)dv\right\}$$

for $i = 1, 2$ and

$$Q_3(t|u) = \exp\left\{-\int_u^t \lambda_3(v|u)dv\right\}, \tag{4}$$

whereas

$$Q(t) = \exp\left\{-\int_0^t (\lambda_1(v) + \lambda_2(v))dv\right\} = Q_1(t)Q_2(t)$$

denotes the probability that the time to the first event — tumour onset or death without tumour — exceeds t. Note that

$$\begin{aligned} S(t) &= Pr\{X(t) = 1\} + Pr\{X(t) = 2\} \\ &= Q(t) + \int_0^t \lambda_1(u)Q(u)Q_3(t|u)du. \end{aligned}$$

Following the work of McKnight and Crowley [18], we can define an average hazard associated with the transition from state 2 to state 3 as

$$\begin{aligned} \lambda^{D|T}(t) &= \lim_{\Delta t \to 0} Pr\{T \in [t, t + \Delta t)|X(t) = 2\}/\Delta t \\ &= \lim_{\Delta t \to 0} \frac{Pr\{T \in [t, t + \Delta t), X(t) = 2\}/\Delta t}{Pr\{X(t) = 2\}} \\ &= \frac{\int_0^t \lambda_1(u)Q(u)\lambda_3(t|u)Q_3(t|u)du}{\int_0^t \lambda_1(u)Q(u)Q_3(t|u)du}. \end{aligned}$$

Similarly, the tumour prevalence function, which is the proportion of live animals with tumour in the population, is defined to be

$$\begin{aligned} \pi(t) &= Pr\{X(t) = 2|T \geq t\} \\ &= \frac{Pr\{X(t) = 2\}}{Pr\{X(t) = 1\} + Pr\{X(t) = 2\}} \\ &= \frac{\int_0^t \lambda_1(u)Q(u)Q_3(t|u)du}{S(t)}. \end{aligned}$$

Finally, we define the lethality function, $l(t)$, to be either the difference of death rates, $\lambda^{D|T}(t) - \lambda_2(t)$, or the relative death rate, $\lambda^{D|T}(t)/\lambda_2(t)$, for different forms of $\lambda_3(t|u)$. With respect to the analysis of data from a carcinogenicity trial, there are two special cases, based on lethality assumptions, which are easily handled; these correspond to tumours which are rapidly lethal and, at the other extreme, tumours which are incidental. In the case of tumours which are rapidly lethal, $\lambda_3(t|u)$ is very large, and death occurs immediately after tumour onset. In this situation, the time of death can be regarded as a surrogate for the time of tumour development. On the other hand, if the tumour is incidental, tumour development has no effect on the death rate and $\lambda_3(t|u) = \lambda_2(t)$. In this case, the analysis of data from the trial can be based on the fact that the proportion of deaths at time t which involve tumour-bearing animals provides an estimate of the tumour prevalence in the population at that time.

2.1 Constructing the Likelihood Function

Now, we outline a general framework for constructing the likelihood function. Let θ represent the full parametric vector. Thus, θ includes parameters which specify the general form of the transition intensities, c.f. Figure 1. In general, the data arising from a carcinogenicity trial will consist of the time and type of failure, i.e., natural death or sacrifice, and an indication of tumour presence or absence at autopsy. Let t_i be the realization of the random variable T for the ith experimental animal, $i = 1, 2, \ldots, n$. If C_i represents the contribution to the likelihood due to the ith animal, then the likelihood function for θ is $L(\theta) = \prod_{i=1}^{n} C_i$. Table 1 identifies the various types of observations which occur and the corresponding contribution to the likelihood.

The terms $A(t)$ and $B(t)$ which appear in Table 1 represent the integrals

$$A(t) = \int_0^t \lambda_1(u)Q(u)\lambda_3(t|u)Q_3(t|u)du$$

and

$$B(t) = \int_0^t \lambda_1(u)Q(u)Q_3(t|u)du,$$

respectively. These same terms can also be expressed in other familiar forms. Let the random variable $W = T - U$ denote the time between tumour onset and subsequent failure. For $t \geq u$, i.e., $w \geq 0$, we can write

$$Pr\{W \in [w, w + dw)|U = u\} = \lambda_3(w + u|u)\overline{G}(w|u)dw$$

Table 1. Contributions to the likelihood function, arising in the analysis of data from a carcinogenicity trial

Observation Type	Outcome	Likelihood Contribution
Death without tumour	$T = t, X(t^-) = 1$	$\lambda_2(t)Q(t)$
Sacrifice without tumour	$T > t, X(t) = 1$	$Q(t)$
Death with tumour	$T = t, X(t^-) = 2$	$A(t)$
Sacrifice with tumour	$T > t, X(t) = 2$	B(t)

where

$$\overline{G}(w|u) = Pr\{W \geq w|U = u\} = Q_3(u + w|u)$$

is the probability of survival to time t given tumour onset at u. Then,

$$A(t) = \int_0^t \lambda_3(w + u|u) f(u) \overline{G}(w|u) du$$

and

$$B(t) = \int_0^t f(u) \overline{G}(w|u) du,$$

where $f(.)$ is the (sub)density function of U, i.e., $f(u) = \lambda_1(u)Q(u)$.

Using parametric formulation for intensities or related quantities, one can analyze data (Dewanji *et al.* [6]). A nonparametric method of estimation based on the likelihood approach is considered in the next section.

3 Non-Parametric Estimation

The observed data for each animal consist of the time of death or sacrifice and an indicator of tumour presence or absence. Suppose there are M distinct death times, denoted by $t_1 < \cdots < t_M$, and let $I_j = (t_{j-1}, t_j], j = 1, 2, ..., M$, where, for completeness, we define $t_0 = 0$ and t_M is the time of terminal sacrifice. Without loss of generality, set $t_j = j$ for $j = 0, 1, ..., M$. The

argument of Kaplan and Meier [13] can be used to show that, without the imposition of distributional restrictions, the likelihood is maximized when the death rates place mass only at the observed times of death. Following Dewanji and Kalbfleisch [5], we can treat death as a discrete process and T as a discrete random variable, and we identify the range of T as $\{1, 2, \ldots, M\}$. Since there is no restriction on the choice of scale for the tumour onset process, it can be considered discrete or continuous. We suppose that the random variable U, which denotes the time of tumour onset, is continuous. Therefore, the resulting model is a mixed scale model for occult tumour trial data (Rai *et al.* [24]). When the time to death variable, T, the time to onset variable U, are both considered discrete, the resulting model is a discrete scale model (Rai and Matthews [23]). Here we are interested in a mixed scale model.

For animal i, there is an associated time of sacrifice, Y_i, which is chosen in advance of the experiment with $Pr(Y_i = j) = q_j$ ($j = 1, 2, \ldots, M$) and $\sum q_j = 1$. This random sacrifice model is analogous to a random censorship model. The event $Y_i = j$ corresponds to a planned sacrifice of animal i at time j^+, so that sacrifices at j are presumed to follow other events that may occur at j. At time M the experiment is terminated and all surviving animals are sacrificed. For this reason, such experiments are referred to as survival/sacrifice experiments.

Let u and t be realizations of U and T respectively. The discrete version of the intensities related to death rates are defined as

$$\lambda_2^*(t) = Pr\{T = t | X(t) = 0, T \geq t\} \quad \text{and}$$
$$\lambda_3^*(t|u) = Pr\{T = t | X(t) = 1, U = u, T \geq t \geq u\}$$

for $t = 1, 2, \ldots, M$ and $u \geq 0$. Furthermore, we define the tumour incidence rate

$$\begin{aligned}\lambda_1^*(j) &= Pr\{U \in I_j | T > j-1, X(j-1) = 0\} \\ &= 1 - \exp\{-\int_{j-1}^{j} \lambda_1(u)du\} \\ &\approx \int_{j-1}^{j} \lambda_1(u)du,\end{aligned}$$

for interval $j = 1, \ldots, M$. By assuming that U is a continuous random variable, we exclude the possibility that tumour onset and death with tumour occur simultaneously.

If there are many sacrifices (Dewanji and Kalbfleisch, [5], McKnight and Crowley [18], parameters of the full model can be estimated non-parametrically

without any further assumption about the form of death rates in tumour-bearing and tumour-free animals. However, most of the tumorigenicity experiments have a few interim sacrifices. Therefore some relationship between these death rates are required. The relationships in the instantaneous death rates (when T and U are assumed to be continuous random variables) in tumour-free and tumour-bearing animals that are proposed by Dinse [7] and Lindsey and Ryan [17] are

$$\lambda_3(t|u) = \lambda_2(t)e^{\gamma} \tag{5}$$

and

$$\lambda_3(t|u) = \lambda_2(t) + \gamma. \tag{6}$$

Since estimation is based on nonparametric formulation, a discrete version of the death rates in tumour-free and tumour-bearing animals using equations (5) or (6) are needed. The model described in equation (5) has some constraints that may not be very realistic (Rai [22]).

Rai and Matthews [21] have proposed the models

$$\frac{\lambda_3^*(t|u)}{1-\lambda_3^*(t|u)} = \frac{\lambda_2^*(t)}{1-\lambda_2^*(t)}e^{\gamma} \tag{7}$$

and

$$\lambda_3^*(t|u) = \lambda_2^*(t) + \gamma, \tag{8}$$

for $t = 1, 2, \ldots, M$, which are based on discrete scale for T.

The former model, which we refer to as the multiplicative failure (MFR) model, corresponds to the discrete version of the proportional hazards model (Cox [3]). The model given in equation (5), the constant risk ratio (CRR) model, is closer to the continuous proportional hazards model. The latter formulation, which we call the additive failure rate (AFR) model, represents a cause-separable hazards model for $\lambda_3^*(t|u)$. For notational convenience, we refer to both models by $\lambda_3^*(t)$. Note that these models assume that the death rate does not depend on the tumour onset time. A detailed comparison of these models is given in Rai [22].

Since both the tumour incidence rate and the death rate without tumour are completely unspecified and events may occur at any time $t = 1, \ldots, M$, we must estimate the $2M$ values of $\lambda_1^*(t)$ and $\lambda_2^*(t)$. If $\lambda_3^*(t)$ was also unspecified, corresponding to a fully nonparametric model of the experiment, we would have to estimate $3M$ parameters; however, the parametric forms for $\lambda_3^*(t)$ prescribed in equations (1) and (2) reduce that total to $2M + 1$.

These relationships were helpful to estimate the parameters even when there is no interim sacrifice. But sometimes different models lead to very

different estimates for tumour onset rates, the primary quantity of interest in tumorigenicity experiments.

In all these models the lethality parameter is considered constant over time. This assumption may not be very realistic. In some experiments there may be more information than just the indicator of tumour presence (Ryan and Orav [25], Parise *et al.* [19]). For example, we may know the grade of the tumour at the time of death and sacrifice. This grade information can be directly linked to model lethality of the tumour. We propose the following semi-parametric models for lethality:

$$\gamma = \gamma_0 + \gamma_1 t, \tag{9}$$

$$\gamma = \gamma_0 + \gamma_1 (t - u), \tag{10}$$

$$\gamma = \gamma_0 + \gamma_1 t + \gamma_2 Z, \tag{11}$$

and

$$\gamma = \gamma_0 + \gamma_1 (t - u) + \gamma_2 Z, \tag{12}$$

for $t = 1, 2, \ldots, M$.

The first relationship leads to a Markov model whereas the second relationship is a semi-Markov model. The last two models incorporate covariate information. This covariate is related to development of tumour and therefore is an internal covariate. Accommodating contributions to the likelihood function from internal covariates can sometimes be cumbersome. In this manuscript we, for simplicity, we consider the first model and other possibilities are considered somewhere else.

We conclude this section by defining some additional notation which will be used in later sections of the paper. Let

$$\begin{aligned} Q(t) &= Pr\{T > t, X(t) = 0\} \\ &= \prod_{j=1}^{t} \{1 - \lambda_1^*(j)\} \prod_{j=1}^{t} \{1 - \lambda_2^*(j)\} \end{aligned}$$

and

$$\begin{aligned} Q_3(t|u) &= Pr\{T > t, X(t) = 1 | U \in I_u\} \\ &= \prod_{j \geq u}^{t} \{1 - \lambda_3^*(j)\}, \end{aligned}$$

for $u \leq t = 1, 2, \ldots, M$. The function $Q(t)$ represents the probability that an experimental animal is alive and tumour-free at time t. This quantity is the

product of two separate terms involving $\lambda_1^*(.)$ and $\lambda_2^*(.)$. This is due to the fact that the variable U is treated as continuous and T is discrete, as described in Rai *et al.* [24]. The quantity $Q_3(t|u)$ corresponds to the conditional probability that an animal which developed a tumour in I_u is still alive at time t. The functions $Q(.)$ and $Q_3(.|.)$ will be used to calculate prevalence and survival functions.

4 Fitting the Semi-parametric Model

We use the EM1 algorithm (Rai and Matthews [21]) to estimate the parameters of the model. This method of estimation is an algorithm of the EM type which was first described by Dempster, Laird, and Rubin [4]. As with all such algorithms, maximum likelihood estimation of the parameters in the model using the observed data (the incomplete data problem) is accomplished by maximizing the conditional expectation, given the data, of the likelihood function generated by the corresponding complete data formulation. Our complete data formulation is based on the procedure given in Rai *et al.* [24]. Note that our complete data formulation provides variance estimates for tumour onset rates and death rates. Here, we briefly describe the estimation procedure.

A Complete Data Formulation

We assume that the time of tumour onset is known to belong to one of the intervals I_t, and that sacrifice is simply a right-censoring of the multistate process. In that case, the data from a sample of n experimental animals can be summarized as the observed values of the counting processes $N_1(t)$, $N_2(t)$, $N_3(t)$, $Y_0(t)$ and $Y_1(t)$ for $t = 1, \ldots, M$. The quantity $N_1(t)$ represents the number of animals with tumour onset time in I_t, whereas $N_2(t)$ identifies the number of tumour-free animals which die at time t. Likewise, $N_3(t)$ indicates the number of tumour-bearing animals which die at time t. The random variables $Y_1(t)$ and $Y_0(t)$ summarize the number of animals with and without tumour, respectively, which are sacrificed at t; thus $n = \sum_{t=1}^{M}\{N_1(t)+N_2(t)+Y_0(t)\}$.

In addition, the variable $R_1(t)$ represents the number of animals at risk of developing a tumour in I_t; likewise, $R_2(t)$ and $R_3(t)$ be the number of tumour-free and tumour-bearing animals, respectively, which are at risk of death at t. Note that due to mixed scale assumption,

$$R_2(t) = R_1(t) - N_1(t).$$

The complete data may be divided into two groups, depending on the type of information available. The first group corresponds to animals in state

1; these either remain in state 1 or move into one of the other two states, i.e., state 2 or state 3. For such animals, the contribution to the likelihood function at time t is

$$L_1(t) \propto \lambda_1^*(t)^{N_1(t)}\{1-\lambda_1^*(t)\}^{R_1(t)-N_1(t)} \times \lambda_2^*(t)^{N_2(t)}\{1-\lambda_2^*(t)\}^{R_2(t)-N_2(t)}.$$

The second group corresponds to animals in state 2; these either remain in state 2 or move into state 3. Thus, animals in state 2 at time t contribute

$$L_2(t) \propto \lambda_3^*(t)^{N_3(t)}\{1-\lambda_3^*(t)\}^{R_3(t)-N_3(t)} \tag{13}$$

to the likelihood function.

Combining these two groups, we obtain the likelihood function for the semi-parametric model based on the complete data, viz.,

$$L \propto \prod_{t=1}^{M} L_1(t)L_2(t). \tag{14}$$

An explicit version of this likelihood function is obtained by replacing $\lambda_3^*(t)$ with various model-specific forms.

The Incomplete Data Problem

Since tumour information can be obtained only at autopsy, the complete data are not available; instead, the observations consist of $N_2(t)$, $Y_0(t)$, $N_3(t)$, and $Y_1(t)$. Let $\theta^T = (\lambda_1(j), \lambda_2(j), \gamma_0, \gamma_1; t = 1, \ldots, M)$ be the vector of parameters in the semi-parametric model. A modified version of the EM algorithm (Rai and Matthews [21]) provides a simple method for estimating θ in two steps: E (expectation) and M1 (one-step maximization). Starting with an initial estimate of θ, say $\theta^{(0)}$, these steps are applied in a strictly alternating sequence until the parameter estimates converge and the log likelihood function of the observed data is maximized.

Suppose that $\theta^{(i-1)}$ represents the value of θ which was obtained at iteration $i-1$ of the algorithm. At the next E-step, we have to evaluate the conditional expectation

$$N_1^{(i)}(j) = E\{N_1(j)|N_3(t), Y_1(t), t = j, j+1, ..., M, \theta = \theta^{(i-1)}\}$$

for $j = 1, \ldots, M$. To compute this expectation, the conditional probability

$$\begin{aligned} P^{(i-1)}(j|t) &= Pr\{U \in I_j | T = t, X(t) = 1\} \\ &= \frac{\lambda_1^*(j)Q(j-1)\lambda_3^*(t|j)Q_3(t-1|j)}{\sum_{l=1}^{t} \lambda_1^*(l)Q(l-1)\lambda_3^*(t|l)Q_3(t-1|l)} \end{aligned} \tag{15}$$

and

$$P_1^{(i-1)}(j|t) = Pr\{U \in I_j | Y = t, X(t) = 1\}$$
$$= \frac{\lambda_1^*(j)Q(j-1)Q_3(t|j)}{\sum_{l=1}^{t} \lambda_1^*(l)Q(l-1)Q_3(t|l)} \tag{16}$$

for $j \leq t$, is required. Note that $Q_3(t-1|t) = 1$, i.e., the probability of surviving at least up to time $t-1$ given that the animal develops a tumour in the interval I_t is one. The dependence of the right-hand side of expression (15) and (16) on $(i-1)$ has been suppressed for notational convenience.

Consequently,

$$N_1^{(i)}(j) = \sum_{t \geq j} \{N_3(t)P^{(i-1)} + Y_1(t)\} P_1^{(i-1)}(j|t).$$

At the ith iteration, the expected values of the risk sets $R_1(.)$, $R_2(.)$ and $R_3(.)$ are evaluated by substituting the quantities $N_1^{(i)}(.)$, $N_2(.)$, $N_3(.)$, $Y_0(.)$ and $Y_1(.)$ in equations (3), (4) and (5), respectively. These expectations are required in order to proceed with the succeeding maximization step.

At the next step of the algorithm we maximize the complete data likelihood specified in equation (7) with respect to the parameter vector θ to obtain the updated value $\theta^{(i)}$. It can easily be seen that the maximum likelihood estimates for the parameters $\lambda_1^*(j)$ are

$$\hat{\lambda}_1^{*(i)}(j) = \begin{cases} N_1^{(i-1)}(j)/R_1^{(i-1)}(j), & \text{if } R_1^{(i-1)}(j) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

$(j = 1, \ldots, M)$. The other parameter estimates $\hat{\lambda}_2^{*(i)}(.)$, $\hat{\gamma}_0^{(i)}$ and $\hat{\gamma}_1^{(i)}$ are updated using an one-step maximization algorithm described in Rai and Matthews [21], as there are no closed-form expressions for these parameters.

Now we summarize our experience concerning the problem of identifiability. As noted previously, the generalized MFR and AFR models discussed in this paper each involve $2M + 2$ unknown parameters. In order to uniquely estimate θ, we therefore require at least $2M + 2$ independent observations. According to the design of the experiment, only two types of event can occur at any time: death with tumour or death without tumour. The data on death information form a multinomial table with $2M$ cell frequencies which specifies the number of deaths with and without tumour at M time points. If there are C interim sacrifices including the terminal sacrifice at M, the data on sacrifice information form C binomial tables with 2 cell frequencies which specifies the number of sacrificed animals with and without tumour. Thus, we will have $2M + C$ independent informations in the data to use in estimating

$2M + 2$ parameters in either semi-parametric model. Note that we need at least one interim sacrifice in addition to the terminal sacrifice for estimating parameters of the general model.

5 Example: Ionizing Radiation and the Occurrence of Glomerulosclerosis

A few tumorigenicity experiments supplement the terminal kill with numerous interim sacrifices; however, many occult tumour studies involve only one interim sacrifice. The requirements of fully nonparametric methods such as those proposed by Dewanji and Kalbfleisch [5] and others are rarely met in these commonly-occurring experimental designs. In this section we have chosen to analyze the data from an occult tumour study in order to illustrate the merits of the proposed semi-parametric methods. The example involves a number of interim sacrifices.

Data for this example is extracted from Table 2 in Dewanji and Kalbfleisch [5]. The data represent a summary, in intervals of 100 days, of the information gleaned from deaths and sacrifices concerning the presence or absence of the disease glomerulosclerosis. This data was also analyzed using discrete scale models in our previous work (Rai and Matthews [23]). Further information concerning this comparative assay regarding the occurrence of glomerulosclerosis following exposure to ionizing radiation may be found in the report of Berlin *et al.* [2]. The results of fitting the basic and generalized versions of MFR and AFR models to these data are summarized below.

The Kaplan-Meier estimate of the survival function for both the control and irradiated groups of mice and the corresponding estimates of the survival function which were derived from the basic and generalized versions of MFR and AFR models are almost identical, except for some time points where the generalized models produce better results than the basic models (due to lack of space these 8 graphs are not produced here). On this basis, it appears that both types of the basic semi-parametric models appear to fit the observed data well, but the generalized versions produce some improvement.

Estimates of the cumulative tumour incidence rate in each of the dose groups for various models are presented in Table 2. As a basis for comparison, the corresponding nonparametric estimates obtained by Dewanji and Kalbfleisch [5], and denoted by the symbol DK, are also included in Table 2. Note that their model is a saturated model for these data. While agreement between the estimates based on the basic MFR and AFR models and the nonparametric estimates (DK) is not so bad in the control group and the irradiated group of mice, the estimates based on generalized versions of the

Table 2. Estimated cumulative tumour incidence rates for the glomerulosclerosis data based on various mixed scale models

Age in	Control group					Irradiated group				
Days	DK[a]	M	A	GM	GA	DK	M	A	GM	GA
0-100	0.195	0.198	0.160	0.195	0.193	0.166	0.190	0.121	0.173	0.130
101-200	0.420	0.432	0.433	0.410	0.427	0.469	0.623	0.637	0.538	0.642
201-300	0.961	0.980	1.040	0.962	0.984	1.104	1.077	1.119	1.052	1.117
301-400	1.266	1.360	1.481	1.380	1.333	1.794	1.649	1.676	1.692	1.670
401-500	2.034	1.894	2.044	1.958	1.894	1.794	2.074	2.047	2.163	2.044
501-600	2.442	2.385	2.529	2.497	2.429	1.794	2.431	2.220	2.183	2.262
601-700	2.750	3.035	3.094	3.130	3.086	1.794	2.431	2.220	2.183	2.262
701-	3.537	3.813	3.672	3.747	3.789	1.794	2.431	2.220	2.183	2.262

[a] DK = Dewanji and Kalbfleisch approach, M = MFR model, A = AFR model, GM = generalized MFR model, GA = generalized AFR model

Table 3. Estimated prevalence functions for the glomerulosclerosis data based on various mixed scale models

Age in	Control group					Irradiated group				
Days	DK[a]	M	A	GM	GA	DK	M	A	GM	GA
0-100	0.194	0.198	0.157	0.194	0.193	0.164	0.194	0.116	0.171	0.127
101-200	0.369	0.387	0.381	0.360	0.375	0.368	0.553	0.561	0.438	0.562
201-300	0.698	0.715	0.752	0.701	0.713	0.758	0.728	0.763	0.692	0.759
301-400	0.775	0.814	0.858	0.817	0.802	0.939	0.868	0.889	0.880	0.885
401-500	0.949	0.904	0.937	0.917	0.905	0.945	0.910	0.926	0.938	0.922
501-600	0.970	0.940	0.966	0.958	0.949	0.937	0.928	0.935	0.949	0.933
601-700	0.970	0.965	0.985	0.983	0.978	0.941	0.895	0.928	0.969	0.921
701-	1.000	0.963	0.992	0.994	0.984	1.000	0.775	0.911	0.988	0.886

[a] as in Table 2

Table 4. Values of the maximized log likelihood and the corresponding number of parameters estimated when the glomerulosclerosis data were fitted with the mixed scale MFR and AFR models involving $\gamma = \gamma_0 + \gamma_1 t$

	$\gamma_1 = 0$			$\gamma_1 \neq 0$		
		Maximized log likelihood			Maximized log likelihood	
Dose group	Number of parameters	MFR	AFR	Number of parameters	MFR	AFR
Control	17	-1778.50	-1782.06	18	-1777.78	-1778.41
Irradiated	17	-2886.39	-2886.79	18	-2880.87	-2886.72

MFR and AFR models are a bit closer to the estimates based on the saturated model. Estimates of the tumour prevalence function, $\pi(t)$, based on various semi-parametric models are summarized in Table 3. The corresponding values derived by Dewanji and Kalbfleisch [5] are once again included for com-

parison purposes. Relative to the nonparametric estimates of $\pi(t)$, the basic MFR model appears to estimate a higher prevalence during the initial periods and lower towards the end. On the contrary, the revers pattern is observed when comparing the estimates obtained using the basic AFR model with the saturated model. The estimates obtained using the generalized versions of these models bridge the gap with those obtained using the saturated model, expect for the first and last time intervals for the irradiated group when using the AFR model.

Table 4 summarizes the values of the maximized log likelihoods and the numbers of parameters estimated for each semi-parametric model that was fitted to the data. The values in the table provide a basis for examining the adequacy of the simpler model for γ with respect to the glomerulosclerosis data. In the case of the MFR model, it appears that the hypothesis $\gamma_1 = 0$ is not contradicted by the data for the control group of mice (p=0.230); however, the same test yields a significance level of 0.001 when applied to the results of fitting the MFR model to the irradiated mice. Somewhat curiously, the reverse of the above remarks summarize the conclusions based on the likelihood ratio tests when the AFR model is fitted to the same data set.

Finally, we also found that incidence rates are different in both the groups for basic and generalized versions of the MFR and AFR models. The results are very similar to those obtained using discrete scale models for tumour onset and death rates (Rai and Matthews [23]).

6 Discussion

As we have previously indicated, in the absence of supplementary information such as that provided by interim sacrifices or the cause of death, realistic approaches to the estimation of tumour incidence rates and carcinogenic potency will necessarily involve modeling assumptions. For the most part, the assumptions on which the discrete scale models of Rai and Matthews [23] were based continue to apply with respect to the mixed scale models which we have just described in preceding sections of this paper. The principle difference has been the choice of scale on which tumour onset is deemed to have occurred.

A benefit which is realized through selection of the mixed scale model concerns the form of the likelihood function. In this case, the complete data formulation can be factored into two components; one of these involves parameters related to tumour onset, while the remaining factor concerns the death rates with and without tumour. Such a separation is not achieved in the corresponding complete data likelihood generated by the discrete scale model.

Note that the choice of a mixed scale model induces a natural preference or ordering of the two events, tumour onset and death, which is not present in corresponding discrete scale models. In effect, because tumour onset occurs on a continuous scale whereas the time to death is a discrete random variable, tumour-free animals are at risk of developing tumour at any time, whereas the same animals are only at risk of death immediately prior to the next point on the discrete scale of the random variable T after remaining tumour-free for the intervening interval. This natural ordering has implications for the resulting estimates of the tumour incidence rate, and the death rates with and without tumour, which we believe are deserving of further study. For example, in the mixed scale model it is possible for animals to die with tumour at the first value in the range of T, i.e., the time of the first observed death or sacrifice, whereas the same event is impossible in a discrete scale model. Consequently, at least one interim sacrifice is required in order to obtain unique parameter estimates in the discrete scale models, whereas no interim sacrifice is required in the case of the mixed scale models.

When there are interim sacrifices in additional to the terminal sacrifice, our generalization of the lethality function using one additional parameter provides adequate results - good fit to the data and comparable estimates of the tumour onset and prevalence rates.

Acknowledgments

The authors wish to thank two referees and the Editor for their comments and suggestions. The research of Dr. Rai and Dr. Hunt was supported in part by a Cancer Center Support grant (CA21765) and the American Lebanese Syrian Associated Charities (ALSAC) and that of Dr. Sun was supported by a grant from the National Institutes of Health.

References

1. H. Ahn, H. Moon and R.L. Kodell, *Applied Statistics* **49**, 157 (2000).
2. B. Berlin, J. Brodsky and P. Clifford, *Journal of the American Statistical Association* **74**, 5 (1979).
3. D.R. Cox, *Journal of the Royal Statistical Society, Ser. B* **34**, 187 (1972).
4. A.P. Dempster, N.M. Laird and D.B. Rubin, *Journal of the Royal Statistical Society, Ser. B* **39**, 1 (1977).
5. A. Dewanji and J.D. Kalbfleisch, *Biometrics* **42**, 325 (1986).
6. A. Dewanji, D. Krewski and M. J. Goddard, *Biometrics* **49**, 367 (1993).
7. G.E. Dinse *Biometrics* **47**, 685 (1991).

8. G.E. Dinse, Biometrics **49**, 399 (1993).
9. G.E. Dinse and S.W. Lagakos, *Biometrics* **38**, 921 (1982).
10. D.M. Finkelstein and Ryan, L.M. *Applied Statistics* **36**, 121 (1987).
11. D.G. Hoel and H.E. Walburg, *Journal of the National Cancer Institute* **49**, 361 (1972).
12. J.D. Kalbfleisch, D.R. Krewski and J. Van Ryzin, *The Canadian Journal of Statistics* **11**, 25 (1983).
13. E.L. Kaplan and P. Meier, *Journal of the American Statistical Association* **53**, 457 (1958).
14. R.L. Kodell and C.J. Nelson, *Biometrics* **36**, 267 (1980).
15. R.L. Kodell, G.W. Shaw and A.M. Johnson, *Biometrics* **38**, 43 (1982).
16. S.W. Lagakos and L.M. Ryan, *Environmental Health Perspectives* **63**, 211 (1985).
17. J. Lindsey and L.M. Ryan, *Applied Statistics* **42**, 283 (1993).
18. B. McKnight and J. Crowley, *Journal of the American Statistical Association* **79**, 639 (1984).
19. H. Parise, G. Dinse and L. Ryan, *Applied Statistics* **50**, 171 (2001).
20. J.C. Portier, *Biometrika* **73**, 371 (1986).
21. S.N. Rai and D.E. Matthews, *Biometrics* **49**, 587 (1993).
22. S.N. Rai, *Biometrical Journal* **39**, 909 (1997).
23. S.N. Rai and D.E. Matthews, *Applied Statistics* **46**, 93 (1997).
24. S.N. Rai, D.E. Matthews and D.R. Krewski, *The Canadian Journal of Statistics* **28**, 65 (2000).
25. L.M. Ryan and E.J. Orav, *Biometrika* **75**, 631 (1988).
26. B.W. Turnbull and T.J. Mitchell, *Biometrics* **40**, 41 (1984).
27. P.L. Williams and C.J. Portier, *Communications in Statistics* **21**, 711 (1992).
28. P.L. Williams and C.J. Portier, *Biometrika* **79**, 711 (1993).

NONLINEAR MIXED EFFECTS MODELS: RECENT DEVELOPMENTS

PODURI S.R.S. RAO AND NICHOLAS ZAINO
Statistics Program, Hylan 915, University of Rochester
Rochester, NY 14627
E-mail: raos@troi.cc.rochester.edu

Illustrations of nonlinear models with fixed and random effects along with their applications for different practical situations are presented. Least squares, maximum likelihood and other procedures suggested in the literature for estimating the fixed effects and variance components are described. Approximations suggested for the likelihood procedure are summarized. Suitable transformations required for these procedures are also presented.

1 Introduction

Nonlinear models are extensively employed, for instance, in bioassay, in radioimmunoassay, for the analysis of dose-response measurements related to medical treatments, for the examination of concentrations and absorptions of chemical compounds in Pharmacokinetics, and to evaluate the effects of herbicides. They are also widely used to study the degradation and reliability of electronic components and integrated systems, to describe time-to-failure distributions and to predict the plateaus of deterioration of the above types of components and systems. Nonlinear models are also suggested to examine the departures from the assumptions such as normality of the linear models.

As in the linear case, some of the coefficients in a nonlinear model can be fixed and the others random. The Analysis of Variance (ANOVA), Minimum Norm and Minimum Variance Quadratic Unbiased Estimation, (MINQUE and MIVQUE), Maximum Likelihood and Restricted Maximum Likelihood (ML and REML) procedures and their modifications are widely used to estimate the variance components and fixed effects of a linear mixed effects model. However, these approaches have to be modified and adapted in the case of a nonlinear mixed effects model. Suitable transformation followed by modifications of the above procedures may provide acceptable estimates for some types of nonlinear models.

The major motivation for the present article comes from Solomon and Cox [20], who have presented examples of nonlinear mixed effects models, examined some transformations to estimate the variance components and fixed effects, suggested an approximation to the likelihood function, and evaluated its appropriateness for estimating the parameters of an exponential regression

model and also for testing a hypothesis related to the 2×2 contingency tables through the logistic regression model. An overview of these and other illustrations and estimation procedures suggested in the literature are presented in the following sections.

Illustrations of nonlinear models along with some of the suggested estimation procedures are presented in Sections 2 and 3. Linear mixed effects models, the corresponding marginal distributions along with the MIVQUE, ML and REML procedures for the fixed effects and variance components are briefly summarized in Section 4. The description in this section leads into Section 5, where a general description of the nonlinear mixed effects model, an approximation to the marginal distribution and the likelihood function for the nonlinear model along with the suggested estimation procedure are presented. Section 6 presents illustrations of the degradation models employed to improve the reliability of electronic and other types of components and systems, and a two-stage procedure of estimation. A summary with discussion appears in Section 7.

2 Four Illustrations of Nonlinear Models

Two commonly used models in a number of applications are the exponential regression model

$$y_{ij} = \exp(\alpha_i + \beta_i x_i) + \epsilon_{ij}, \tag{2.1}$$

and the logistic regression model

$$y_{ij} = \exp(\alpha_i + \beta_i x_i) + \epsilon_{ij}, \tag{2.2}$$

For both the models, $i = 1, 2, ..., k$ denotes the experimental units and $j = 1, 2, ..., n_i$ denotes the number of observations on the i^{th} unit.

In these models, the response y_{ij} is related to the *fixed* variable x_i, which can be the covariate or the concomitant variable. The *error* ϵ_{ij} is usually assumed to follow the normal distribution with zero mean and unit standard deviation. When the coefficients α_i and β_i are fixed parameters, they can be estimated through the nonlinear estimation and similar procedures. However, in some applications, one or both the coefficients are assumed to be fixed or random.

A model used in economic analysis for relating production (P) to labor (L) and capital (C) is the Cobb-Douglas *production function*

$$P_{ij} = \alpha_i (L_{ij})^{\beta_i} (C_{ij})^{\gamma_i} + \epsilon_{ij}, \tag{2.3}$$

$i = 1, 2, ..., m$ industries, and $j = 1, 2, ..., t$ time periods, for instance. One or more of the coefficients α_i, β_i and γ_i may be assumed to be random to perform meta-analysis, for instance.

To analyze the effects of herbicides, the dose-response model considered by Rudemo *et al.* [16] takes the form of

$$y_{ij} = C + (D - C)/[1 + (x_{ij}/x_{0i})^{\alpha_i}] + \sigma\epsilon_{ij} \ , \tag{2.4}$$

$i = 1, 2, ..., 8$ herbicides, and $j = 1, 2, ...6$ doses. In this model, x_{ij} is the concentration of herbicide, x_{0i} is the half-effect of the herbicide, and y_{ij} is the weight of plants. The coefficients C and D are the same for all the herbicides, and α_i represents the fixed effect. The random error ϵ_{ij} is assumed to have mean zero and unit variance.

When the half-effect varies with the herbicide absorption by the plants, the random-coefficient model considered by the above authors takes the form of

$$y_{ij} = C + (D - C)/[1 + (\beta_i x_{ij})^{\alpha_i}] + \sigma\epsilon_{ij}, \tag{2.5}$$

where α_i is fixed and β_i random. For the case of replications, with the additional subscript k, the random effect r_k was added to the model.

Box-Cox type transformation, Taylor's approximation to the right side of the model and suitably weighting the error term followed by the likelihood procedure were examined to estimate the parameter of the transformation. Data on four pairs of herbicides and the dry weights (y) of 150 plants were used for this purpose.

The above authors have also mentioned that when there are errors of measurement in the controlled variable x_{ij}, it can be replaced in the above models by $x_{ij}(obs) = x_{ij}(true) + \sigma_x\delta$. Now the model would contain an additional random effect.

3 Nonlinear Models to test normality or to validate transformations

In the one-way mixed effects model, $y_{is} = \mu + A_i + B_{is}$, the random effect A_i and the error term B_{is} are usually assumed to be independently normally distributed with zero means and variances σ_A^2 and σ_B^2 respectively. To represent departures from normality, Solomon and Cox [20] considered

$$y_{is} = \mu + A_i + B_{is} + \alpha_{20}A_{is}^2 + \alpha_{11}A_iB_{is} + \alpha_{02}B_{is}^2 \ , \tag{3.1}$$

where $i = 1, 2, ...m$ represents the units and $s = 1, 2, ..., r$ represents measurements on the ith unit. The above authors have demonstrated that "skewness

of the random effects and heterogeneity of the within-group variation" are related to the fixed parameters α_{20}, α_{11} and α_{02}. To test the hypothesis that these coefficients are zero, it was suggested that a suitable permutation test or the likelihood ratio procedure can be employed. Estimation of these parameters and the variance components σ_A^2 and σ_B^2 through the third order moments obtained from the observations y_{is} was described.

To discuss transformations, Solomon [19] considered the mixed-effects model $y_{is}^{1/k} = \mu + A_i^* + B_{is}^*$, where A_i^* is the random effect and B_{is}^* is the error term. Solomon and Cox [20] noted that retaining the quadratic terms, this model can be expressed approximately as the nonlinear model in (3.1) with $\alpha_{20} = \alpha_{02} = \alpha_{11}/2$, and examined the approximation with data from 16 systolic and diastolic blood pressure measurements on each of 25 patients. This relationship for the three fixed parameters was found to be satisfactory for their estimates obtained as described above with the actual measurements as well as their square root and log transformations.

4 Linear Mixed Effects Model, Marginal Distribution and the Likelihood Estimation

This section briefly summarizes the estimation procedures for the linear mixed effects model, and it is intended as an introduction to the estimation procedures described in the following two sections.

Consider the case in which $f(y_{ij}|\alpha_i)$ is normal with mean $\mu_i = \mu + \alpha_i$ and variance σ^2, and $f(\alpha_i)$ is normal with mean zero and variance σ_α^2. This situation can be expressed in terms of the linear model

$$y_{ij} = \mu_i + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij} \,, \tag{4.1}$$

$i = 1, 2, ..., k$ and $j = 1, 2, ..., m$, where ϵ_{ij} and α_i follow independent normal distributions with zero means and variances σ^2 and σ_α^2 respectively.

The joint distribution of the $n = km$ observations is

$$f(y_{11}, ..., y_{km} \mid \alpha_1, \alpha_2, ..., \alpha_k) f(\alpha_1, \alpha_2, ..., \alpha_k) \,, \tag{4.2}$$

which can be expressed as

$$g_1(y_{11}, ..., y_{km}; \mu, \sigma^2) g_2(\alpha_i, \bar{y}_i, \mu, \sigma^2, \sigma_\alpha^2), \tag{4.3}$$

where $\bar{y}_i$ is the mean of the m observations of the i^{th} group.

The marginal distribution of $(y_{11}, ..., y_{km})$, which provides the likelihood function $L(\mu, \sigma^2, \sigma_\alpha^2)$, is obtained by integrating the joint distribution in (4.2) with respect to α_i, that is, by integrating out g_2 in (4.3). As pointed by Solomon and Cox [20], this integral is the same as $E[L(\mu, \sigma^2, \sigma_\alpha^2, \alpha_i)]$ where the

expectation is taken with respect to α_i. ¿From this integration, the marginal distribution of $(y_{11}, ..., y_{km})$ becomes the product of k multinormal distributions each with mean μ, variance $\sigma_\alpha^2 + \sigma^2$ and covariance σ_α^2, and it provides the above likelihood function. Note that this likelihood corresponds to the linear mixed effects model $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ in (4.1).

In general, the Linear Mixed Effects Model can be expressed as

$$Y = X\beta + \epsilon = X\beta + U\xi + U_0\xi_0 \,. \tag{4.4}$$

In this model, Y is an n-vector of observations, ϵ is an n-vector which includes the random effects and the residuals, X is an $(n \times p)$ known matrix, and β is the $(p \times 1)$ vector of the fixed parameters. In the right hand side of (4.4), suggested by C.R. Rao [13], ξ represents the random effects and ξ_0 the residuals.

As suggested by Harville [5], the Linear Mixed Effects Model can also be expressed as

$$Y = X\beta + Zb + \epsilon \,. \tag{4.5}$$

In this representation, the vector of the random effects b is assumed to have mean zero and dispersion $D(b) = D$, and the vector of residuals ϵ is assumed to have mean zero and dispersion $D(\epsilon) = R$. As a result, the dispersion of Y becomes $V = ZDZ' + R$.

The marginal distribution of Y is obtained by integrating the joint distribution $f(Y|b)f(b)$ with respect to b. When $f(Y|b)$ and $f(b)$ follow the multivariate normal distributions, this marginal distribution becomes the multivariate normal distribution with mean $X\beta$ and covariance V given by

$$f(Y; \beta, \sigma) = (2\Pi)^{-n/2}|V|^{-(1/2)} \exp[-(1/2)(Y - X\beta)'W(Y - X\beta)] \,. \tag{4.6}$$

In this expression, $\sigma = (\sigma_{ij})$ denotes the elements of D and R, and $W = V^-$. This distribution provides the likelihood function for β and σ, $L = L(\beta, \sigma)$.

¿From (4.6), the likelihood estimates of β and σ are obtained from

$$\beta = (X'WX)^- X'WY \tag{4.7}$$

and

$$tr[V(dV/d\sigma_{ij})] = (Y - X\beta)'W(dV/d\sigma_{ij})W(Y - X\beta) \,. \tag{4.8}$$

Note that if $V = \Sigma\sigma_i{}^2 V_i$, $(dV/d\sigma_i^2) = V_i$.

The joint density $f(Y|b)f(b)$ was considered by Hartley and J.N.K. Rao [4] for the estimation of the fixed effects and the variance components and by Henderson [7] for the estimation of β and predicting the random effect b.

The MINQUE and MIVQUE procedures as well as the REML method for estimating the variance components are presented in C.R. Rao and Kleffé

[14] and P.S.R.S. Rao [15], and the REML method by Harville [6]. As shown by all these authors, the Best Linear Unbiased Predictor (BLUP) of the random effect vector b is obtained from $DZ'W(Y - X\beta)$. With the normality assumption for b and ϵ, the BLUP is the same as $E(b|Y)$. For the variance components of the Generalized Linear Model (GLM) considered by Schall [17], Lee and Chaubey [8] examined the MINQUE type of estimation.

5 Nonlinear Mixed Effects Models and an Approximate Likelihood Estimation

Following (4.5), for the nonlinear case, Vonesh and Carter [23] and Gumpertz and Pantula [3], for instance, consider mixed effects models of the type

$$Y = F(X\beta) + Zb + \epsilon\,, \tag{5.1}$$

where $F(X\beta)$ is a nonlinear function of the fixed effects β. Note that Zb with the random effect b and the error ϵ are additive.

Sheiner and Beal [18], Lindstrom and Bates [9], Solomon and Cox [20], Wolfinger [24], and others consider models of the form

$$Y = F(X\beta + b) + \epsilon \tag{5.2}$$

and

$$Y = F(X, \beta, Z, b) + \epsilon = F + \epsilon. \tag{5.3}$$

The nonlinear models of the type presented in Sections 2 and 3 can be represented through these general models or their modifications. As in Section 4, the dispersions of b and ϵ can be represented by D and R respectively.

5.1 Estimation through the Approximate Likelihood

For the estimation of the fixed effects and variance components of the nonlinear models in Section 2, 3 and 5, least squares, Analysis of Variance, Maximum Likelihood and similar methods may be adapted. For instance, for the estimation of the fixed effects of the model in (3.1), Solomon and Cox [20] considered the method of moments approach. Transformations of one or both sides of the models before applying these types of approaches may provide acceptable estimates in some situations.

However, unlike in the linear case described in Section 4, for the nonlinear mixed-effects models of the types in Sections 2, 3 and 5, the integral of g_2 can not be expressed in an explicit form. As a consequence, the marginal distribution of the observations and the corresponding likelihood function will not be available.

Solomon and Cox [20] suggest expressing g_2 in a Taylor's series and retaining the leading terms, before integrating out the random effects. If g_2 is of the exponential type, these terms correspond to the *Laplace expansion.* Now, the fixed effects and variance components can be estimated from the resulting marginal distribution obtained as described in Section 4. The validity of this approximate procedure was examined by the above authors for the likelihood corresponding to the exponential model

$$Y_{js} = \exp(\theta + A_j)x_s + B_{js}, \tag{5.4}$$

$j = 1, 2, ..., m$ and $s = 1, 2, ..., r$. The random effect A_j and the error term B_{js} were assumed to be independently normally distributed with zero means and variances σ_A^2 and σ_B^2 respectively. The approximate likelihood was found to be close to the exact likelihood, unless σ_A^2 is very large.

The above authors have also investigated the approximation to the likelihood for testing the difference between the effects of two treatments with the responses arranged in several 2×2 tables. For the hypothesis of equality of the means of the effects of the treatments, the above type of approximation for the likelihood corresponding to the logistic model was considered. The power of the *score test* with the approximation was compared with the Mantel-Haenszel test.

To adopt the approach for the approximate likelihood $L(\beta, \sigma)$ of (5.3), Wolfinger and Lin [25] consider the joint distribution

$$g = C|R|^{-1/2} \exp\{-(Y - F)'R^{-1}(Y - F)/2\}$$
$$|D|^{-1/2} \exp\{-b'D^{-1}b/2\}\,. \tag{5.5}$$

Let, $F_0 = F_0(X, \beta) = F[(X, \beta, Z, b)|b = 0]$, $Z_0 = dF/db'|b = 0$ and $Vo = Z_0 D Z_0' + R$. The approximation results in the maximization of the likelihood corresponding to the mixed effects model

$$Y = F_0(X, \beta) + Z_0 b + \epsilon\,. \tag{5.6}$$

With the "pseudo observations", $Y_0 = Y - F_0(X, \beta) + X_0\beta$, where $X_0 = dF_0/d\beta$, the estimate of β is now obtained from

$$\beta_0 = (X_0' W_0 X_0)^{-}(X_0' W_0 Y_0), \tag{5.7}$$

where $W_0 = (V_0)^{-1}$. This expression resembles (4.7) for the linear mixed effects models. Estimates of the variance components σ for (5.1), (5.2) and (5.3) are obtained from (4.8) with X_0, Y_0, V_0 and β_0. The same approximation can be considered for the REML estimates of the variance components. The BLUP for b is obtained from $b_0 = DZ_0' W_0(Y_0 - X_0\beta)$. For the empirical BLUP, β in this expression is replaced by the estimate in (5.7).

The above authors examined the approximation in (5.6) for two models. The first model considered was the logistic model

$$y_{ij} = (\beta_1 + b_{i1})/\{1 + \exp[-(t_{ij} - \beta_2 - b_{i2})/\beta_3\} + \epsilon_{ij}, \tag{5.8}$$

for $i = 1, 2, ..., 15$ subjects, with $j = 1, 2, ...10$ observations on the i^{th} subject. The coefficients β_1 and β_2 are the fixed effects, (b_{i1}, b_{i2}) are the random effects and ϵ_{ij} is the error. Ten different values were considered for t_{ij}; same for each subject. This model was examined earlier by Pinheiro and Bates [11]. The second model examined was the pharmacokinetic model for the concentration of a drug given by

$$\log y_{ij} = \log F + \epsilon_{ij}, \tag{5.9}$$

where $F = \{10k_a[\exp(-k_i t_{ij}) - \exp(-k_a t_{ij})]/v_i(k_a - k_i)\}\exp(\epsilon_{ij})$. In this model k_a is the fixed effect and k_i is the random effect. The investigation showed that for the approximate likelihood method, estimators of the fixed effects are almost unbiased for the first model, but biased for the second.

Shun and McCullagh [21] discuss the conditions suitable for the Laplace approximation. Vonesh [22] applied the Laplace approximation only to the random effects, and found that the resulting estimates are consistent; the rate of convergence was found to depend on both the number of subjects (i) and the number of observations (j).

For the repeated measures data, Lindstrom and Bates [9] considered the model

$$y_{ij} = F(X_i\beta + Z_i b_i) + \epsilon_{ij}\ . \tag{5.10}$$

For the Taylor's approximation, the derivative of F at the current estimate of $b' = (b_1, b_2, ...)$ was considered; Sheiner and Beal [18] considered the derivative at $b' = 0$. Ramos and Pantula [12] also discuss the estimation procedures for the nonlinear random coefficient models.

6 Nonlinear Models for Reliability and Degradation

Degradation, which is the cumulative damage or deterioration, provides valuable information on the reliability, for instance, of an electronic component. An important aspect of degradation is the estimation of the time-to-failure distribution for a specified level of deterioration. Degradation over time can be linear or nonlinear. We describe below one illustration for the linear case and two illustrations for the nonlinear case.

Linear degradation can be described by the model

$$y_{ij} = \alpha_i + \beta_i t_j + \epsilon_{ij}, \tag{6.1}$$

for $i = 1, 2, ..., n$ components (sample paths), and $j = 1, 2, ..., m_i$ measurements on the i^{th} component, where t_j represents the time of the j^{th} measurement. When β_i in (6.1) follows the Weibull, normal or lognormal distribution or when (α_i, β_i) follow the bivariate normal distribution, Lu and Meeker [10] present explicit expressions for the time-to-failure distributions.

A simple description of nonlinear degradation can be obtained from the Paris Law, $g = cs^m$, where g is the growth of fatigue cracks, s is the stress intensity factor, and (c, m) are constants. Starting with this equation, the above authors considered the model

$$y_{ij} = -(1/\beta_{2i}) \log[1 - (.90)^{\beta_{2i}} \beta_{1i} \beta_{2i} t_j] + \epsilon_{ij}. \tag{6.2}$$

In this model, y_{ij} is log (crack length at time t_j), and (β_{1i}, β_{2i}) are random. The time-to-failure distribution was obtained through a two-stage procedure. At the first stage, estimates of (β_{1i}, β_{2i}) were obtained. For the second stage, the sampling variances and covariances of these estimates along with the variances and covariances of (β_{1i}, β_{2i}) were considered. Estimates of the expected values of (β_{1i}, β_{2i}) and the time-to-failure distribution were obtained with the weights obtained from the between and within variances and covariances. We note that for the linear mixed effects model, the estimator obtained from (4.7) for the fixed parameter is a weighted average of the least squares estimators, where the weights are inversely proportional to the unconditional variances of the least squares estimators. A similar procedure was followed for the above two-stage estimation.

The second illustration is related to the *Plateau*, maximum degradation over time. Highly reliable systems such as lasers, integrated circuits and optical communication devices do not fail over time, but their performance can deteriorate and reach a plateau. Boulanger and Escobar [1] describe the experiment of Gumpertz and Pantula [3] and consider the Weibull sigmoidal random coefficient model for degradation,

$$y_{ij}(t) = \alpha_{ij}[1 - \exp(-(\beta_{ij} t)^{\gamma}] + \epsilon_{ij}(t). \tag{6.3}$$

In this model, $y_{ij}(t)$ is the change in propagation delay up to time t of the j^{th} device at the i^{th} stress level x_i, α_{ij} is the plateau for the j^{th} device which is random over the devices, β_{ij} is related to α_{ij}, and γ is a fixed constant.

The problems of importance considered by the above authors were to (a) estimate α_{ij}, (b) optimize the stress levels and the proportion of devices to allocate for each stress level, and (c) optimize the times for measuring the degradation at each stress level. At the first stage, Maximum Likelihood Estimates of $(\alpha_{ij}, \beta_{ij})$ for each device at the i^{th} stress level x_i were obtained.

At the second stage, the model

$$\log(\alpha_{ij}) = A + Bx_i + \delta_{ij} \tag{6.4}$$

was considered. In this model, $V(\delta_{ij})$ includes the sampling variance from the first stage and the variation across the devices. Estimates of A and B were obtained from this model.

7 Summary and Discussion

Nonlinear mixed effects models are applicable in several practical situations, and some of the important illustrations are presented in this article. To estimate the variance and covariance components of the linear mixed effects models, the ANOVA, MIVQUE, ML and REML procedures or some of their modifications are known to have desirable properties. To adapt the above type of procedures for a nonlinear mixed effects model, suitable transformations of the model or approximations to the likelihood function are required, as seen in the previous sections. In addition, we note that C.R.Rao [13] presents a nonlinear model for the two-way classification, along with the approximations to the ANOVA type of sums of squares and the corresponding estimation procedure.

As expected, some of the approximations have been found to perform well under suitable conditions. For the computations, programs such as SAS NLINMIX can be used. For some types of nonlinear models, acceptable estimates are obtained through the two-stage procedures of the type described in Section 6. However, it was mentioned that these procedures may require a large amount of computer time to obtain confidence limits for the fixed effects. For the different types of nonlinear mixed effects models employed in practical situations, further research is required to evaluate the suitability of the transformations and estimation procedures briefly summarized in this article.

Acknowledgments

The authors would like to thank Dr. Yogendra Chaubey for his interest in this topic and for suggestions on the first draft of this paper. Thanks also to Joan Robinson for patiently and promptly typing this article in LaTeX.

References

1. M. Boulanger and L.A. Escobar, *Technometrics* **36**, 260 (1994).

2. M.B. Carey and R.H. Koenig, *IEEE Transactions on Reliability* **40**, 499 (1991).
3. M. Gumpertz and S.G. Pantula, *Journal of the American Statistical Association* **87**, 201 (1992).
4. H.O. Hartley and J.N.K. Rao, *Biometrika* **54**, 93 (1967).
5. D.A. Harville, *Journal of the American Statistical Association* **72**, 320 (1977).
6. D.A. Harville, In *Advances in Statistical Methods for Genetic Improvement of Livestock* (Springer-Verlag, New York, 1990).
7. C.R. Henderson, *Biometrics* **9**, 226 (1975).
8. H.S. Lee and Y. P. Chaubey, *Communications in Statistics, Theory and Methods* **25 (6)**, , (1375)1996.
9. M.J. Lindstrom and D.M. Bates, *Biometrics* **46**, 673 (1990).
10. C.J. Lu and W.Q. Meeker, *Technometrics* **35**, 161 (1993).
11. J.C. Pinheiro and D.M. Bates, *Journal of the Computational Graphical Statistics* **4**, 12 (1995).
12. R.Q. Ramos and S.G. Pantula, *Statistics and Probability Letters* **24**, 49 (1995).
13. C.R. Rao, *Linear Statistical inference and its Applications*, 2nd Edition (John Wiley and Sons, New York, 1973).
14. C.R. Rao and J. Kleffé , *Estimation of Variance Components and Applications* (North Holland Publishing Co, Amsterdam, 1988).
15. P.S.R.S. Rao, *Variance Components Estimation: Mixed Models, Methodologies and Applications* (Chapman & Hall/CRC Press, 1997).
16. M. Rudemo, D. Ruppert and J.C. Streibig, *Biometrics* **45**, 349 (1989).
17. R. Schall, *Biometrika* **78**, 719 (1991).
18. L.B. Sheiner and S.L. Beal, *Journal of Pharmacokinetics and Biopharmacokinetics* **8**, 553 (1980).
19. P.J. Solomon, *Biometrika* **72**, 233 (1985).
20. P.J. Solomon and D.R. Cox, *Biometrika* **79**, 1 (1992).
21. Z. Shun and P. McCullagh, *Journal of the Royal Statistical Society* **B57**, 749 (1995).
22. E.F. Vonesh, *Biometrika* **83**, 447 (1996).
23. E.F. Vonesh and R.L. Carter, *Biometrics* **48**, 1 (1992).
24. R.D. Wolfinger, *Biometrika* **80**, 791 (1993).
25. R.D. Wolfinger and X. Lin, *Computational Statistics and Data Analysis* **25**, 465 (1997).

THE SAMPLING BIAS OF HECKMAN'S SAMPLE BIAS ESTIMATOR

PAUL RILSTONE
York University, Department of Economics, North York, Canada, M3J 1P3
E-mail: pril@yorku.ca

AMAN ULLAH
University of California, Riverside, Department of Economics, Riverside, CA, 92521, U.S.A.
E-mail: aman.ullah@ucr.edu

Heckman's [4] two-step estimator for sample selection models can be poorly behaved in some situations due to a form of multicollinearity. The high dispersion of the estimator in these contexts can be deduced by inspection of the asymptotic covariance matrix. However, large–sample theory suggests that the estimator is asymptotically unbiased. In this paper, we derive the second–order bias of Heckman's estimator and demonstrate that this is substantial in similar circumstances. This is reflected in reported simulations.

1 Introduction

Numerous authors, including Wales and Woodland [14], Nelson [5], Paarsch [8], Nawata and Nagase [6], have provided substantial practical and simulation evidence that Heckman's [4] two–step estimator for sample selection models can be poorly behaved in certain circumstances. This typically occurs when the variables in "Heckman's lambda" are a subset of the other explanatory variables, resulting in a form of multicollinearity. The impact of this on the dispersion of Heckman's estimator can be seen by inspection of the asymptotic covariance matrix of the estimator. However, the standard distributional results from large–sample theory indicate that the estimator is asymptotically unbiased, implying that "on average" the estimator is correct. In this paper we derive the second–order bias of Heckman's estimator and demonstrate that, in finite samples, the bias can also be quite substantial.

We consider the basic sample selection model which consists of a linear regression equation (referred to as the "wage equation" given its predominant use in this sort of analysis) and a decision equation which determines whether the dependent variable of the wage equation is observed. To estimate the parameters of this model it is usually assumed that the disturbances from

these two equations have a joint normal distribution. [a] The parameters of the decision equation are estimated in a first–stage probit approach. The inverse Mills's ratio or "Heckman's lambda" is evaluated at the estimated parameter values and this is then inserted as an additional regressor into the wage equation.

Several researchers have considered the effects on the sampling properties of Heckman–like corrections when the normality assumptions are relaxed. Goldberger [3] and Arabmazar and Schmidt [1,2] have shown that the simple Tobit estimator may be biased in the presence of nonnormality or heteroscedasticity. Schafgans [12] has shown that this bias carries over to Heckman's model, as one would expect. Our concern is with the bias of this estimator, even in the context of a correctly specified model. The results we derive are applicable to more general models with the normal distribution of the disturbances of the decision equation replaced by another. Our results depend on the nonlinearities in the model and multicollinearity problems, rather than with any particular distributional assumptions.

The discussion proceeds as follows. In Section 2 we derive the second–order bias of the Heckman estimator. In Section 3 we examine the finite sample bias with a simulated model and apply the bias corrections to a data set. Section 4 summarizes the conclusions.

2 Derivation of the second–order bias

The higher–order properties of nonlinear estimators have been considered by several authors including Pfanzagl and Wefelmeyer [9], Skovgaard [13] and Rilstone, Srivastava and Ullah [11]. In this paper we find it convenient to use the approach in Rilstone, Srivastava and Ullah [11] to derive the second–order bias of Heckman's estimator. The equation of interest is written

$$Y_i = X_i'\alpha + \epsilon_i \tag{1}$$

where Y_i is observed only if a latent variable $Z_i^* = W_i'\gamma + \nu_i$ is greater than zero. X_i and W_i are $k_x \times 1$ and $k_w \times 1$ observable vectors of explanatory variables. Also observed is an indicator variable $Z_i = 1\,[W_i'\gamma + \nu_i > 0]$. ϵ_i and ν_i have a joint normal distribution with correlation coefficient ρ. ϵ_i has variance σ_ϵ^2 and the variance of ν_i is normalized to one.

[a] Olsen [7] showed that Heckman's estimator is still consistent if the joint normality assumption is relaxed to an assumption of normality of the disturbance term in the decision equation combined with linearity of the expectation of the wage equation conditional on the decision equation disturbance.

It is well known that, conditional on $Z_i = 1$, ϵ_i has a truncated normal distribution and that $\mathcal{E}[\epsilon_i \mid Z_i = 1] = \rho\sigma_\epsilon\lambda_i$ where

$$\lambda_i = \frac{\phi(W_i'\gamma)}{\Phi(W_i'\gamma)} \tag{2}$$

is the inverse Mills's ratio, ϕ and Φ denoting the density and distribution functions of a standard normal random variable. We denote the first– and second–order derivatives of λ_i by $\lambda_i^{(1)}$and $\lambda_i^{(2)}$ respectively, noting that $\lambda_i^{(1)} = -\lambda_i(\lambda_i + W_i'\gamma)$. It is also useful to define

$$\overline{\lambda}_i = \frac{\phi(W_i'\gamma)}{(1 - \Phi(W_i'\gamma))}. \tag{3}$$

Augmenting the right hand side of (1) with the inverse Mills's ratio we have the regression:

$$Y_i = X_i'\alpha + \eta\lambda(W_i'\gamma) + u_i \tag{4}$$

where $\eta = \rho\sigma_\epsilon$ and u_i is a residual. Note that $\mathrm{Var}[u_i \mid Z_i = 1] \equiv \sigma_i^2 = \sigma_\epsilon^2(1 + \rho^2\lambda_i^{(1)})$. The score function for probit estimation of γ is

$$S_i(\gamma) = (Z_i\lambda_i - (1 - Z_i)\overline{\lambda}_i)W_i. \tag{5}$$

Heckman's estimator for this model may be written as the solution to

$$\frac{1}{N}\sum q_i(\widehat{\beta}) = 0 \tag{6}$$

where

$$q_i(\beta) = \begin{pmatrix} Z_i X_i u_i(\beta) \\ Z_i\lambda(W_i'\gamma)u_i(\beta) \\ S_i(\gamma) \end{pmatrix} \tag{7}$$

and $\beta = (\alpha', \eta, \gamma')'$ is a $k \times 1$ vector of parameters with $k = k_x + k_w + 1$.

The expressions derived in Rilstone, Srivastava and Ullah [11] are functions of the derivatives of $q_i(\beta)$ and their expectations. We indicate the $k \times k$ and $k \times k^2$ matrices of first– and second–order derivatives by $\bigtriangledown q_i(\beta)$ and $\bigtriangledown^2 q_i(\beta)$[b] respectively. The second–order bias is obtained via a second–order stochastic expansion of the difference between $\widehat{\beta}$ and the true value of β and then taking the expectation of that difference. In Rilstone, Srivastava and Ullah [11] it is

[b]We note that these are defined such that the l'th row of $\bigtriangledown q_i(\beta)$ contains the gradient vector of the l'th element of $q_i(\beta)$, the l'th row of $\bigtriangledown^2 q_i(\beta)$ contains the vectorized Hessian matrix of the l'th element of $q_i(\beta)$.

shown that the second–order bias of estimators solving moment equations of the form given in equation (6) can be written

$$B(\widehat{\beta}) = \frac{1}{N}\left(\mathcal{E}\left[\nabla q_i\right]\right)^{-1}\left\{\mathcal{E}\left[V_i d_i\right] - \frac{1}{2}\mathcal{E}\left[\nabla^2 q_i\right]\mathcal{E}\left[d_i \otimes d_i\right]\right\} \tag{8}$$

in which

$$V_i = \nabla q_i - \mathcal{E}\left[\nabla q_i\right], \qquad d_i = \left(\mathcal{E}\left[\nabla q_i\right]\right)^{-1} q_i. \tag{9}$$

Sufficient conditions under which $B(\widehat{\beta})$ is a valid expression for the second–order bias are discussed in Rilstone, Srivastava and Ullah [11] and consist of smoothness restrictions on the model as well as existence, uniformly in β, of the fourth moments of the random variables appearing in $\nabla q_i(\beta)$ and $\nabla^2 q_i(\beta)$. Since $q_i(\beta)$ is infinitely differentiable in this case, the smoothness conditions are satisfied. We assume that the moment conditions are satisfied.

The matrix of first derivatives is

$$\begin{aligned}\nabla q_i\left(\beta\right) &= \begin{pmatrix} Z_i X_i \nabla u_i\left(\beta\right) \\ Z_i\left(\lambda_i \nabla u_i\left(\beta\right) + u_i\left(\beta\right) \nabla \lambda_i\right) \\ \nabla S_i(\gamma) \end{pmatrix} \\ &= \left(\begin{pmatrix} -Z_i X_i X_i' \\ -Z_i \lambda_i X_i' \\ 0 \end{pmatrix}\begin{pmatrix} -Z_i X_i \lambda_i \\ -Z_i \lambda_i^2 \\ 0 \end{pmatrix}\begin{pmatrix} -Z_i \eta X_i \lambda_i^{(1)} W_i' \\ -Z_i\left(u_i\left(\beta\right) - \eta\lambda_i\right)\lambda_i^{(1)} W_i' \\ (Z_i\lambda_i^{(1)} - (1-Z_i)\overline{\lambda}_i^{(1)}) W_i W_i' \end{pmatrix}\right).\end{aligned} \tag{10}$$

Since u_i has zero mean, the expectation of this last expression may be written

$$\begin{aligned}\mathcal{E}\left[\nabla q_i\right] &= \mathcal{E}\left(\begin{pmatrix} -X_i X_i' Z_i \\ -\lambda_i X_i' Z_i \\ 0 \end{pmatrix}\begin{pmatrix} -X_i \lambda_i Z_i \\ -\lambda_i^2 Z_i \\ 0 \end{pmatrix}\begin{pmatrix} -\eta X_i \lambda_i^{(1)} W_i' Z_i \\ -\eta \lambda_i \lambda_i^{(1)} W_i' Z_i \\ -\lambda_i \overline{\lambda}_i W_i W_i' \end{pmatrix}\right) \\ &= -\mathcal{E}\begin{pmatrix} X_i^* X_i^{*'} Z_i & \eta\lambda_i^{(1)} X_i^* W_i' Z_i \\ 0 & \lambda_i\overline{\lambda}_i W_i W_i' \end{pmatrix}.\end{aligned} \tag{11}$$

where $X_i^* = (X_i', \lambda_i)'$. The inverse of (11) may be written as

$$\left(\mathcal{E}\left[\nabla q_i\right]\right)^{-1} = -\begin{pmatrix} a & b \\ 0 & c \end{pmatrix}, \tag{12}$$

where

$$\begin{aligned} a &= \left(\mathcal{E}\left[X_i^* X_i^{*'} Z_i\right]\right)^{-1} - \eta\left(\mathcal{E}\left[X_i^* X_i^{*'} Z_i\right]\right)^{-1}, \\ b &= \mathcal{E}\left[\lambda_i^{(1)} X_i^* W_i' Z_i\right]\left(\mathcal{E}\left[\lambda_i\bar{\lambda}_i W_i W_i'\right]\right)^{-1}, \\ \text{and } c &= \left(\mathcal{E}\left[-\lambda_i\overline{\lambda}_i W_i W_i^{'}\right]\right)^{-1}. \end{aligned}$$

We also have

$$\mathcal{E}\left[q_i q_i'\right] = \begin{pmatrix} \mathcal{E}\left[\sigma_i^2 X_i^* X_i^{*'} Z_i\right] & \rho\mathcal{E}\left[\lambda_i^2 X_i^* W_i' Z_i\right] \\ \rho\mathcal{E}\left[\lambda_i^2 W_i X_i^{*'} Z_i\right] & \mathcal{E}\left[\lambda_i \overline{\lambda}_i W_i W_i'\right] \end{pmatrix}. \tag{13}$$

With expressions (12) and (13) it is useful to review what is known about the asymptotic distribution of $\widehat{\beta}$. Heckman [4] showed that $\sqrt{N}(\widehat{\beta} - \beta) \rightarrow \mathrm{N}(0, \mathcal{V})$ where, using our notation, we can write

$$\mathcal{V} = (\mathcal{E}[\bigtriangledown q_i])^{-1}\mathcal{E}[q_i q_i'](\mathcal{E}[\bigtriangledown q_i])^{-1'} \equiv \mathcal{E}[d_i d_i'] \tag{14}$$

By inspection, we can characterize when the standard errors associated with $\widehat{\beta}$, that is the square roots of the diagonal elements of $\mathcal{V}$, may be large. We concentrate on the first $k_x + 1$ elements of $\mathcal{V}$. There are several standard situations, such as when the X_i's are collinear and/or σ_ϵ^2 is large. What is of interest are those situations which are particular to the sample selection model: there are two such interesting cases. The standard errors will be large ($\mathcal{E}[\bigtriangledown q_i]$ approaches singularity) when the columns of X_i^* are closely collinear; in particular when λ_i can be written as a linear combination of the X_i's. Since λ_i can be well approximated by a linear function of its argument,[c] $W_i'\gamma$, collinearity will be a problem when the W_i's are a subset of the X_i's or are highly correlated with them.[d]

The second instance where the asymptotic standard errors may be large is, again by inspection, when ρ is large. This corresponds to situations when there is a large degree of simultaneity in the model. The hypothesis that ρ is zero is one that is often tested via significance tests on the estimator of η. We note that both of these situations when the standard errors may be large are quite common in empirical studies. There is often very little theoretical justification for assuming that, for example, there are elements of W_i which are not contained in X_i or that there is no simultaneity between the wage and decision equations.

As is evident from equation (8), the second–order derivatives of the moment equations may make a substantial contribution to the second–order bias. We derive explicit representations for these in the following equations. Put

$$\bigtriangledown^2 q_i(\beta) = \begin{pmatrix} q^{11} & q^{12} & q^{13} \\ q^{21} & q^{22} & q^{23} \\ q^{31} & q^{32} & q^{33} \end{pmatrix},$$

[c] Tobin recognized this in his original work on his model.

[d] Some authors state as an identification condition that the W_i's contain at least one element not in the X_i's. Strictly speaking this is not necessary, although from a practical perspective, if this is not the case, one can expect to have large standard errors.

$$q^{11} = 0_{k_x \times (k_x k)}, \qquad q^{12} = \left(0_{k_x \times k_x} \quad 0_{k_x \times 1} \quad - X_i \lambda_i^{(1)} W_i'\right),$$

$$q^{13} = \left(0_{k_x \times (k_x k_w)} \quad - X_i \lambda_i^{(1)} W_i' \quad - \eta X_i \lambda_i^{(2)} \left(W_i' \otimes W_i'\right)\right),$$

$$q^{21} = \left(0_{1 \times k_x^2} \quad 0_{1 \times k_x} \quad - \lambda_i^{(1)} \left(X_i' \otimes W_i'\right)\right), \qquad q^{22} = \left(0_{1 \times k_x} 0_{1 \times 1} - 2\lambda_i \lambda_i^{(1)} W_i'\right),$$

$$q^{23} = \left(-\lambda_i^{(1)} \left(W_i' \otimes X_i\right) - \lambda_i \lambda_i^{(1)} W_i' D\right),$$

$$D = \left(u_i(\beta)\lambda_i^{(2)} - \eta(\lambda_i^{(1)})^2 - \lambda_i \lambda_i^{(2)} - (\lambda_i^{(1)})^2\right)\left(W_i' \otimes W_i'\right),$$

$$q^{31} = \left(0_{k_w \times k_x^2} \quad 0_{k_w \times k_x} \quad 0_{k_w \times k_w k_x}\right), \qquad q^{32} = \left(0_{k_w \times k_x} \quad 0_{k_w \times 1} \quad 0_{k_w \times k_w}\right),$$

$$q^{33} = \left(0_{k_w \times k_x k_w} \quad 0_{k_w \times k_w} \quad (Z_i \lambda_i^{(2)} + (1 - Z_i)\overline{\lambda}_i^{(2)}) W_i (W_i' \otimes W_i')\right).$$

It is readily shown that

$$\mathcal{E}\left[\nabla^2 q_i\right] = \mathcal{E}\begin{pmatrix} & q^{11} & & (0_{k_x \times k} 0_{k_x \times 1} - X_i \lambda_i^{(1)} W_i') & \overline{q}^{13} \\ \left(0_{1 \times k^2} \quad 0_{1 \times k} \quad - \lambda_i^{(1)} \left(X_i' \otimes W_i'\right)\right) & & \left(0_{1 \times k} 0_{1 \times 1} - 2\lambda_i \lambda_i^{(1)} W_i'\right) & \overline{q}^{23} \\ & q^{31} & & q^{32} & \overline{q}^{33} \end{pmatrix},$$

where

$$\overline{q}^{13} = \mathcal{E}\left(0_{k_x \times (k_x k_w)} \quad X_i \lambda_i^{(1)} W_i' \quad \eta X_i \lambda_i^{(1)} \left(W_i' \otimes W_i'\right)\right),$$

$$\overline{q}^{23} = \mathcal{E}\left(-\lambda_i^{(1)} (W_i' \otimes X_i) \quad - \lambda_i \lambda_i^{(1)} W_i' (-\lambda_i \lambda_i^{(2)} \quad - (\lambda_i^{(1)})^2)(W_i' \otimes W_i')\right),$$

$$\overline{q}^{33} = \mathcal{E}\left(0_{k_w \times k_x k_w} \quad 0_{k_w \times k_w} \quad (\Phi_i \lambda_i^{(2)} + (1 - \Phi_i)\overline{\lambda}_i^{(2)}) W_i \left(W_i' \otimes W_i'\right)\right),$$

$$\left(\mathcal{E}\left[\nabla q_i\right]\right)^{-1} = \begin{pmatrix} H^{11} & H^{11} & H^{13} \\ H^{21} & H^{22} & H^{23} \\ 0 & 0 & H^{33} \end{pmatrix}.$$

Since

$$\mathcal{E}\left[V_i d_i\right] = \mathcal{E}\left[\nabla q_i \left(\mathcal{E}\left[\nabla q_i\right]\right)^{-1} q_i\right]$$

and

$$\left(\mathcal{E}\left[\bigtriangledown q_i\right]\right)^{-1} q_i = \begin{pmatrix} H^{11}X_i u_i + H^{12}\lambda_i u_i + H^{13}S_i \\ H^{21}X_i u_i + H^{22}\lambda_i u_i + H^{23}S_i \\ H^{33}S_i \end{pmatrix},$$

$$\mathcal{E}\left[V_i d_i\right]$$

$$= \begin{pmatrix} -\mathcal{E}\left[X_i X_i' H^{13}S_i + X_i\lambda_i H^{23}S_i + X_i\lambda_i^{(1)}W_i' H^{33}S_i\right] \\ -\mathcal{E}\left[\lambda_i X_i' H^{13}S_i + \lambda_i^2 H^{23}S_i + \lambda_i\lambda_i^{(1)}W_i' H^{33}S_i - u_i(\beta)\lambda_i^{(1)}W_i' H^{33}S_i\right] \\ \mathcal{E}\left[(Z_i\lambda_i^{(1)} - (1-Z_i)\overline{\lambda}_i^{(1)})W_i W_i' H^{33}S_i\right] \end{pmatrix}.$$

We can now examine in what circumstances the second–order bias may be large. It is not possible to put a sign on the terms of this bias, but several qualitative observations may be made in this respect. There are two terms in $B(\widehat{\beta})$, both scaled by $\left(\mathcal{E}\left[\bigtriangledown q_i\right]\right)^{-1}$. The first term, $\mathcal{E}\left[V_i d_i\right]$, is essentially the correlation between the residual $\bigtriangledown q_i - \mathcal{E}\left[\bigtriangledown q_i\right]$ and q_i. The second term is a linear combination of $\mathcal{E}\left[d_i \otimes d_i\right]$, where the weights are functions of the Hessian of $q_i(\beta)$. Note that $\mathcal{E}[d_i \otimes d_i]$ is simply the vectorization of $\mathcal{V}$, the asymptotic covariance matrix of $\widehat{\beta}$. Therefore, in a qualitative sense, the second–order bias will be large in those cases where the estimators have large standard errors. From above, we note that this corresponds to situations of near non–identification and simultaneity. (The parameters of the wage equation are identified if there is a unique minimum to the sum of squared residuals from the wage equation. This will not be the case if λ_i is linear and W_i is contained in X_i.) The first problem is very similar to that of multicollinearity, but this is not a completely accurate depiction of the problem, since even in the presence of severe (but not perfect) collinearity, least squares estimates are unbiased, even if poorly behaved.

3 Simulation and Empirical Results

To get a sense for how well the second–order bias derived in Section 2 reflects the sampling bias of the Heckman estimators we conducted a small set of simulation experiments consisting of a wage equation in the form of equation (1) and a decision equation. Each has an intercept and one conditioning variable. The disturbances are constructed as jointly normal. This was parameterized so that there were five regression parameters including the coefficient on λ_i and the variance from the wage equation. (This was set equal to unity. Recall that the variance term from the decision equation is normalized to one

in these models.) The intercept terms are set to zero, the coefficients on X_i and W_i set to one. We altered two components of the model across the experiments. First for obvious reasons, the correlation between the disturbances was set at a variety of values. Second, as noted, the bias and variance of the parameters of interest depend on the collinearity of the X_i's and λ_i. Since the latter is roughly linear in its argument, the correlation between the W_i's and the X_i's is crucial. The W_i's were constructed as uniform on [-1,1]. The X_i 's were then constructed as a linear combination mixture of the W_i's and another independent, mean zero, uniform random variable: $X_i = \rho_{xw} W_i + v_i$. The support of the distribution of v_i was altered across experiments in order to keep the variance of X_i constant and equal to that of W_i, across the experiments.

The observed sampling properties of the estimators are summarized in the following tables for a variety of values for ρ and ρ_{xw} as well as sample sizes of $N = 50$, 100. We focus on the estimates of the regression parameters of the wage equation and consider their properties relative to what would be an unbiased least squares estimator (albeit infeasible in this case), a "Heckman estimator" with the true values of γ used in the wage equation.[e] The tables are self-explanatory. The sampling bias (*i.e.* averaging over the 1000 replications of the experiment) of the estimates is given in Table 1. We note that these reflect the analytical results of the previous section: the biases are absolutely increasing functions of ρ and ρ_{xw}, and decreasing functions of the sample size. Note that the intercept and the coefficient on the inverse Mills's ratio are particularly affected. Since it is often the case in practice that the X_i's and the W_i's in these kinds of models contain almost the same random variables, we would often be in situations corresponding to ρ_{xw}=.75. With the sample size $N = 50$, the bias can be particularly large.

Table 2 gives the mean squared error of the estimates of the usual Heckman estimators, relative to the infeasible estimators. As would be expected, these are also increasing functions of the correlation parameters.

To provide an empirical example we used a data set of 50 observations taken from the 1990 Integrated Public Use Microdata series employed in a study of white–black wage differentials by Chia–Hui Chiu. The variables in the decision equation consisted of an intercept, marital status and working status. The dependent variable of the wage equation was log wages, the explanatory variables consisted of an intercept and experience. This is undoubtably a rather simplistic model, but one that qualitatively reflects many which are

[e]The comparison could have been made with other benchmark estimators, notably efficient maximum likelihood estimators, but we feel that this benchmark is more comparable. Moreover, the MLE is not necessarily unbiased.

Table 1. Sampling Bias of $\widehat{\beta}$.

Parameter	$N = 50$			
	$\rho,\ \rho_{xw}$			
	.25, .25	.25, .75	.75, .25	.75, .75
α_1	-0.02882	-0.03893	-0.09583	-0.11844
α_2	-0.00258	-0.00222	0.00019	0.00152
η	0.03511	0.04815	0.11268	0.13676
γ_1	-0.00165	-0.00165	-0.00165	-0.00165
γ_2	0.05878	0.05878	0.05878	0.05878
Parameter	$N = 100$			
	$\rho,\ \rho_{xw}$			
	.25, .25	.25, .75	.75, .25	.75, .75
α_1	0.00222	-0.00540	-0.01935	-0.03019
α_2	0.00491	0.00746	0.00637	0.01001
η	-0.00216	0.00730	0.02655	0.03942
γ_1	0.00380	0.00380	0.00380	0.00380
γ_2	0.02578	0.02578	0.02578	0.02578

used in empirical work where identification might be rather tenuous. Table 3 reports the usual point estimates and standard errors as well as the bias estimates. Since the bias depends on the true parameters and population moments, the bias estimates themselves are based on the point estimates and sample moments. [f] We notice that the bias estimates are quite large relative to the point estimates and the standard errors. Such a situation could lead one to make substantially different inferences.

One final point should be made here. One should interpret these estimates very carefully. It is well known that higher–order approximations can be very poor, be they based on stochastic expansions as done here or via other techniques such as Edgeworth expansions. Phillips and Park [10] have a discussion of this. It is a reasonable conjecture that this corresponds to situations where the higher terms in the expansions make a strong contribution to the sampling biases. In the course of the Monte Carlo experiments we calculated the analytic second–order biases and bootstrap estimates of the biases. Both of these approaches performed very poorly, particularly in those cases when the results in Section 2 would indicate potentially high biases.

[f] In principle, using these estimated values rather than true population values should not effect the rate of convergence of the bias estimate since the estimates themselves are $\sqrt{N}$–consistent.

Table 2. Sampling Inefficiency of $\widehat{\beta}$

Parameter	$N = 50$			
	ρ, ρ_{xw}			
	.25, .25	.25, .75	.75, .25	.75, .75
α_1	2.09845	1.95023	4.16079	3.61555
α_2	1.00134	1.00014	1.00092	0.99973
η	1.67572	1.72793	2.87769	2.78589
γ_1	1.00000	1.00000	1.00000	1.00000
γ_2	1.00000	1.00000	1.00000	1.00000
Parameter	$N = 100$			
	ρ, ρ_{xw}			
	.25, .25	.25, .75	.75, .25	.75, .75
α_1	1.23581	1.18577	1.39542	1.27143
α_2	0.99946	0.99866	0.99979	1.00024
η	1.67572	1.72793	2.87769	2.78589
γ_1	1.00000	1.00000	1.00000	1.00000
γ_2	1.00000	1.00000	1.00000	1.00000

Table 3. Parameter point and bias estimates, standard errors

	Point Estimate	Bias Estimate	Standard Error
Parameter			
α_1	-0.39561	-2.70729	0.59017
α_2	0.06672	-0.12628	0.05453
η	5.65843	13.81740	2.92799
γ_1	-0.28100	-0.47202	0.73259
γ_2	0.64172	0.24726	0.27583
γ_3	0.12921	0.17037	0.25398

4 Conclusion

This paper has derived the second–order bias of Heckman's sample selection estimator. The message, from the analytical results which are reflected in the simulations, is that this bias can be large when the estimators have large standard errors and/or the parameters are poorly identified. The lesson for applied researchers is that there is even more reason to be careful in interpreting point estimates in these situations.

Acknowledgments

The authors would like to thank the workshop participants at the University of British Columbia, the 1996 Berkeley/ NSP Conference on Bootstrapping

and the India and Southeast Asia Meeting of the Econometric Society for helpful remarks. Research funding for Rilstone and Ullah was provided by the Natural Sciences and Engineering Research Council of Canada and Academic Senate,UCR, respectively.

References

1. A. Arabmazar and P. Schmidt, *Journal of Econometrics* **17**, 253 (1981).
2. A. Arabmazar and P. Schmidt, *Econometrica* **50**, 1055 (1982).
3. A.S. Goldberger, in *Studies in Econometrics, Time Series and Multivariate Statistics*, eds. S. Karlin, T. Amemiya and L.A. Goodman (John Wiley, New York, 1983).
4. J.J. Heckman, *Econometrica* **47**, 153 (1979).
5. R.D. Nelson, *Journal of Econometrics* **24**, 181 (1984).
6. K. Nawata and N. Nagase, *Econometric Reviews* **15**, 387 (1996).
7. R.J. Olsen, *Econometrica* **48**, 1099 (1980).
8. H.J. Paarsch, *Journal of Econometrics* **24**, 197 (1984).
9. J. Pfanzagl and W. Wefelmeyer, *Journal of Multivariate Analysis* **8**, 1 (1978).
10. P.C.B. Phillips and Joon Y. Park, *Econometrica* **55**, 1065 (1988).
11. Paul Rilstone, V. K. Srivastava and Aman Ullah, *Journal of Econometrics* **75**, 369 (1996).
12. Marcia M.A. Schafgans, *Semiparametric estimation of a sample selection model: A simulation study* (London School of Economics and Political Science, Department of Economics, 1996).
13. I.M. Skovgaard, *Scandinavian Journal of Statistics* **8**, 227 (1981).
14. T. J. Wales and A.D. Woodland, *International Economic Review* **21**, 437 (1980).

COMPUTATIONAL SEQUENCE ANALYSIS: GENOMICS AND STATISTICAL CONTROVERSIES

PRANAB K. SEN

Department of Biostatistics and Statistics, University of North Carolina, Chapel Hill, NC 27599-7420, USA

E-mail: pksen@bios.unc.edu

In genomics, and more generally, in computational biology, principles of molecular genetics govern computational sequence analysis, providing room for stochastics to comprehend the basic differences between mathematical exactness and biological diversity. With a large number of sites having categorical (qualitative) responses with imprecise interrelationships, conventional (discrete or continuous) multivariate statistical modeling and analysis may encounter roadblocks of various kinds. Limitations of likelihoods and their variants are appraised in this context. Alternative approaches that take into account underlying biological implications to a greater (and parametrics to a lesser) extent are appraised in the light of validity and robustness perspectives.

1 Introduction

At the dawn of bioinformatics (and *genomic* science too), *biostochastics* is in an interdisciplinary phase. The conventional approach of planning biomedical studies (in a reasonably controlled setup), formulating statistical models (from an existing pool of standard ones), and carrying out standard statistical analysis may no longer be universally adoptable in bioinformatics setups. At the present it is not precisely known what constitutes the core of bioinformatics. We may quote from a very recent text by Ewens and Grant [9]:

> We take bioinformatics to mean the emerging field of science growing from the application of mathematics, statistics, and information technology, including computers and the theory surrounding them, to study and analysis of very large biological and in particular, genetic data sets. The field has been fueled by the increase in the DNA data generation.

As such, within the broader setup of bioinformatics, we consider here some aspects of computational sequence analysis (CSA) pertaining to human *genomes*. Principles of molecular genetics, as well as, various environmental factors govern CSA. At the current stage, gene scientists can not scramble fast enough to keep up with the genomics, emerging at a furious rate and in astounding detail. The data dusts need to be settled in their proper places

before methodological issues can be viewed in a proper perspective. CSA, and bioinformatics in general, do not aim to lay down fundamental mathematical laws that govern biological systems parallel to those laid down in physics. Such laws, if they exist, are a long way from being determined for biological systems. Biodiversity, extreme variability in human immunity, and genetic complexities (in number and activity) have rather confounding impacts on such biological (theoretical) laws. There is, at this stage, mathematical utility in the creation of tools that investigators can use to analyze biological systems data. Because of underlying stochastic evolutionary forces, such tools involve statistical modeling of biological systems, which in turn, requires the incorporation of probability theory, statistics, and stochastic processes. Although *knowledge discovery and data mining* (KDDM) procedures (or statistical learning, by another name, Hastie *et al.* [11]) are increasingly being used in CSA, it might not be proper to jump on statistical conclusions based on data analysis alone (Cox [7], Breiman [5]) unless the algorithms could be justified from statistical perspectives. There is a genuine need to grasp the genetic and molecular biologic bases of CSA, and in the light of that to formulate statistical modeling and analysis schemes. Primarily driven by this motivation, we (Sen [21]) designate *Biostochastics to deal with stochastic modeling and analysis (i.e., stochastics) for very large biological (including genetic and genomic) data sets.* As such, biostochastics extends to large biological systems which may not have predominant genetic factors; neuronal spatio-temporal model (Sen [22])s are noteworthy examples. In this study, however, we confine ourselves to biostochastics pertaining to CSA that has a dominant genetic and genomics flavor; but our inclination is to emphasize the underlying methodology with a view to facilitate applications.

With a large number of sites, with categorical responses, statistical modeling may become quite complex, for which classical likelihood approaches may stumble into conceptual as well as computational difficulties. Conventional (discrete or continuous) multivariate analysis may also encounter roadblocks due to excessively high dimensionality as well as imprecise specification of underlying dependence patterns. Without an acceptable topology that defines *neighborhoods*, there might not be enough incentive for standard statistical modeling and analysis. Alternative approaches that take into account underlying biological implications to a greater (and parametrics to a lesser) extent are appraised on *validity* and *robustness* considerations.

Section 2 outlines the molecular biological background, and based on this motivation, statistical approaches are outlined in Section 3. Variants of likelihoods, such as the pseudolikelihood, are discussed in Section 4. Section 5 deals with some nonparametrics along with some general remarks.

2 Molecular Biological Perspectives

Within the broader domain of computational biology, we pay especial attention to CSA pertaining to human genomes. *Gene*, in the Mendelian setup, is the basic unit of inheritance. Genes occur at definite sites or *loci*, on *chromosomes*, which are strings of DNA (Deoxyrinucleic acid), the basic genetic material in a *cell* and the carrier of genetic information for all organisms, except for some viruses. DNA is a double-helical model; it is a polymer, madeup of nucleotides which are four in number, and can be distinguished by the four bases: A (adenine), C (cytosine), G (guanine) and T (thymine). Like the DNA, Ribonucleic acid (RNA) and proteins are also macromolecules of a cell, though they differ in their forms and constitution. Like DNA, RNA is a nucleic acid, but with T replaced by U (uracil), and it has a single strand. Proteins are also polymers, and there are 20 amino acids. Most human cells contain 46 chromosomes, in 23 pairs; one pair relates to the *sex* chromosomes, while the other 22 homologous pairs are termed *autosomes*. There are about 30,000 to 40,000 genes embedded within the human genome. *Genetic data*, for duploid organisms, relate to traits determined by autosomal Mendelian loci, so that DNA plays a basic role in genetic data analysis (Waterman [25], Lange [14], Ewens and Grant [9]).

Principles of molecular genetics, such as the central dogma that DNA makes RNA makes protein, govern CSA. The transfer of genetic information from DNA to DNA (called *replication*) means that the molecule can be copied; the loop from DNA to RNA called *transcription* precedes the loop from RNA to protein, called *translation*. The RNA which is translated into protein is termed the *messenger* RNA (or ${}_mRNA$), and the *transfer* RNA (or ${}_tRNA$) translates the genetic code into amino acids. If we accept the basic role of DNA as the genetic information carrier, then it is natural to conclude that evolution is directly related to changes in DNA This is the genesis of molecular evolution. Substitutions, such as $A \leftrightarrow G$ or $C \leftrightarrow T$ are called *transitions*, while $A \leftrightarrow C$, $A \leftrightarrow T$, $G \leftrightarrow C$, and $G \leftrightarrow T$ are called *transversions* (recall that in a DNA, A pairs with T and G pairs with C).

Physiochemical studies, electron microscopy, and X-ray diffraction analyzes have established that most DNA molecules are long, flexible, threadlike structures, having a nearly constant diameter with regularly spaced and repeated structures, irrespective of the base composition (A, C, G, T) or their order. In the usual form of the double-helix structured DNA, the *purine* and *pyrimidine* rings in each chain are stacked 0.34nm, i.e., every ten base pairs. There are two external grooves: major (wider) and minor (narrower). Next, note that amino acids are encoded by triplets of nucleotides, called *codons*. Let

us define $\mathcal{N}_R = \{A, C, G, U\}$, and let $\mathcal{C} = \{(x_1, x_2, x_3) : x_j \in \mathcal{N}_R, j = 1, 2, 3\}$ be the codon. Finally, let $\mathcal{A}$ be the set of aminoacids and termination codon. Then the *genetic code* can be defined as a map: $g : \mathcal{C} \to \mathcal{A}$, $g \in \mathcal{G}$, so that $\mathcal{G}$ is the set of all genetic codes. As the human genome project is heading for a completion, there are some formidable statistical tasks which are surfacing into the research efforts to fathom out the mystery of the genomic code.

3 Statistical Motivation

Most problems in CSA are essentially statistical. Amidst a *chaos of random mutation* and *natural selection*, stochastic evolutionary forces act on genomes, resulting in *genetic drifts*, and thus requiring stochastic modeling and analysis for quantitative studies. In genomic sequence analysis, typically, we encounter data on a large number (K) of positions or *sites*, and in each position, we have a purely qualitative (nucleotides or amino acid labels) categorical response with 4 to 20 categories depending on the DNA or protein sequence. the spatial (functional as well as stochastic) dependence (or association) patterns of these sites may not be known, nor can they be taken to be *stochastically independent*. Also, as has been mentioned before, regular and nearly identical structures of the DNA solicitate statistical appraisal based on other variational properties which exhibit more statistical variation and information too. In this way, we are more in the domain of biostochastics than biomathematics or purely computer algorithms.

In this high-dimensional qualitative response setup, it is difficult to incorporate standard (discrete or continuous) multivariate analysis tools, in a parametric formulation (as the number of associated parameters may be exceedingly large and the underlying model may not be that well specified or anticipated). As such a (complete/partial/conditional/profile/pseudo) likelihood approach may be hopeless, unless there is a sample, drawn objectively, of enormously large size. On both counts, we may have difficulties in adopting conventional tools. If we restrict ourselves to a single site, in most cases (particularly, for viral sequences), there is little statistical information. In a multiple site context, treating the sites as independent could lead to serious misspecification of the model. As such, we need to consider high-dimensional qualitative categorical data models preserving intersite dependence, and then to proceed to CSA statistical appraisals. There has been some attempts to incorporate suitable *quasi-likelihood* methods based on generalized estimating equations (GEE) and invoking *Markov chain monte Carlo* (MCMC) methodology that amends easily to Gibbs sampling and Metropolis-Hastings algorithms (Durbin *et al.* [8]). However, these procedures are yet to have a solid

methodological foundation in such a complex setup. In this study, we confine ourselves to some specific models, and appraise variants of likelihoods and alternative approaches towards their resolutions.

4 CSA : Likelihoods and Alternatives.

Let us consider a setup of K sites, and let $\mathbf{X} = (X_1, \ldots, X_K)'$ be the response (stochastic) vector, where each X_j is categorical with anywhere between 4 and 20 qualitative categories. If we denote the joint probability function of $\mathbf{X}$ by $p(\mathbf{x})$, we may express it as

$$p(\mathbf{x}) = p(x_1) \prod_{j=2}^{K} p(x_j | x_i, i < j). \tag{1}$$

If these sites were ordered in some way and had a *Markov chain* (MC) property, then we could have written

$$p(\mathbf{x}) = p(x_1) \cdot p(x_2|x_1) \cdots p(x_K|x_{(K-1)}). \tag{2}$$

Things could have been even simpler if these conditional probabilities $p(x_j|x_{(j-1)})$ were the same for all $j = 2, \ldots, K$. However, lacking such an ordering, a MC property can not be taken for granted, and on top of that the stationarity of the transition probabilities may need critical appraisals. If we are able to validate the MC property, but stationarity may not hold, then the number of parameters associated with the joint probability law jumps from $C(C+1)-$ to $C-1+(K-1)C^2$, and this in turn may demand an excessively large same size (as K is large) to justify conventional statistical methods to be valid and efficient. In the more likely event, we do not even have an ordered index set $\mathcal{J} = \{1, \ldots, K\}$, so that it may not be feasible to adopt a MC model.

Of prime scientific interest is the statistical comparison of sets of genomic sequences from *human immunodeficiency virus* (HIV). It may be noted that a *retrovirus*, like HIV, has the ability to reverse the normal flow of genetic information from genomic DNA, and that the genetic variability of HIV is relatively high compared to other retroviruses. In this way, we have a spatial (or spatio-temporal) model in a general multivariate analysis of variance (MANOVA) setup. Typically, there are molecular epidemiologic studies of genomic sequences that pertain to different epidemiologic strata, so that *external CSA* like MANOVA becomes pertinent. On the other hand, comparing the *variability* at different sites for the same individual (or their *covariability*) constitutes the *internal CSA* (like the canonical analysis in multivariate

statistical analysis). In this way, we encounter both MANOVA and canonical analysis type problems in a very high dimensional categorical data model.

The conventional *product multinomial* model can be adopted to formulate suitable likelihood functions. With K sites and $C(\geq 4)$ categories for each site response, we have typically C^K cells, so that the number of parameters involved in a single multinomial law is equal to $C^K - 1$, and if there are G groups, this number jumps to $G(C^K - 1)$. Clearly, in this setup, with large K the above number increases exponentially (even for $G = 1$ or 2 and $C = 2$. This in turn requires the sample size(s) astronomically large in order that standard discrete multivariate analysis tools (Agresti [1], Sen and Singer [23]) can be validly used to draw statistical conclusions from acquired data sets (satisfying suitable objective sampling schemes). Faced with this roadblock, one may naturally wonder whether the model can be represented in terms of a comparatively smaller number of parameters, and suitable *pseudolikelihood* formulations can be made instead of the classical likelihood one to avoid some of these difficulties. We present some of these in a very simple setup, and appraise their suitability in CSA.

For the K-site model, we introduce a vector $\mathbf{Y} = (Y_1, \ldots, Y_K)'$ where Y_k is equal to 1 or 0 according as there is a mutation at site k or not, for $k = 1, \ldots, K$. Also, let $\Omega = \{(i_1, \ldots, i_K) : i_k = 0, 1,\ k = 1, \ldots, K\}$; note that the cardinality of Ω is 2^K. Further let $P(\mathbf{y}) = P\{\mathbf{Y} = \mathbf{y}\}, \mathbf{y} \in \Omega$. Let us define then $Q(\mathbf{y}) = \log\{P(\mathbf{y})/P(\mathbf{0})\}$, so that by definition,

$$P(\mathbf{y}) = e^{Q(\mathbf{y})} / \sum_{\mathbf{z} \in \Omega} e^{Q(\mathbf{z})}. \tag{3}$$

Led by a basic representation of multivariate binary random variables due to Bahadur [3], Liang, Zeger and Qaqish [15] advocated the following representation:

$$P(\mathbf{y}) = \exp\{\sum_{k=1}^{K} u_k y_k + \sum_{1 \leq s < k \leq K} y_s y_k u_{sk} + \ldots + y_1 \ldots y_K u_{1\ldots,K}\}, \tag{4}$$

where the u_k are the conditional logits, u_{sk} are the conditional log -odd ratio, etc.. If in this representation, we let all the second and higher order interaction coefficients ($u_{ijk..}$) to be null, we end up with a pairwise dependence model wherein

$$Q(\mathbf{y}) = \sum_{k=1}^{K} \alpha_k y_k + \sum_{1 \leq s < k \leq K} \gamma_{sk} y_s y_k, \tag{5}$$

where the α_k and γ_{sk} are respectively the main effect and first order interactions.

We conceive of n independent sequences $\mathbf{Y}_i$, $i = 1, \ldots, n$ and denote the $K \times n$ matrix of observations by $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$, and denote its kth row (vector) by $\mathbf{Y}_n(k)$, $k = 1, \ldots, K$. Then, following Besag [4], we formulate a *pseudolikelihood function* as

$$L_P(\boldsymbol{\theta}, \mathbf{Y}) = \prod_{k=1}^{K} P(\mathbf{Y}_n(k) = \mathbf{y}_k | \mathbf{Y}_s(n), s \neq k, \boldsymbol{\theta}), \tag{6}$$

where $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_K, \gamma_{12}, \ldots, \gamma_{K-1,K})$. Writing $\gamma_{kk} = 0, k = 1, \ldots, K$, we may show that for this pairwise dependence model,

$$L_P(\boldsymbol{\theta}, \mathbf{Y}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ \frac{e^{y_{ik}(\alpha_k + \sum_{j=1}^{K} \gamma_{kj} y_{ij})}}{1 + e^{y_{ik}(\alpha_k + \sum_{j=1}^{K} \gamma_{kj} y_{ij})}} \right\}. \tag{7}$$

This is termed the *autologistic model.* It is not clear how in general, can this pseudolikelihood function be interpreted as a conditional, partial or even profile likelihood function ?

The *maximum pseudolikelihood estimator* (MPLE) $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ is a point of maxima of the pseudolikelihood function. This MPLE can be taken more in the spirit of generalized M-estimators (as it is based on a model different from the likelihood function); moreover, using the structure in (7), it can be shown that the MPLE is consistent and asymptotically normal. However, because of the basic model difference, robustness perspectives of such MPLEs need to be assessed properly. Further, for the same reason, the MPLE's efficacy is unknown, and expected to be lower than that of the MLE if the latter were obtainable from the data. Moreover, this also raises the robustness issues for the MPLE for model departures and distortion due to the choice of the particular autologistic model. For the MPLE, there are computational difficulties, and moreover, there is no direct way of obtaining the (asymptotic) dispersion matrix of the MPLE (which depends on the unknown likelihood function). The deficiency of the MPLE stems from the fact that the associated estimating equations (EE) are not isomorphic (even asymptotically) to the EE for the MLE. Also, the observed information (matrix) may not be available in this formulation. Thus, there are some genuine concern over the use of MPLE on the ground of robustness and efficiency. From the EE perspectives, one may use the results in Liang *et al.* [15] to study large sample properties of the MPLE, although for large values of K, it remains to appraise how far such asymptotic properties are tenable.

We may notice that in the above formulation of the autologistic function

$$P(\boldsymbol{\theta}, \mathbf{Y}) = C(\boldsymbol{\theta}) F(\mathbf{Y}, \boldsymbol{\theta}), \tag{8}$$

where $C(\boldsymbol{\theta}) = P(\mathbf{Y} = \mathbf{0}, \boldsymbol{\theta})$ and

$$\log F(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \{\alpha_k + \sum_{j=k+1}^{K} \gamma_{kj} y_{ij}\}. \tag{9}$$

This representation makes it possible to use the *simulated likelihood ratio* technique, using the Gibbs sampling or the Metropolis algorithm (or more generally the Hastings algorithm) to generate a simulated sequence of random vectors to carry out the maximization on simulated data. Note that the above sketched MCMC procedure, advocated by Geyer and Thompson [10], is highly computation incentive, and as in the current CSA setup, usually we have a very large dimensional $\boldsymbol{\theta}$ even such MCMC techniques may run into computational complexities.

If we consider internal CSA problems, in an autologistic model, we need to confine our attention to the sub-vector $\boldsymbol{\gamma} = (\gamma_{jk}, 1 \leq j < k \leq K)$, there being $\binom{K}{2}$ such parameters. As such, for large K we have a larger number of parameters, and all the problems referred to in connection with $\boldsymbol{\theta}$ show up in this case too. In addition, for testing independence of the K positions an autologistic model may have good relevance, but for canonical analysis, such models may be questionable. The basic difficulty in incorporating the classical canonical analysis (Anderson [2]) in this context is the lack of linearly combinability of the coordinates of $\mathbf{Y}_i$ so as to justify the basic interpretation of canonical correlations and canonical variates. We may refer to Sen [22] for some discussion of canonical analysis in neuronal spatio-temporal models, and the present setup is quite akin to it.

If we want to have external CSA, such as the MANOVA for several independent groups of sequences, for testing homogeneity of the groups, the autologistic model may serve a good purpose. If we denote the gth group parameter vector by $\boldsymbol{\theta}_g = (\boldsymbol{\alpha}_g', \boldsymbol{\gamma}_g')'$, $g = 1, \ldots .G$, then primarily we may be interested in testing for the homogeneity of the $\boldsymbol{\alpha}_g$. As in the normal theory MANOVA, we may be tempted to assume that the $\boldsymbol{\gamma}_g$ are homogeneous, but in view of a lack of parametric orthogonality of $\boldsymbol{\alpha}_g$ and $\boldsymbol{\gamma}_g$ such an assumption may have to carefully appraised. Since these parameters are not directly related to the full likelihood functions, the usual likelihood ratio type test may not work out well in this setup. Pseudolikelihood based inference would be subject to the same criticism as in internal CSA. Wald-type test based on the MPLE of the $\boldsymbol{\alpha}_g$ may conceived of. But, as it is cumbersome to estimate the (asymptotic) covariance matrix of the $\hat{\boldsymbol{\theta}}_g$, a formulation of the Wald-type test statistic may not be that convenient either. Moreover, with such estimated covariance matrices, we may have genuine concern regarding their robustness

properties (against plausible model departures), and as a result, such tests are not expected to be robust. However, it is possible to introduce a suitable permutation test for the homogeneity that we shall consider later on.

Suppose now that instead of the binary response we consider the full model involving C^K possible response vectors for K positions, each with C possible (qualitative) outcomes. Even for a single position, instead of a conventional *logit* model (for binary outcome), we need to consider a generalized logit model for polychotomous responses. For multiple sites, a precise formulation of the pseudolikelihood function becomes quite involved, involving too many parameters. Neither the Bahadur [3] representation nor the Liang , Zeger and Qaqish [15] formulation (for the binary case) provides a good resolution. Besides, the question of robustness, validity and efficacy of such procedures becomes even more pertinent in this general case. As such, we take recourse to suitable alternative external CSA procedure.

Let us now consider a permutation test for homogeneity of G groups of independent sequences in an external CSA setup. Consider G groups, where the gth group consists of n_g independent sequences, for $g = 1, \ldots, G$. In a pseudolikelihood formulation, we consider the autologistic model, and denote the parameter vector for the gth group by $\boldsymbol{\theta}_g$, $g = 1, \ldots, G$. We further let $n = n_1 + \cdots + n_G$, so that if we pool all these groups together, we would have a set of n sequences. Because of the complications in the computation of the (asymptotic) covariance matrix of these estimates, the usual pseudolikelihood approach may stumble into impasses. As such, we deemphasize the likelihood formulation, but, nevertheless use the individual group estimates in a permutation setup to generate suitable (conditional) permutation tests which might have simplicity in formulation and affinity to the Wald type of test based on the MPLE's for the pooled sequences. We denote the ith sequence in the gth group by $\mathbf{Y}_{gi}$, for $i = 1, \ldots, n_g$, $g = 1, \ldots, G$. We denote the collection matrix (of order $K \times n$) for the pooled sample by $\mathbf{Y}^o = (\mathbf{Y}_{11}, \ldots, \mathbf{Y}_{1n_1}, \ldots, \mathbf{Y}_{G1}, \ldots, \mathbf{Y}_{Gn_G})$. Thus under the null hypothesis of homogeneity of the G groups, the columns of $\mathbf{Y}^o$ are independent and identically distributed random vectors (i.i.d.r.v.), each vector can have 2^K possible realizations (over the set of mutation or not, at each site). We write

$$\mathbf{Y}^o = (\mathbf{Y}_1^o, \ldots, \mathbf{Y}_n^o). \tag{10}$$

Let now $\mathbf{R}_n = (R_1, \ldots, R_n)$ be any permutation of $(1, \ldots, n)$, so that there are $n!$ possible realizations of $\mathbf{R}_n$; we denote this set by $\mathcal{R}$. Let then

$$\mathbf{Y}^o(\mathbf{R}_n) = (\mathbf{Y}_{R_1}^o, \ldots, \mathbf{Y}_{R_n}^o), \tag{11}$$

for $\mathbf{R}_n \in \mathcal{R}$. Thus, corresponding to the sample point $\mathbf{Y}^o$, we generate an *orbit* of $n!$ sample points, and we denote this orbit by $\mathcal{O}(\mathbf{Y}^o)$. It follows by standard rank-permutation arguments (Chatterjee and Sen [6]) that under the null hypothesis (H_0) of homogeneity of the G groups,

$$P\{\mathbf{Y}^o = \mathbf{Y}^o(\mathbf{R}_n)|\mathcal{O}(\mathbf{Y}^o)\} = (n!)^{-1}, \tag{12}$$

for every $\mathbf{R}_n \in \mathcal{R}$. This (discrete) uniform permutation (conditional) probability measure is denoted by $\mathcal{P}_n$.

For the autologistic model, the pseudolikelihood function (for the gth group), $L_P^{(g)}(\boldsymbol{\theta}_g, \mathbf{Y}_g)$ is defined as in (7), and this leads to the *pseudo-score statistic* (vector) $\mathbf{U}_{n_g}^{(g)}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}$ as

$$\mathbf{U}_{n_g}^{(g)}(\boldsymbol{\theta}) = \sum_{i=1}^{n_g}\sum_{k=1}^{K}\{(\partial/\partial\boldsymbol{\theta})[(\, y_{ik}^{(g)}\{\alpha + \sum_{j=1}^{K}\gamma_{jk}y_{ij}^{(g)}\}\,) - \log(1+\exp\{y_{ik}^{(g)}\{\alpha_k + \sum_{j=1}^{K}\gamma_{jk}y_{ij}^{(g)}\}\,)]\}, \tag{13}$$

for $g = 1, \ldots, G$. Next, we consider the pooled sample of n sequences, and using (7) we compute the pooled MPLE of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}_n$; we may need to use the Gibbs sampling or MCMC tools for numerically achieving this. But this is to be done only for the pooled sample, not for the G samples separately. Let us then denote the *aligned* pseudo-scores by

$$\hat{\mathbf{U}}_{n_g}^{(g)} = \mathbf{U}_{n_g}^{(g)}(\hat{\boldsymbol{\theta}}_n),\ g = 1, \ldots, G. \tag{14}$$

Note that by construction, the MPLE $\hat{\boldsymbol{\theta}}^n$ is a symmetric function of the n stochastic vectors $\mathbf{Y}_r^o, r = 1, \ldots, n$, and hence, the MPLE is $\mathcal{P}_n$-invariant, i.e., for every point on the orbit $\mathcal{O}(\mathbf{Y}^o)$, the MPLE remains the same. Therefore, under the null hypothesis, for each $g(= 1, \ldots, G)$, the n_g terms appearing as summands of $\hat{\mathbf{U}}_{n_g}^{(g)}$ are marginally identically distributed and they are interchangeable or exchangeable random vectors (across the G groups as well). As such, we express the aligned pseudo-scores as

$$\hat{\mathbf{U}}_{n_g}^{(g)} = \sum_{i=1}^{n_g}\mathbf{U}(\mathbf{Y}_{gi}, \hat{\boldsymbol{\theta}}_n) = \sum_{i=1}^{n_g}\hat{\mathbf{U}}_{n_g,i},\ \text{ say, } g = 1, \ldots, G, \tag{15}$$

where under the permutation setup, the MPLE is invariant, and hence, permutation distribution of the $\hat{\mathbf{U}}_{n_g}^{(g)}$ can be generated by the same permutation law , formulated in (12), but applied to the $\hat{\mathbf{U}}_{n_g,i}$ instead of the $\mathbf{Y}_{gi}$.

We define the pooled sample vector $\hat{\mathbf{U}}_n^o$ in the same manner as in $\mathbf{Y}^o$, and this way, the orbit $\mathcal{O}(\mathbf{Y}^o)$ is mapped onto an orbit $\mathcal{O}(\hat{\mathbf{U}}_n^o)$, and hence, the permutation law $\mathcal{P}_n$ is generated by the column permutations of $\mathbf{R}_n$ over the set $\mathcal{R}$, i.e., permutations of $(1, \ldots, n)$. Note that by virtue of the choice of the MPLE, we have

$$\sum_{i=1}^{n} \hat{\mathbf{U}}_i^o = \mathbf{0}, \tag{16}$$

so that we have

$$E_{\mathcal{P}_n}\{\hat{\mathbf{U}}_{ng}^{(g)}\} = \mathbf{0}, \; g = 1, \ldots, G. \tag{17}$$

Let us further define the permutation pseudo-score covariance matrix by

$$\mathbf{V}_n = \frac{1}{n-1} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \hat{\mathbf{U}}_{n_g,i} \hat{\mathbf{U}}'_{n_g,i}, \tag{18}$$

and note that this matrix is $\mathcal{P}_n$-invariant. Also, let $\mathbf{\Lambda}_n$ be a $G \times G$ matrix with elements

$$\lambda_{gg',n} = (n_g(n\delta_{gg'} - n_{g'}))/n, \; g, g' = 1, \ldots, G, \tag{19}$$

where δ_{ij} is the Kronecker delta (i.e., equal to 1 or 0 according as $i = j$ or not). Then, by standard argument (as in Chatterjee and Sen [6]), we obtain on writing the rolled-out vector $\hat{\mathbf{U}}_n = (\hat{\mathbf{U}}_{n_1}^{(1)'}, \ldots, \hat{\mathbf{U}}_{n_G}^{(G)'})'$ that

$$E_{\mathcal{P}_n}(\hat{\mathbf{U}}_n \hat{\mathbf{U}}'_n) = \mathbf{V}_n \bigotimes \mathbf{\Lambda}_n, \tag{20}$$

where $\bigotimes$ refers to the Kronecker product of the two matrices. Note that $\mathbf{\Lambda}_n$ is of rank $G - 1$, and noting that $n^{-1}\mathbf{\Lambda}_n$ has the same structure as the multinomial sample covariance matrix, its generalized inverse can be taken as

$$\mathbf{\Lambda}_n^- = \mathrm{diag}(\frac{1}{n_1}, \ldots, \frac{1}{n_G}) - \frac{1}{n}\mathbf{1}\mathbf{1}'. \tag{21}$$

Note that $\boldsymbol{\theta}$ has $K(K+1)/2 = M$ (say) elements, so that the pseudo-scores $\hat{\mathbf{U}}_{n_g}$ are M-dimensional vectors, and as a result $\mathbf{V}_n$ is an $M \times M$ matrix. We denote the generalized inverse of $\mathbf{V}_n$ by $\mathbf{V}_n^-$, and note that under fairly general regularity conditions (as the elements of $\boldsymbol{\theta}$ are linearly independent), $\mathbf{V}_n$ is positive definite, in probability, as n increases, and hence, it is of full rank in probability. Further, note that by (16),

$$\mathbf{1}'\hat{\mathbf{U}}_n = 0, \tag{22}$$

so that we may consider the following quadratic form as a suitable (Wald-type) aligned test statistic:

$$\mathcal{L}_n = \sum_{g=1}^{G} \frac{1}{n_g} \hat{\mathbf{U}}'_{n_g} \mathbf{V}_n^- \hat{\mathbf{U}}_{n_g}. \tag{23}$$

Note that essential for the computation of the test statistic $\mathcal{L}_n$ is the MPLE $\hat{\boldsymbol{\theta}}_n$ which can be obtained by using the algorithms mentioned before. Once this is done, the computation of the aligned pseudo scores vectors $\hat{\mathbf{U}}_{n_g}$ and the covariance matrix $\mathbf{V}_n$ does not pose any horrendous task, so that like the Rao score test statistic in the classical likelihood ratio type tests, computationally the major task is the MPLE for the pooled sample. In this sense, $\mathcal{L}_n$ is more like an aligned M-statistic; we refer to Rao and Sen [19] for the general asymptotics for such aligned pseudo-score statistics, considered in a directional data model setup.

For finite sample sizes, the exact permutation distribution of $\mathcal{L}_n$, under $\mathcal{P}_n$, can be obtained by direct enumeration. This task becomes prohibitively laborious as the n_g increase. For this reason, there is a genuine need to consider suitable large sample approximation to the permutation distribution of $\mathcal{L}_n$ In the present setup, note that the $Y_{ij}^{(g)}$ are binary, and hence, the coordinates of each $\hat{\mathbf{U}}_{n_g,i}^{(g)}$ are all bounded random variables. This makes it possible to use the celebrated Wald-Wolfowitz-Noether-Hoeffding-Motoo permutational central limit theorem on the individual pseudo-scores statistics. Omitting these details of algebraic manipulations (quite similar to those in Puri and Sen [18] (Ch. 5, p. 186)), we arrive at the following :

When the individual sample sizes $n_g, g = 1, \ldots, G$ all increase, subject to the condition that $n_g/n \to \rho_g, g = 1, \ldots, G$, where the ρ_g are positive numbers adding upto one, the permutational (conditional) distribution of $\mathcal{L}_n$ converges, in probability, to the central chi-square distribution with degrees of freedom (DF) $M^* = (G-1)K(K+1)/2$.

Based on this basic convergence result, asymptotically, the permutation test can be replaced by an unconditional test based on $\mathcal{L}_n$ having the central chi-square distribution with M^* DF (under the null hypothesis of homogeneity of the G groups). Consistency properties and asymptotic power under local alternatives follow conventional lines of arguments, and hence, the details are omitted.

In the context of genomics, usually K is large, and hence, (even for G as low as 2 or 3) M^* may be quite large. For example, often K is in the range of 200 to 500, so that even for $G = 2$, M^* may be in the range of 20,000 to 75,000. Even, for K as small as 50, and $G = 2$, M^* is more than 750. With

larger DF, the accuracy of the chi-square approximation to the permutation distribution of the test statistic may generally require increasing large sample sizes. Although in some cases in CSA this can be justified, in general, it might create some roadblocks. One alternative is to partition the parameter vector $\boldsymbol{\theta}_g = (\boldsymbol{\alpha}_g', \boldsymbol{\gamma}_g')'$, for each $g = 1, \ldots, G$, and then to test for the null hypothesis of homogeneity of the $\boldsymbol{\alpha}_g$ assuming further that the $\boldsymbol{\gamma}_g$ are homogeneous. This is quite analogous to the multivariate normal compound symmetry testing problem (Anderson [2]) where the overall hypothesis H_{MVC} is partitioned into H_M and H_{VC}, and while testing for the null hypothesis H_M (of homogeneity of the means), the homogeneity of the variances and covariances is presumed, although here we are to deal with multivariate binary data model where the parametric orthogonality does not hold, and exact likelihood formulations are difficult. Thus, we need to proceed through appropriate permutation formulations with suitable alignments. In the present case, again we need to consider the MPLE of $\boldsymbol{\theta}$ from the pooled sample (of all the n sequences), and then consider the pseudo scores pertaining to the $\boldsymbol{\alpha}$ component, for individual groups, aligned at the pooled MPLE. The rest of the manipulations are quite similar, and the resulting test statistic, denoted by $\mathcal{L}_n^{1)}$, would have then a permutation distribution which is asymptotically, in probability, chi-square with $(G-1)K$ DF.

5 Nonparametrics and General Comments

As has been mentioned earlier, the autologistic model may lack proper statistical justifications from genuine likelihood perspectives. As such, routinely using the autologistic model may lead to nonrobust and inefficient statistical resolutions, and this is particularly of serious concern when K is large. Further, in the more general case of very high dimensional categorical data with C categories for each position (and $C \geq 4$), an autologistic model needs considerable modifications (to accommodate such categorical outcomes, instead of binary ones), resulting in a much larger number of associated parameters. While such a case may still be treated in a permutation setup, much of the charm of the autologistic model would disappear, and the inefficiency and nonrobustness aspects may even be more pertinent. Because of this reason, we consider some alternative procedures that attach less emphasis on the likelihood approach, and more on alternative measures that deal with similar homogeneity problems. We may refer to Pinheiro *et al.* [16,17] for some resolutions, and we will mainly add some further comments to their work.

Consider a general CSA with K sites, each one having a categorical response with $C(\geq 2)$ qualitative categories, indexed as $1, \ldots, C$. For the ith

sequence, let $\mathbf{X}_i = (X_{i1}, \ldots, X_{iK})'$ be a random vector of responses where X_{ik} denotes the category outcome $c(= 1, \ldots, C)$ at site $k(= 1, \ldots, K)$. Recalling that these sites may not be stochastically independent, we need to have a measure of divergence which takes into account the inter-site stochastic dependence to a certain extent. The primary motivation for using a diversity measure stems from the fact that HIV or some other retrovirus have the ability to have higher mutation rates which can be traced with a diversity index, without going through some likelihood formulations. With that in mind, we define the *Hamming distance* between a pair (i, i') of sequences as

$$D_{ii'} = \frac{1}{K} \sum_{k=1}^{K} I(X_{ik} \neq X_{i'k}), \tag{24}$$

so that $D_{ii'}$ is the proportion of sites where $\mathbf{X}_i$ and $\mathbf{X}_{i'}$ do not match. Since the K coordinate indicator functions are not necessarily independent, this measure attempts to take into account their dependence, albeit in a symmetric manner. It is easy to see that the expected value of $D_{ii'}$ is the average (over the K positions) *Gini-Simpson* [24] diversity indexes, which we denote by Δ_H. It is also possible to employ other measures of diversity which have nonparametric flavour; for details, we refer to Pinheiro *et al.* [16]).

It may be remarked that an optimal nonparametric estimator of Δ_H is the Hoeffding [12] U-statistic (corresponding to the kernel D_{ij} of degree 2):

$$\bar{D}_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} D_{ij}. \tag{25}$$

This U-statistic formulation enables us to use conventional statistical tools for testing homogeneity of the Δ_H for different groups. In conventional way of using ANOVA procedures based on U-statistics, such tests for homogeneity of the Δ_H for the different groups were considered by Pinheiro *et al.* [16,17]; their test is based on the 'between group' and 'within group' sum of squares of the sample Hamming distances. Since the kernel is bounded, asymptotic results were not so difficult to derive. However, because of the fact that under the null hypothesis of homogeneity, for the ANOVA test statistic based on the divergence of the statistics in (25), we have stationarity of order 1 (Hoeffding [12]), and as a result, the asymptotic null hypothesis distribution is not a conventional chi-squared or variance ratio distribution, and its use needs further simulation results.

Following the decomposition of the Gini-Simpson index (Sen [20]), we consider here a somewhat different formulation. For the gth group of n_g independent sequences, we define the sample Hamming distance as in (25) and denote

it by $\bar{D}_{n_g}^{(g)}$, for $g = 1, \ldots, G$. Also, we pool the G groups into a combined set of n sequences, and compute the Hamming distance, denoted by $\bar{D}_n^{(o)}$. Note that the Hamming distances are nonnegative quantities, bounded from above by 1, and it is possible to express

$$\begin{aligned}\bar{D}_n^{(o)} &= \text{within group component} + \text{between group component} \\ &= \sum_{g=1}^{G} (n_g/n)\bar{D}_{n_g}^{(g)} + D_n(B), \text{ say,} \end{aligned} \tag{26}$$

where $D_n(B)$ is nonnegative and is stochastically small when the G groups of sequences are homogeneous. Denoting the first term on the right hand side of (26) as $D_n(W)$, we may compare the two and formulate a test statistic as

$$\mathcal{L}_H = D_n(B)/D_n(W). \tag{27}$$

The same permutation principle invoked in Section 4 can be incorporated to find the critical level. However, for large sample sizes, such a test may not have a desired chi-square or normal approximation (mainly due to the fact that $D_n(B)$ when standardized may not be asymptotically normal or chi-square). For that reason, some alternative test procedures have been considered by Sen [21]; these have simple asymptotic distributions under the permutation setup, as well as, unconditionally.

For the internal CSA, as a very first step, Karnoub *et al.* [13] formulated a conditional test for independence of mutations in the case of $K = 2$. This area merits specific statistical attention to handle the case of general K. Likelihoods are difficult to be formulated precisely, and hence, autologistic or other models may stumble into conceptual difficulties. Nonparametrics have a much better prospect. High-dimensionality and categorical data setups pose the basic challenges for such resolutions.

Acknowledgments

This work was supported by the Cary C. Boshamer Professorship Research Fund at the University of North Carolina, Chapel Hill. The author is grateful to the reviewers for their helpful comments on the manuscript.

References

1. A. Agresti, *Categorical Data Analysis* (Wiley, New York, 1990).
2. T. W. Anderson, *An Introduction to Multivariate Statistical Analysis* (Wiley, New York, 1958)

3. R.R. Bahadur, In *Studies in Item Analysis and Prediction*, ed. H. Solomon (Stanford University Press, Calif., 1961).
4. J. Besag, *J. Roy. Statist. Soc.* **B 48**, 192 (1974).
5. L. Breiman, *Ann. Statist.* **28**, 374 (2000).
6. S.K. Chatterjee and P.K. Sen,*Calcutta Statist. Assoc. Bull.* **13**, 18 (1964).
7. D.R. Cox, *J. Roy. Statist. Soc.* **A 158**, 455 (1995).
8. R. Durbin, S. Eddy, A. Krogh and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models for Proteins and Nucleic Acids* (Cambridge University Press, UK, 1998).
9. W.J. Ewens and G.R. Grant, *Statistical Methods in Bioinformatics: An Introduction*, (Springer, New York, 2001).
10. C.J. Geyer and E.A. Thompson, *J. Amer. Statist. Assoc.* **90**, 909 (1995).
11. T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, (Springer, New York, 2001).
12. W. Hoeffding, *Ann. Math. Statist.* **19**, 293 (1948).
13. M.C. Karnoub, F. Seillier-Moiseiwitsch and P.K. Sen, *Inst. Math. Statist. Lect. Mono. Ser.* **33**, 221 (1999)
14. K. Lange, *Mathematical and Statistical Methods in Genetic Analysis*, (Springer, New York, 1997).
15. K. Liang, S.L. Zeger and B. Qaqish, *J. Roy. Statist. Soc.* **B 54**, 3 (1992).
16. H. Pinheiro, F. Seillier-Moiseiwitsch, P.K. Sen and J. Eron, In *Handbook of Statistics, Volume 18 : Bioenvironmental and Public Health Statistics*, eds. P. K. Sen and C. R. Rao (Elsevier, Amsterdam, 2000).
17. H. Pinheiro, F. Seillier-Moiseiwitsch and P. K. Sen, In preparation (2001).
18. M. L. Puri and P. K. Sen, *Nonparametric Methods in Multivariate Analysis* (Wiley, New York, 1971).
19. C.R. Rao and P.K. Sen, *J. Nonpar. Statist.* (to appear, 2002).
20. P.K. Sen, *Calcutta Statist. Assoc. Bull.* **49**, 1 (1999).
21. P.K. Sen, *Excursions in Biostochastics: Biometry to Biostatistics to Bioinformatics*, (Lecture Notes, Academia Sinica Inst. Statist. Sci., Taipei, ROC, 2001).
22. P.K. Sen, *Scientiae Mathematicae Japonicae 55* (to appear, 2002).
23. P.K. Sen and J.M. Singer(1993). *Large Sample Methods in Statistics: An Introduction with Applications* (Chapman-Hall, UK, 1993).
24. E.H. Simpson, *Nature* **163**, 688 (1949).
25. M. Waterman, *Introduction to Computational Biology: Maps, Sequences and Genomes* (Chapman and Hall, UK, 1995).

UNIVERSAL OPTIMALITY OF COMPLETELY RANDOMIZED DESIGNS

KIRTI R. SHAH
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
E-mail: kshah@uwaterloo.ca

BIKAS K. SINHA
Indian Statistical Institute, Kolkata, India
E-mail: bksinha@isical.ac.in

It is reasonable to expect that in a completely randomized design when the number of experimental units is not a multiple of the number of treatments, the most symmetrical allocation would be optimal. However, the available proofs are w.r.t. specific optimality criteria and are not always straightforward. In this paper, we show that the most symmetrical allocation is universally optimal both for the estimation of the treatment effects as well as for the estimation of the treatment contrasts.

1. Introduction

We consider the problem of comparing v treatments using n experimental units when the design adopted is the completely randomized design (CRD). When n is a multiple of v, it is easy to see that the design where every treatment is applied to n/v e.u.'s is optimal. When n is **not** a multiple of v, the design where every treatment is applied to $[\frac{n}{v}]$ or to $[\frac{n}{v}]+1$ e.u.'s may be expected to be optimal. Such a design is known as the design with the most symmetrical allocation (MSA)

Known optimality results for the MSA relate only to the specific optimality criteria. For some criteria such as the distance optimality criterion, the proof is somewhat involved (Liski *et al.* [2]).

In this paper, we use a theorem of Shah and Sinha [4] to show that the MSA is in fact universally optimal (UO) both for the estimation of the treatment effects as well as for the estimation of the treatment contrasts.

2. Universal Optimality

The concept of universal optimality (UO) was first introduced by Kiefer [1]. The motivation is that for some settings a particular design is optimal w.r.t. very many optimality criteria. In such cases it would be useful to try to establish

optimality w.r.t. a **class** of optimality criteria. This would obviate the need for establishing optimality w.r.t. specific criteria. When the class of criteria is the set of **all** criteria satisfying some minimal reasonable conditions, if a design is optimal w.r.t each of these criteria, it is called universally optimal. Kiefer [1] exploited this idea to establish UO of the balanced block designs and of the generalized Youden design within the appropriate design classes.

Kiefer's notion of UO was somewhat modified by Shah and Sinha [3] who used a slightly different set of conditions on the class of optimality criteria. We present these here for the sake of completeness.

Let $\mathcal{D}$ denote the class of designs and for $d \in \mathcal{D}$, let $\boldsymbol{C}_d$ denote the information matrix for the set of effects to be estimated. Let $\phi(\boldsymbol{C}_d)$ denote an optimality functional which we seek to minimize. In what follows, we shall omit the suffix d. The conditions to be satisfied by $\phi(\cdot)$ are

(i) $\phi(\boldsymbol{C}) = \phi(\boldsymbol{C}_g)$ for all $g \in \mathcal{G}$ where $\boldsymbol{C}_g$ is obtained by applying permutation g to the rows and the columns of $\boldsymbol{C}$. (This reflects the symmetry of the problem. The choice of $\mathcal{G}$ depends upon the set up at hand. In most cases, it is the entire permutation group).

(ii) If $\boldsymbol{C}_1 \geq \boldsymbol{C}_2$ i.e. if $\boldsymbol{C}_1 - \boldsymbol{C}_2$ is non-negative definite then, $\phi(\boldsymbol{C}_1) \leq (\boldsymbol{C}_2)$. (This condition is very appealing. In fact, one would wish to impose only this condition and no others. Unfortunately, this leaves the class of designs much too wide which makes it hard to find optimal designs.)

(iii) $\phi(\boldsymbol{C}_1) > \phi(\boldsymbol{C}_2) \Longrightarrow \phi(t\boldsymbol{C}_1) > \phi(t\boldsymbol{C}_2)$ for all positive integers t. (This is reasonable. If d_2 is better than d_1, then t copies of d_2 are better than the same number of copies of d_1).

(iv) $\phi(\sum t_g \boldsymbol{C}_g) \leq \phi((\sum t_g)C)$ where t_g's are non-negative integers. (This is a form of convexity which is very appealing. It is weaker than the usual convexity condition. See Shah and Sinha [3] for a detailed discussion of this).

A well known result due to Kiefer [1] states that if a design d^* is such that (i) $\boldsymbol{C}_{d^*}$ is completely symmetric and (ii) $\boldsymbol{C}_{d^*}$ has a maximum trace, then d^* is UO. (A matrix of the form $\alpha I + \beta J$ is called completely symmetric.) This has been the most widely used result for proving universal optimality. This result is considerably strengthened by the following result due to Shah and Sinha [4].

Theorem 2..1. Let $d^* \in \mathcal{D}$. If for a design d, we can get non-negative integers t_g (not all zero) such that

$$\left(\sum t_g\right) C_{d^*} - \sum t_g C_{dg} \text{ is non-negative definite,}$$

then d^* is at least as good as d w.r.t. every criterion satisfying conditions (i) to (iv).

If this holds for **all** $d \in \mathcal{D}$, d^* is said to be universally optimal. The above theorem is an extension of a result due to Yeh [5] modified for the Shah-Sinha formulation. Yeh required the difference to be a null matrix. In fact, he conjectured that this condition is necessary. However, in our application to CRD, the difference is a non-null matrix.

3. Optimality of MSA CRD's

Here, we have a total of n experimental units to be used for studying the effects of v treatments. If n is divisible by v, it is easy to show that the design where each treatment is equally replicated is UO.

We shall now assume that n is **not** divisible by v. Let n_i denote the number of e.u.'s receiving the treatment i. For the most symmetrical allocation (MSA), $n_i = x$ or $x+1$ where x is the largest integer not exceeding n/v. Thus, $|n_i - n_j| \leq 1$ for all i, j for the MSA design.

We now show that the MSA design is UO. We deal separately with the following two cases.

Case I: Estimation of the treatment effects

Our model here is

$$E(y_{ij}) = \mu_i, j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, v$$
$$V(y_{ij}) = \sigma^2.$$

Here y_{ij} denotes the j-th observation on treatment i and the observations are assumed to be uncorrelated.

It is easy to see that the C-matrix for the estimation of the μ_i's is diag $(n_1, n_2, \ldots, n_v)$.

For the MSA design, let p denote the number of treatments for which $n_i = x+1$. Thus, $n = \sum n_i = p(x+1) + qx$ where $q = v - p$.

For the competing design, we assume that $n_1 \geq n_2 \geq \cdots \geq n_v$. Let

$y = \sum_{1}^{p} n_i/p$ and let $z = \sum_{p+1}^{v} n_i/q$. By successive applications of condition (iv) for optimality criteria, it follows that the design with replication numbers given by $(p!q!)y$ for the first p treatments and $(p!q!)z$ for the remaining q treatments is at least as good as $p!q!$ repetitions of the original design. Thus, it is enough to compare the MSA design with a design with y replications for the first p treatments and z for the others. In view of condition (iii) we may ignore the fact that y and z may not be integers.

We now try to express $(x+1,\ldots,x+1,x,\ldots,x)$ as a convex combination of permutations of $(y,\ldots,y,z,\ldots,z)$ where the coefficients are non-negative rational numbers. Here, the vectors are partitioned in two parts containing p and q terms respectively. We can then apply our theorem to show that the MSA design is superior to $(y,\ldots,y,z,\ldots,z)$ and hence to $(n_1,n_2,\ldots,n_v)$. Here, the term "superior" should be interpreted as "at least as good as" w.r.t. any criterion satisfying conditions (i) to (iv).

We first deal with the case where $p \geq q$. The case where $p < q$ is handled similarly.

We consider the case where q of the y's go to the last q positions and these are replaced by the z's. Number of such permutations is $s = \binom{p}{q}$. We give weight β to each of these and weight α to $(y,\ldots,y,z,\ldots,z)$. All others are given weight zero so that $\alpha + s\beta = 1$. We try to see if this combination gives $(x+1,\ldots,x+1,x,\ldots,x)$. Equating these, we get,

$$\begin{aligned} \alpha y + \beta(tz + (s-t)y) &= x+1 \\ \alpha z + \beta s y &= x \end{aligned}$$

where $s = \binom{p}{q}$ and $t = \binom{p-1}{q-1}$. It can be shown that these have solution $\alpha = (y-x)/(y-z)$ and $\beta = (x-z)/s(y-z)$. Since $y \geq x \geq z$ and since $y > z$, it follows that we can get non-negative integers t_g such that

$$(\sum t_g)(x+1,\ldots,x+1,x,\ldots,x) = \sum t_g(n_1,\ldots,n_v)g$$

where $(n_1,\ldots,n_v)g$ is a permutation of $(n_1,\ldots,n_v)$. This establishes the UO property of the MSA design for the estimation of the treatment effects.

Case II: Estimation of the treatment differences

The same model as above is still applicable. It is easy to see that the

information matrix for the estimation of the contrasts in the μ_i's is given by

$$C_d = diag(n_1, n_2, \ldots, n_v) - \frac{1}{n}\begin{pmatrix} n_1 \\ n_2 \\ \vdots \\ n_v \end{pmatrix}(n_1, n_2, \ldots, n_v)$$

Let d^* denote the MSA design. It is enough to show that given a design d, for a suitable choice of non-negative integers t_g

$$(\sum t_g) C_{d^*} \geq \sum t_g C_{dg}$$

where C_{dg} is obtained by applying permutation g to the rows and columns of the matrix C_d.

Again we assume that $n_1 \geq n_2 \geq \cdots \geq n_v$ and take the same choice of t_g's as in the previous case. If we define $x_d = (n_1, n_2, \ldots, n_v)$, it is enough to show that

$$\sum t_g G_g x_d x_d' G_g' \geq (\sum t_g) x_{d^*} x_{d^*}'$$

where G_g is the permutation matrix for the permutation g. (This follows from the fact that $\sum t_g G_g x_d = (\sum t_g) x_{d^*}$).

We now consider a v-variate distribution where we assign probability $t_g/(\sum t_g)$ to $G_g x_d$ with the t_g's as described above. We have seen that

$$\sum \frac{t_g}{(\sum t_g)} G_g x_d = x_{d^*}$$

Since the dispersion matrix of $G_g x_d$ is non-negative definite, it follows that

$$\sum \frac{t_g}{(\sum t_g)} G_g x_d x_d' G_g' \geq x_{d^*} x_{d^*}'.$$

It now follows that $(\sum t_g) C_{d^*} - \sum t_g C_{dg}$ is a non-negative definite matrix. This establishes the UO property of the MSA design for the estimation of the treatment effects differences.

Remark 3..1. The above establishes the optimality of the MSA with respect to all the known criteria such as A-, D-, E-, MV-, (M,S)- and SD- optimality. Though most of these were known, the derivation for some of these such as SD- optimality were rather involved.

Yeh [5] had conjectured that

$$C_{d^*} - \sum a_g C_{dg} = 0$$

is a necessary condition for the universal optimality of the design d^*. Here, a_g's are non-negative real numbers. Our above result for the CRD shows that the above matrix could be non-null. However, this leaves open the possibility that for some other choice of a_g's the difference matrix is null.

We now give a revised conjecture that for d^* to be universally optimal for any design d, there exist non-negative integers t_g such that

$$\sum t_g C_{d^*} - \sum t_g C_{dg} \text{ is a n.n.d. matrix.}$$

The application to the CRD set up appears to be the first known case where the matrix is non-null.

In any case, The theorem used here is a powerful tool for establishing UO. It is hoped that it will be found useful in many investigations.

Acknowledgment

This work was partially supported by a grant from the Natural Science and Engineering Research Council of Canada. This support is gratefully acknowledged.

References

1. J. Kiefer, In *A Survey of Statistical Design and Linear Models*, ed. J.N. Srivastava (North-Holland, Amsterdam, 1975).
2. E.P. Liski, N.K. Mandal, K.R. Shah and B.K. Sinha, *Topics in Optimal Designs* (To appear in Lecture Notes in Statistics Series, Springer-Verlag, New York, 2001).
3. K.R. Shah and B.K. Sinha, *Canadian Journal of Statistics* **17**, 345 (1989).
4. K.R. Shah and B.K. Sinha, In *Recent Advances in Experimental Designs and Related Topics*, ed. S. Altan and J. Singh (Nova Sciences Publishers, Inc. Huntington, New York, 2001).
5. C.M. Yeh, *Biometrika* **73**, 701 (1986).

A NONPARAMETRIC COMPARISON OF TUMOR INCIDENCES WITH INTERMEDIATE LETHALITY AND DIFFERENT DEATH RATES

JIANGUO SUN, QIANG ZHAO

Department of Statistics, University of Missouri, 222 Math Sciences Building, Columbia MO 65211

E-mail : tsun@stat.missouri.edu

SHESH N. RAI

Department of Biostatistics, St. Jude Children's Research Hospital, 32 N. Lauderdale St., Memphis, TN 38105-2794

The comparison of incidence rates of occult tumors is usually one of main objectives of tumorigenicity experiments. For the problem, a common practice is usually to assume that tumors are lethal or nonlethal (Hoel and Walburg [9]; Sun [17]), to fit incidence rate data to certain parametric or semiparametric models (Dewanji *et al.* [4]), or to treat tumors with intermediate, but known lethality (Lagakos and Louis [11]). In this paper, a simple nonparametric test, which allows tumors to have intermediate and unknown lethality, is proposed for the comparison of incidence rates. The method also allows distributions of death times to depend on treatments or doses. The proposed method is applied to data arising from a tumorigenicity experiment.

1 Introduction

This paper considers statistical analysis of tumorigenicity experiments with focus on the comparison of different treatments or doses with respect to rates of development of tumor (*e.g.*, Dinse and Lagakos [7]; Dewanji and Kalbfleisch [3]; Dinse [6]). In these situations, the variable of interest is usually the time to tumor onset, which is often not directly observable. Instead, only death time of animals under study and the status of tumor onset at the death time are observed. For such treatment comparison problems, an important factor that has to be considered and has a great deal of impact on the comparison is tumor lethality, measuring the effect of tumor onset on the death rate of animals. Two extreme cases are that tumors are lethal and nonlethal, respectively. The former means that tumor onset kills the animal right away, while the latter means that tumor onset has no effect on the death rate and does not alter the risk of death from other causes.

If tumors are lethal or nonlethal, there exist a number of parametric and nonparametric test procedures. This is especially the case for lethal tumors

since tumor onset time and death time can be treated as being identical and thus are observed or right-censored. In other words, many existing survival methods could be directly applied for the comparison of tumor rates in this case (Kalbfleisch and Prentice [10]). For discussion on nonlethal tumors, see, among others, Dinse [6], Dinse and Lagakos [7], Peto and Peto [13], Sun [17] and Sun and Kalbfleisch [18]. If tumor is between lethal and nonlethal, for the comparison, a common method is to fit some parametric or semiparametric three-states regression models to observed data and then to apply the resulting score test (Dinse and Lagakos [7], Dewanji and Kalbfleisch [3]; Rai, Matthews and Krewski [15]). The three-states model includes alive without tumor, alive with tumor and dead. Animals could move from either of the first two states to the last state. A detailed description of the three-state model for tumorigenicity experiments is given in the article by Rai, Sun and Hunt in this volume. The goal of this paper is to develop a nonparametric test procedure without making any model assumption about tumor rates.

To compare tumor rates, another factor that needs to be considered is animal death time, which serves as observation or censoring time and could depend on treatments. A comparison not accounting for animal death time difference could overestimate or underestimate treatment difference (Dinse and Lagakos [7]). For example, consider a situation in which animals in one group have longer survival times and higher tumor rates than animals in the other group. Then tests assuming the same death distribution could overestimate tumor rate difference. In other words, the survival difference if existing needs to be adjusted for treatment comparison.

The remainder of the paper is organized as follows. We begin in Section 2 with introducing notation and assumptions that will be used throughout the paper. Section 3 discusses the comparison of incidence rates of occult tumors with intermediate lethality and presents a simple nonparametric test procedure. For the problem, the Cox [2] proportional hazards model is assumed for tumor lethality and treatment effect on death rates. The proposed methodology allows estimation and test of tumor lethality as well as treatment effect on animal death. The method is a generalization of that proposed in Sun [17], who considered the same problem for nonlethal tumors. In Section 4 the methodology is applied to a real data set from a tumorigenicity experiment. Some discussions and concluding remarks are given in Section 5.

2 Notation and Assumptions

Consider a tumorigenicity experiment involving n independent animals who are tumor-free at time $t = 0$ and are randomly assigned to one of $p + 1$

treatment or dose groups. For the ith animal, let U_i denote the time of tumor onset, T_i^* the time of death and C_i the censoring or sacrifice time. It will be assumed that the development of a tumor is an irreversible event. In practice, since tumors are occult, we do not observe the U_i's. Instead, we observe only $T_i = \min\{T_i^*, C_i\}$, the smallest of death and sacrifice times. Define $N_i(t) = I(U_i \leq t)$, the indicator of the absence ($N_i(t) = 0$) or presence ($N_i(t) = 1$) of a tumor in animal i at time t, and $\tilde{N}_i(t) = I(T_i \leq t)$, indicating if the animal is dead (by 1) at time t, $i = 1, ..., n$. Then observed data are $\{T_i, N_i(T_i), \delta_i = I(T_i = T_i^*); i = 1, ..., n\}$. Let Z_i denote the p-dimensional vector of treatment or dose group indicators associated with the ith animal. Our main goal is to test the hypothesis $H_0 : E\{N_i(t) \mid Z_i\}$ is independent of i.

As mentioned before, in the following, we will focus on tumors that are between lethal and nonlethal. To model tumor lethality, we will assume that given Z_i and U_i, the hazard function of the death times T_i^*'s is given by the following proportional hazards model (Cox, 1972),

$$\lambda_i(t) = \lambda_i(t \mid Z_i) = I(T_i \geq t)\, \lambda_0(t)\, e^{\tau' Z_i + \beta I(U_i \leq t)}, \qquad (1)$$

where $\lambda_0(t)$ denotes the baseline hazard function, τ is a p-dimensional vector of unknown regression parameters characterizing the group or dose effect on death rate, and β is a scalar parameter describing tumor lethality, $i = 1, ..., n$. The above model allows that animal death rates could be affected by both doses and the onset of a tumor. Note that $\beta = 0$ means that the tumor is nonlethal and $\tau = 0$ means that the dose has no effect on the death rate. If both $\beta = 0$ and $\tau = 0$, the death times T_i^*'s then follow the same distribution and are independent of the N_i's and the Z_i's for all animals. Thus the model allows both the assessment and estimation of possible dose effect and tumor lethality.

Models similar to model (1) have been discussed, for example, by Dinse [5], Lindsey and Ryan [12] and Rai and Matthews [14]. In particular, Sun [17] considered the problem of testing the hypothesis H_0 under model (1) with $\beta = 0$, that is, for nonlethal tumors. To test H_0 for current situation, there exist several difficulties compared to the case of $\beta = 0$. One major problem is estimation of the parameter τ, which can be easily obtained if $\beta = 0$, and the parameter β. This is because model (1) involves the unknown tumor onset times U_i's. In the following, we will develop an EM-type algorithm for this. It will be assumed that the censoring times C_i's are independent of tumor onset and death times and that they follow the same distribution as the death times when there is no dose and tumor effect.

3 Statistical Methods

In this section, we discuss the test of the hypothesis H_0 and estimation of parameters τ and β. Let $S_0(t) = \exp\{-\int_0^t \lambda_0(s)\,ds\}$ be the baseline survival function for the T_i^*'s. To construct a test statistic for H_0, first assume that τ, β and S_0 are known. Note that we can rewrite $N_i(T_i)$ as $N_i(T_i) = \int_0^\infty N_i(t)\,d\tilde{N}_i(t)$. Thus under H_0 and the assumptions,

$$E\{N_i(T_i) \mid Z_i\} = \{e^{(\tau' Z_i + \beta)} + 1\} \int_0^\infty \lambda_0(t)\,\{S_0(t)\}^{\exp(\tau' Z_i + \beta)}\,\mu(t)\,S_0(t)\,dt$$

conditional on Z_i, where $\mu(t)$ denotes the mean function of the $N_i(t)$'s under H_0. Therefore, under H_0, we have that

$$E\left\{\frac{N_i(T_i)\,\{e^{(\tau' Z_i+\beta)}+1\}^{-1}}{S_0(T_i)^{\exp(\tau' Z_i+\beta)}} \mid Z_i\right\} = \int_0^\infty \lambda_0(t)\,\mu(t)\,S_0(t)\,dt\,.$$

This motivated the following test statistic

$$X(\tau,\beta,S_0) = \sum_{i=1}^n (Z_i - \bar{Z})\,\frac{N_i(T_i)\,\{e^{(\tau' Z_i+\beta)}+1\}^{-1}}{S_0(T_i)^{\exp(\tau' Z_i+\beta)}}\,, \tag{2}$$

which has expectation zero under H_0, where $\bar{Z} = \sum_{i=1}^n Z_i/n$.

Among others, Charles and Wei [1] discussed similar test statistics for the comparison of two treatments in the context of balanced repeated measurements. Sun [17] considered the test of H_0 and proposed a test statistic similar to $X(\tau,\beta,S_0)$ replacing unknown parameters by their consistent estimates. Let $\alpha = (\tau,\beta,S_0)$. It can be shown using the method in Sun [17] that under H_0, $X(\alpha)$ has asymptotically a normal distribution with mean zero. Thus using the statistic $X(\hat{\alpha})$, the hypothesis H_0 should be rejected for large values of $X(\hat{\alpha})$, where $\hat{\alpha}$ denotes a consistent estimate of α.

To apply the statistic $X(\hat{\alpha})$, we need to estimate α under H_0. For this purpose, note that if the U_i's are known, τ and β can be easily estimated by the solution to equations $X_\tau(\tau,\beta) = 0$ and $X_\beta(\tau,\beta) = 0$, where

$$X_\tau(\tau,\beta) = \sum_{i=1}^n \int_0^\infty \left\{ Z_i - \frac{\sum_{j=1}^n I(t \le T_j)\,e^{\tau' Z_j + \beta I(U_j \le t)}\,Z_j}{\sum_{j=1}^n I(t \le T_j)\,e^{\tau' Z_j + \beta I(U_j \le t)}} \right\} d\,N_i^*(t)$$

and

$$X_\beta(\tau,\beta) = \sum_{i=1}^n \int_0^\infty \left\{ I(U_i \le t) - \frac{\sum_{j=1}^n I(t \le T_j)\,e^{\tau' Z_j + \beta I(U_j \le t)}\,I(U_j \le t)}{\sum_{j=1}^n I(t \le T_j)\,e^{\tau' Z_j + \beta I(U_j \le t)}} \right\} d\,N_i^*(t)\,,$$

which are score functions from the partial likelihood function of τ and β. In the above, $N_i^*(t) = I(T_i \leq t, \delta_i = 1)$ and $N^*(t) = \sum_{i=1}^n N_i^*(t)$. Given τ and β, the baseline survival function S_0 can be conveniently estimated by the consistent estimator

$$\hat{S}_0(t;\tau,\beta) = \exp\left\{ - \int_0^t \frac{d\,N^*(s)}{\sum_{j=1}^n I(s \leq T_j)\, e^{\tau' Z_j + \beta I(U_j \leq t)}} \right\} .$$

This motivates the following algorithm for estimating α under H_0.
Step 0. Choose initial values for the U_i's.
Step 1. Estimate the marginal survival distribution of the tumor onset time under H_0 based on the U_i's and denote it by $\hat{S}_U$.
Step 2. Estimate τ and β by solving the equations $X_\tau(\tau,\beta) = 0$ and $X_\beta(\tau,\beta) = 0$ and then estimate S_0 by $\hat{S}_0(t;\hat{\tau},\hat{\beta})$. This yields an estimate, say $\hat{S}_{T^*|U}$, of the conditional survival function of the T_i^* given the U_i's.
Step 3. Let U_i be equal to its conditional expectation $E\{\, U_i \,|\, T_i, N_i(T_i), \hat{S}_U, \hat{S}_{T^*|U}\,\}$ under the current values of the parameters, $i = 1, ..., n$.
Step 4. Go back to step 1 until the convergence of the estimates of τ, β and S_0.

In the above algorithm, a natural choice for the initial value of U_i is to set $U_i = T_i/2$ if $\delta_i = 1$ and $U_i = T_i$ if $\delta_i = 0$. Steps 2 and 3 correspond to the maximization step of EM algorithm and Step 4 can be regarded as the expectation step of EM algorithm. An alternative to steps 0 and 1 is to estimate S_U based on interval-censored data $[\,(T_i, N_i(T_i)\}\,;\, i = 1, ..., n\,]$ on the U_i's (Sun [16]; Turnbull [19]). In this way, in step 1 after the first iteration, S_U can be estimated by the Kaplan-Meier estimate (Kalbfleisch and Prentice [10]).

Once estimates $\hat{\alpha}$ are obtained, we propose to use the simple bootstrap procedure given in Efron [8] to determine the p-value for testing H_0 using $\hat{X}(\hat{\alpha})$ as well as the distributions and variance estimates of $\hat{\alpha}$. Specifically, the bootstrap procedure can be carried as follows. For a given integer K and each $k = 1, ..., K$,
Step 1. Draw from observed data $\{\, T_i, N_i(T_i), \delta_i\ ,\ i = 1, ..., n\,\}$ with replacement n random samples $\{\, T_i^{(k)}, N_i^{(k)}(T_i^{(k)}), \delta_i^{(k)}\ ,\ i = 1, ..., n\,\}$.
Step 2. Obtain an estimate of α denoted by $\hat{\alpha}^{(k)} = (\hat{\tau}^{(k)}, \hat{\beta}^{(k)}, \hat{S}_0^{(k)})$ and let $\hat{X}^{(k)} = \hat{X}(\hat{\alpha}^{(k)})$.
Step 3. Go back to step 1 if $k < K$.

The distribution of $\hat{\alpha}$ and $\hat{X}(\hat{\alpha})$ can be estimated by $\hat{\alpha}^{(1)}, ..., \hat{\alpha}^{(K)}$ and $\hat{X}^{(1)}, ..., \hat{X}^{(K)}$ for large enough K, respectively. For the test of the hypothesis H_0, the p-value can be calculated as $\sum_{k=1}^K I(|\hat{X}^{(k)}| \geq |\hat{X}(\hat{\alpha})|)\,/K$. Some-

times it is also interest to test $\tau = 0$ and $\beta = 0$, which correspond to no treatment effect on death rate and the tumor under study being nonlethal, respectively. Suppose that we accept H_0. Then for the test of $\tau = 0$, the p-value can be calculated as $\sum_{k=1}^{K} I(|\hat{\tau}^{(k)}| \geq |\hat{\tau}|)/K$ for large K and the same can be done for $\beta = 0$. Note that here both estimation and test about τ and β are performed under H_0, which is the main interest of the paper. The general inference about τ and β could be carried out in the similar way.

4 An Application

To illustrate the proposed method in the previous section, we apply it to the tumorigenicity experiment reported by Hoel and Walburg [9] on lung tumors on 144 male RFM mice. The experiment involves two treatments, conventional environment (96 mice) and germfree environment (48 mice). The observed data include observation times (T_i's) of the mice, lung tumor presence or absence indicators at the observation times ($N_i(T_i)$'s) and treatment indicators. Note that the observation times here are either the death or sacrifice times of the animals. The data were given in Table 5 of Hoel and Walburg [9] and also analyzed by Lagakos and Louis [16] and Sun [17]. One of the objectives of the study was to compare the lung tumor incidence rates of the two treatments.

To compare lung tumor incidence rates, define $Z_i = 0$ if an animal was given conventional environment and $Z_i = 1$ otherwise. Applying the method given in Section 3, we got that $\hat{\tau} = -2.0314$, $\hat{\beta} = 0.4008$ and $X(\hat{\alpha}) = 38.1738$. To obtain the p-value for testing the equality of tumor development rates between the two groups, the bootstrap procedure given in Section 3 with $K = 10000$ was used and yielded a p-value of 0.0673 for H_0. The results indicate that there is a moderate difference between the lung tumor incidence rates of the mice under the two treatments and that the mice in germfree environment had a little higher rate of tumor development. These are similar to those given by Hoel and Walburg [9], Lagakos and Louis [11] and Sun [17]. Assuming the nonlethality, Hoel and Walburg [9] and Sun [17] gave p-values of 0.01 and 0.028, respectively. Note that the difference between the methods used in Hoel and Walburg [9] and Sun [17] is that the latter adjusts for survival difference between the two groups, while the former does not. In contrast, the method given here adjusts for both the survival difference and lethality. Lagakos and Louis [11] assumed known lethalities and obtained p-values of around 0.07.

We also considered the tests of the hypotheses $\tau = 0$ and $\beta = 0$. Using the same bootstrap samples as those for $X(\hat{\alpha})$, we obtained a p-value of almost zero for testing $\tau = 0$, indicating that the two groups had significantly

different survival rates. The same was also pointed out by Lagakos and Louis [11] and Sun [17]. In particular, the latter obtained and plotted the separate Kaplan-Meier estimates of the survival functions for death times of the mice in the two treatment groups, which suggested that the mice in the germfree environment had significantly longer survival time than those in the conventional environment. For $\beta = 0$, the bootstrap procedure yielded a p-value of 0.3722. This is consistent with the fact that lung tumors are usually regarded as relatively nonlethal (Hoel and Walburg [9]).

5 Concluding Remarks

A nonparametric test procedure is proposed for the comparison of tumor incidence rates, which is often one of the main objectives of tumorigenicity experiments. Compared to the existing methods, a main feature of the presented approach is that it allows tumors having unknown intermediate lethality. Also it adjusts for the possible effect of doses or treatments on animal survival rates. To implement the method, a bootstrap procedure is developed for the determination of p-values.

Note that in the above, Z_i was assumed to be treatment indicators for simplicity. In fact, the presented method also applies to the situation where the vector Z_i includes some covariates whose effect one may want to examine. The focus of this paper is on the comparison of tumor development rates. Sometimes it may be of interest to describe and estimate tumor development rates. For these situations, certain models usually need to be assumed (Dewanji *et al.* [4]; Lindsey and Ryan [12]; Rai, Matthews and Krewski [15]).

More work remains to be done. One problem for future research is to study asymptotic properties of the proposed test procedure and the point estimates of τ and β. In this paper, we have assumed that death time of an animal can be described by the proportional hazards model (1). It would be useful to develop similar test procedures for other commonly used models such as additive failure rate model and multiplicative failure rate model studied by Rai, Matthews and Krewski [15] among others. A related problem is model checking or selection, for which there seems no established method.

Acknowledgments

The authors wish to thank two referees for their comments and suggestions. The research of J. Sun was supported by a grant from the National Institutes of Health and that of S.N. Rai was supported in part by a Cancer Center Support grant (CA 21765) and the American Lebanese Syrian Associated Charities.

References

1. S.D. Charles and L.J. Wei, *Biometrics* **44**, 1005 (1988).
2. D.R. Cox, *Journal of the Royal Statistical Society, Ser. B* **34**, 187 (1972).
3. A. Dewanji and J.D. Kalbfleisch, *Biometrics* **42**, 325 (1986).
4. A. Dewanji, D. Krewski and M.J. Goddard, *Biometrics* **49**, 367 (1993).
5. G.E. Dinse, Biometrics **42**, 325 (1991).
6. G.E. Dinse, *Statistics in Medicine* **13**, 689 (1994).
7. G.E. Dinse and S.W. Lagakos, *Applied Statistics* **32**, 236 (1983).
8. B. Efron, *Journal of the American Statistical Association* **76**, 312 (1981).
9. D.G. Hoel and H.E. Walburg, *Journal of the National Cancer Institute* **49**, 361 (1972).
10. J.D. Kalbfleisch and R.L. Prentice, *The Statistical Analysis of Failure Time Data* (John Wiley and Sons, New York, 1980).
11. S.W. Lagakos and T.A. Louis, *Applied Statistics* **37**, 169 (1988).
12. L. Lindsey and L.M. Ryan *Applied Statistics* **42**, 283 (1993).
13. R. Peto and J. Peto, *Journal of the Royal Statistical Society* **A135**, 185 (1972).
14. S.N. Rai and D.E. Matthews, *Applied Statistics* **46**, 93 (1997).
15. S.N. Rai, D.E. Matthews and D.R. Krewski, *The Canadian Journal of Statistics* **28**, 65 (2000).
16. J. Sun, *Interval Censoring*, in *Encyclopedia of Biostatistics* (John Wiley and Sons, New York, 1998).
17. J. Sun, *Journal of the Royal Statistical Society* **B61**, 243 (1999).
18. J. Sun and J. D. Kalbfleisch, *Biometrics* **52**, 726 (1996).
19. B.W. Turnbull, *Journal of the Royal Statistical Society* **B38**, 290 (1976).

GENERALIZED SMOOTHED ESTIMATING FUNCTIONS WITH CENSORED OBSERVATIONS

A. THAVANESWARAN
Department of Statistics, University of Manitoba, Winnipeg , Manitoba R3T 2N2 CANADA
Email: thavane@ccu.umanitoba.ca

JAGBIR SINGH
Department of Statistics, Fox School of Business, Temple University, Philadelphia, PA1922 USA
Email: jagbir@sbm.temple.edu

Recently there has been a growing interest in the estimation of intensity of continuous time analog of regression models (semimartingales). Counting process model, a special case of the semimatingale models, is widely used to study the estimate of the hazard function in Aalen [1], Ramlau-Hansen [8] *etc.*. Following a recent work of Novak [5] on generalized kernel density estimators, a new class of nonparametric estimators for the intensity of a semimartingale process is introduced. It includes the kernel smoother for the hazard rate based on optimal estimating function studied in Thavaneswaran and Singh [12] as a special case. The asymptotic properties of the smoothed estimator with censored observations are also discussed in some detail.

1 Introduction

A semimartingale is a stochastic process which can be represented as the sum of a process of bounded variation and a local martingale. In the case of continuous time, a typical example of semimartingale is a process $(X(t),\ t \geq 0)$ with independent increments for which $E|X(t)|$ is finite and a function of locally bounded variation. The class of semimartingales includes point processes, Ito processes, diffusion processes, etc. Consider a continuous time stochastic process $(X(t),\ t \geq 0)$ defined on $(\Omega, \mathcal{A}, P)$, a complete probability space for each P in a family $\{\mathcal{P}\}$ of probability measures, and a family $F = [F_t,\ t \geq 0]$, of σ algebras $F_s \subseteq F_t \subseteq \mathcal{A}$ for $s \leq t$, F_0 augmented by sets of measure zero of $\mathcal{A}$, and $F_t = F_{t+}$, where $F_{t+} = \cap_{s>t} F_s$. We denote by D the space of right-continuous functions $(x(t),\ t \geq 0)$ having limit on the left and use $X = (X(t), F_t)$ to denote an F_t -adapted random process $(X(t))$ with trajectories in the space D. For simplicity we assume that $X(0) = 0$. We shall denote by $M^2_{loc}(F, P)$ a class of locally square integrable martingales $(H(t), F_t)$. Assume that the process $(X(t), F_t)$ is a semimartingale for each P,

that is for each P it can be represented in the form

$$X(t) = V(t) + H(t) \tag{1}$$

where $V(t)$ is a locally bounded variation process and $H(t) \in M^2_{loc}(F, P)$. When we allow $V(t)$ and $H(t)$ to depend on $P \in \{\mathcal{P}\}$ only through θ, the model (refeq1.1) can be written as

$$X(t, \theta) = V(t, \theta) + H(t, \theta)$$

When $\theta \in \mathcal{R}$, optimal as well as recursive estimates have been studied in Thavaneswaran and Thompson [11]. Asymptotic properties of the parametric estimators for vector valued multiparameter semimartingales had been studied in Hutton and Nelson [3].Here we consider the following semimartingale model of the form

$$dX(t) = \alpha(t)Y(t-)dR(t) + dM(t). \tag{2}$$

where $\alpha(t)$ is an unobservable deterministic part of the intensity process of the semimartingale $X(t)$, $\{X(s), Y(s), R(s), 0 \leq s \leq t\}$ are observable processes, $M(t) \in M^2_{loc}(F, P)$ with predictable variance process $< M >_t = \int_0^t C(s)dR(s)$, and $C(s)$is a known function of the observations and $\alpha(s)$. For a similar restriction that the conditional mean and the conditional variance of $X(t)$ are absolutely continuous with respect to $R(t)$, see Hutton and Nelson [3]. When $R(t) = t$, equation (2) turns out to be the Aalen's [1] multiplicative intensity model for counting processes.

Example 1.1. *Poisson process:* When $X(t)$ is a right continuous process having jumps of size 1 and $\lambda(t) = \int_o^t \alpha(s)Y(s)ds$, with $Y(s) = 1$, is a deterministic function, the semimartingale model (2) becomes a non homogeneous Poisson process model with cumulative intensity $\lambda(t)$.

Example 1.2. *Multiplicative intensity model:* When $X(t)$ denotes the number of deaths up to time t, $\alpha(t)$ is a hazard rate, $Y(t-)$ is the number of individuals at risk before time t, then the semimartingale model (2) turns out to be the multiplicative intensity model introduced by Aalen [1]. The process $M(t)$ is a zero mean square integrable with variance process$< M >_t = \int_0^t \alpha(s)Y(s)ds$ provided there are no simultaneous deaths. This model has been widely applied to such phenomena as life history data or arrivals at an intensive care unit of a hospital. The counting process model and the martingale limit theory have been used to study the asymptotic properties of the estimators of the hazard function and the survival function with censored observations.

In a recent paper, Novak [5] discussed the problem of obtaining a generalized kernel density estimator with independent observations. The kernel function smoothing approach is a useful tool for hazard function estimation as well.

Recently, efforts have been made to extend the smoothing methodology in various directions, see Ramlau-Hansen [8], Thavaneswaran [10], Thavaneswaran and Singh [12]. The purpose of this paper is to explore the implications of the result of Novak [5] for estimating the intensity function of a semimartingale model. We also apply it to hazard function estimation with censored observations by generalizing the Novak's [5] result.

Let $X_1, ..., X_n$ be a random sample from the distribution of a random variable (r.v.) X with density $f(x) > 0$. The classical kernel density estimator of $f(x)$, at any given x, proposed by Rosenblatt [7], and studied by Parzen [6] and Nadaraya [4] is:

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} f_\gamma \left(\frac{(X_i - x}{h} \right), \qquad n \in \mathbb{N}$$

The kernel f_γ, the density of some symmetric r.v. γ, and the smoothing parameter (band width) $h = h(n)$ are to be chosen by the statistician. It is assumed that the density f is continuous at x, f_γ vanishes outside the interval $[-T, T]$ and has at most a finite number of discontinuity points. The generalized kernel estimator studied in Novak [5] is:

$$f_{n,\alpha}(x) = \frac{1}{n} \sum_{i=1}^{n} f_{h\gamma}\left((X_i - x) f^\alpha(X_i)\right) f^\alpha(X_i) I_i \tag{3}$$

where $\alpha \in \mathcal{R}$, $I_i = I\{|x - X_i| f^\alpha(x) < hT_+\}$, and T_+ is a constant greater than T.

For any continuous function g of f we propose the following extended version of the Novak's estimator as

$$f_{n,g}(x) = \frac{1}{n} \sum_{i=1}^{n} f_{h\gamma}\left((X_i - x) g(f(X_i))\right) g(f(X_i)) I_i \tag{4}$$

where $I_i = I\{|x - X_i| g(f(x)) < hT_+\}$.

If $g(x) = f^\alpha(x)$, then estimator (1.6) reduces to the estimator (3) and for $\alpha = \frac{1}{2}$ it coincides with Abramson [2] estimator. Thus, the class of proposed estimators $\{f_{n,g}(x)\}$ generalizes that of kernel density estimators including Novak's as well. In order to compute the estimator, Novak suggested that one can substitute a consistent estimator of $f(X_i)$. In the case of censored data it is more appropriate to substitute an estimate which incorporates censoring. Thus, it is of interest to note that our approach allows to smooth the density with censored data by letting $g(f(x))$ as the survival function and substituting the product limit estimator in the kernel. The following lemma will be used in Section 2.

Lemma 1.1. As $n \to \infty$

$$Ef_{n,g}(x) \to f(x)$$

$$nhDf_{n,g}(x) \to \nu f(x) g(x) \tag{5}$$

where D denotes the variance and $\nu = \int f_\gamma^2(x)\, dx$.

The proof of the lemma follows by the dominated convergence theorem.

In the next section, a modified version of this generalized kernel density estimator is studied for a semimartingale intensity.

2 Estimating function based kernel estimate

Based on a single realization, a smoothed optimal estimating equation studied in Thavaneswaran and Singh [12] for estimating the semimartingale intensity $\alpha(t)$ at t_0 is:

$$\int_o^1 f_\gamma\left((t_0 - s)h^{-1}\right) h^{-1} dG_s^0 = 0 \tag{6}$$

where

(i) $G_t^0 = \int_0^t a_{s,\theta}^0 dM_{s,\theta}$, as in Thavaneswaran and Thompson [11], is an optimal estimating function defined through $a_{s,\theta}^0 = (\frac{Y(s)J(s)}{C(s)})$, where $J(s) = I(Y(s) > 0, C(s) > 0)$, and

(ii) f_γ is a non-negative integrable kernel function with band width h .

In analogy with (6), we propose a smoothed estimate for $\alpha(t_0)$ as a solution of

$$\int_0^1 f_{h\gamma}((t_0 - s)g(\alpha(X(s))))g(\alpha(X(s)))dG_s^0 = 0, \tag{7}$$

where g is a continuous function of $\alpha(t)$ through the observations as in (4). When $g(x) = 1$, (7) reduces to (6). The explicit form of the resulting estimator from the estimating equation (2.2) can be written as:

$$\theta^0(t) = \frac{\int_0^1 f_{h\gamma}\left((t_0 - s)g(X(s))\right) g(X(s)) \frac{Y(s)J(s)}{C(s)} dX(s)}{\int_0^1 f_{h\gamma}\left((t_0 - s)g(X(s))\right) g(X(s)) \frac{Y^2(s)J(s)}{C(s)} dR(s)}.$$

2.1 Asymptotics: Strong consistency and asymptotic normality

The optimal smoother for $\alpha(t_0) = \theta_0$ for fixed t_0 from a sequence of semi-martingales indexed by n, (i.e.)

$$dX_n(t) = \alpha(t)Y_n(t-)dR(t) + dM_n(t),$$

for $0 \leq t \leq 1$, can be written as

$$\begin{aligned}\theta_n^0(t) &= \frac{\int_0^1 f_{h\gamma}((t_0 - s)g(X_n(s)))g(X_n(s))\frac{Y_n(s)J_n(s)}{C_n(s)}dX_n(s)}{\int_0^1 f_{h\gamma}((t_0 - s)g(X_n(s)))g(X_n(s))\frac{Y_n^2(s)J_n(s)}{C_n(s)}dR(s)} \\ &= \theta_0 + \frac{\int_0^1 f_{h\gamma}\left((t_0 - s)g(X_n(s))\right)g(X_n(s))\frac{Y_n(s)J_n(s)}{C_n(s)}dM_n(s)}{\int_0^1 f_{h\gamma}\left((t_0 - s)g(X_n(s))\right)g(X_n(s))\frac{Y_n^2(s)J_n(s)}{C_n(s)}dR(s)} \\ &= \theta_0 + \frac{\bar{M}_n}{A_n}\end{aligned}$$

where

$$\bar{M}_n = \int_0^1 f_{h\gamma}\left((t_0 - s)g(X_n(s))\right)g(X_n(s))\frac{Y_n(s)J_n(s)}{C_n(s)}dM_n(s)$$

and

$$A_n = \int_0^1 f_{h\gamma}\left((t_0 - s)g(X_n(s))\right)g(X_n(s))\frac{Y_n^2(s)J_n(s)}{C_n(s)}dR(s)$$

For any sample size n, it can be easily shown that the estimate is unbiased for models with deterministic intensity. The following two theorems, analogous to the ones given in Thavaneswaran and Singh [11] are established.

Theorem 2.1. Let $m_n = \sum_{i=1}^n (\Delta\bar{M}_i/A_i)$. Under the assumptions that

(i) the predictable variance process of m_n , $< m_n >_\infty < \infty$ a.s. and

(ii) for the predictable process A_n, $A_\infty = \infty$ a.s. .

The optimal smoother $\theta_n^0(t) \to \alpha(t)$ a.s. for all fixed t as $n \to \infty$

The assumptions (i) and (ii) are somewhat restrictive for a general semi-martingale model. However these can be easily verified for an autoregressive model of order one as in Shiryayev [9](p. 489).

Proof: The process

$$H_n(s) = f_{h\gamma}\left((t_0 - s)g(X_n(s))\right)g(X_n(s))\frac{Y_n(s)J_n(s)}{C_n(s)}$$

is predictable and $M_n(s)$ is a zero mean square integrable martingale. Hence it follows from the properties of stochastic integrals, the integral

$$\bar{M}_n = \int_0^1 H_n(s)dM_n(s)$$

is a zero mean square integrable martingale. Furthermore,

$$A_n = \int_0^1 f_{h\gamma}((t_0 - s)g(X_n(s)))g(X_n(s))\frac{Y_n^2(s)J_n(s)}{C_n(s)}dR(s)$$

is a Lebesgue-Stieltjes integral of a predictable process with respect to a bounded variation function $R(s)$ and is predictable. Therefore, $\bar{M}_n/A_n$ is a martingale sequence. The proof now follows by applying the strong law of large numbers for the martingale sequence $\bar{M}_n/A_n$, as in Shiryayev [9](p. 487).

Theorem 2.2. Assume that

(i) $\frac{g(X_n(s)Y_n(s)J_n(s)}{C_n(s)} \to \frac{1}{\sigma(s)}$, as $n \to \infty$ in probability,

(ii) the functions α, g and σ are continuous at the point t,

(iii) $R(t) = t$, $\lim \frac{g(X_n(s))Y_n^2(s)J_n(s)}{nC_n(s)} = \gamma_s$ in probability uniformly in a neighborhood of t.

Then,

$$\sqrt{nh_n}(\theta_n^0(t) - \theta_0) \xrightarrow{D} N\left(0, \frac{\nu}{\gamma_t}\right) \text{ as } n \to \infty,$$

where $\nu = \int_0^1 f_\gamma^2(t)dt$ as in (5).

Proof: We apply the martingale central limit theorem given in Shiryayev [9](p. 511). In order to apply the theorem we verify the necessary conditions. Recall that

$$\sqrt{nh_n}(\hat{\theta}_n^0(t) - \theta_0) = \frac{\frac{\int_0^1 H_n(s)dM_n(s)}{\sqrt{nh_n}}}{\frac{A_n}{nh_n}} = \frac{\hat{M}_n}{B_n}$$

where $H_n(s)$, A_n are as defined earlier and$\{\hat{M}_n\}$ is a sequence of martingales with variance process,

$$< \hat{M}_n >= \frac{1}{nh_n}\int_0^1 H_n^2(s)d < M >_n (s) = \frac{1}{nh_n}\int_0^1 H_n^2(s)C_n(s)ds$$

$$= \int_0^1 \frac{f_{h\gamma}^2\left((t_0 - s)g(X_n(s))\right) g^2(X_n(s))Y_n^2(s)J_n(s)}{C_n^2(s)}\frac{C_n(s)}{nh_n}ds$$

$$= \int_0^1 \frac{f_\gamma^2((u)g(X_n(t-h_nu)))\, g^2(X_n(t-h_nu)Y_n^2(t-h_nu)J_n(t-h_nu)}{C_n^2(t-h_nu)} \frac{C_n(t-h_nu)}{n} du$$

$$\to \nu\ \gamma_t.$$

Thus the corresponding variance process converges in probability and satisfies one of the necessary conditions for the martingale central limit theorem.

In order to verify the second condition, consider

$$[|H_n(s)| > \epsilon\sqrt{nh_n}] = \left[|f_{h\gamma}((t-s)g(X_n(s)))\, g(X_n(s))\frac{Y_n(s)J_n(s)}{C_n(s)}| > \epsilon\sqrt{nh_n}\right],$$

as $n \to \infty$, $h_n \to 0$, $\frac{g(X_n(s))Y_n(s)J_n(s)}{C_n(s)} \to \frac{1}{\sigma(s)}$, uniformly in a neighborhood of t. Thus

$$[I|H_n(s)| > \epsilon\ ] \to 0$$

in probability.

Applying the martingale central limit theorem, as $n \to \infty$, $\hat{M}_n(t) \to N(0, \nu\ \gamma_t)$ in distribution. Moreover, $B_n \to \gamma_t$ in probability. Hence,

$$\sqrt{nh_n}(\theta_n^0(t) - \theta_0) \to N\left(0, \frac{\nu}{\gamma_t}\right)$$

in distribution.

3 Conclusion

In this paper we have introduced generalized kernel estimators for the intensity of a semimartingale process. It extends the results of Thavaneswaran and Singh [12] on strong consistency and asymptotic normality. Generalized kernel estimator of the hazard function with censored observations turns out to be a special case.

Acknowledgments

Research of both authors supported in part by Stan Alton and R.W. Johnson Pharmaceutical Research Institute. First author's research supported in part by the National Science and Engineering Council of Canada.

References

1. O.O. Aalen, *Ann. Statist.* **6**, 701 (1978).
2. I.S. Abramson, *Ann. Math.Statist.* **10**, 1217 (1982).
3. J.E. Hutton and P.I. Nelson, *Stochastic Process. Appl.* **22**, 245 (1986).
4. E.A. Nadaraya, *Theory Probab. Appl.* **19**, 133 (1974).
5. S.Y. Novak, *Theory of Probability and Applications* **90**, 301 (2000).
6. E. Par Zen, *Ann. Math. Statist.* **33**, 1065 (1962).
7. M. Rosenblatt, *Ann. Math.Statist.* **27**, 832 (1956).
8. H. Ramlau-Hansen, *Ann. Statist.* **11**, 453 (1983).
9. A.N. Shiryayev, *Probability* (Springer, New York, 1984).
10. A. Thavaneswaran, *Stochastic Process. Appl.* **28**, 81 (1988).
11. A. Thavaneswaran and M.E. Thompson, *Journal of Applied Probability* **23**, 409 (1986).
12. A. Thavaneswaran and J. Singh, *Ann. Inst. Statist. Math.* **45**, 721 (1993).

LARGE SAMPLE ASYMPTOTIC PROPERTIES OF ORDINARY LEAST SQUARES, TWO STAGE LEAST SQUARES AND LIMITED INFORMATION MAXIMUM LIKELIHOOD ESTIMATORS IN SIMULTANEOUS EQUATIONS MODELS

R. TIWARI

Department of Statistics, University of Jammu, Jammu-180006, India
E-mail: rjtiwari@yahoo.com

V.K. SRIVASTAVA

Department of Statistics, Lucknow University, Lucknow-226007, India

This article discusses the large sample asymptotic properties of three estimators, *viz.*, OLS, TSLS and LIML, for the coefficients of a single structural equation.

1 Introduction

For the coefficients in a single structural equation of a complete simultaneous equation model, Nagar [5] has considered k- class estimators and has derived large sample asymptotic approximations for their bias vector and mean squared error matrix. He has restricted his attention to those k - class estimators which are consistent and have same asymptotic distribution according to large sample asymptotic theory. These estimators are specified by the characterizing scalar k such that k is nonstochastic and $(1-k)$ is of order $O(T^{-1})$, T being the number of observations. Such estimators do include the case of two stage least squares (TSLS) estimator but fail to include two popular estimators. One is the ordinary least squares estimator for which the value of k is nonstochastic but $(1-k)$ does not satisfy the requirement of being of order $O(T^{-1})$. The other is limited information maximum likelihood (LIML) estimator for which the value of $(1-k)$ is of order $O_p(T^{-1})$ but the value of k is stochastic. The purpose of this paper is to provide large sample asymptotic approximations for the bias vector and mean squared error matrix for these two estimators and to make a comparative study of the performance properties of the three popular estimators (*viz.*, ordinary least squares, two stage least squares and limited information maximum likelihood) employing the large sample asymptotic theory.

The plan of this paper is as follows. Section 2 describes the simultaneous equations model and presents the estimators for the coefficients in a structural equation. Section 3 discusses the asymptotic properties of the ordinary least

squares estimator while Section 4 provides the large sample asymptotic properties of the limited information maximum likelihood estimator and compares them with those of the two stage least squares estimators. Lastly, Section 5 gives the derivation of the results.

2 Model Specification And Estimators

Consider a simultaneous equation model consisting of a set of M linear structural equations in M jointly dependent and Λ predetermined variables.

$$YB + X\Gamma = U \tag{1}$$

Where Y is a $T \times M$ matrix of T observations on M jointly dependent variables, B is a $M \times M$ nonsingular matrix of coefficients associated with them, X is a $T \times \Lambda$ matrix of T observations on Λ predetermined variables, Γ is a $\Lambda \times M$ matrix of coefficients associated with them and U is a $T \times M$ matrix of structural disturbances. It is assumed that the row vectors of U are independently and identically distributed, each following a multivariate distribution with mean vector 0 and variance covariance matrix Σ.

With no loss of generality, let us suppose that we are interested in the estimation of parameters in the first structural equation of the model (1), expressible as

$$\begin{aligned} Y &= Y_1\beta + X_1\gamma + u \\ &= A\delta + u; \quad A = (Y_1 \vdots X_1) \quad ; \quad \delta = \begin{pmatrix} \beta \\ \cdots \\ \gamma \end{pmatrix} \end{aligned} \tag{2}$$

where y is a $T \times 1$ vector of T observations on the jointly dependent variable to be explained, Y_1 is a $T \times m$ submatrix of Y, β is a $m \times 1$ vector of coefficients associated with $m(< M)$ explanatory jointly dependent variables, X_1 is a $T \times l$ submatrix of X, γ is a $l \times 1$ vector of coefficients associated with $l(\leq \Lambda)$ explanatory predetermined variables and u is the first column vector of U.

The reduced form corresponding to explanatory jointly dependent variables in (2) is expressible as

$$Y_1 = X\Pi_1 + UH_1 \tag{3}$$

where Π_1 is a $\Lambda \times m$ matrix of reduced form coefficients determined by the elements of the matrix $-\Gamma B^{-1}$ and H_1 is a $M \times m$ matrix obtained from the elements of the inverse matrix B.

We can write

$$\begin{aligned} A &= (Y_1 \vdots X_1) \\ &= \left(X\Pi_1 \vdots X_1\right) + \left(UH_1 \vdots 0\right) \\ &= X\Pi + UH, \end{aligned} \tag{4}$$

where

$$\Pi = \left[\Pi_1 \vdots \begin{pmatrix} I \\ \dots \\ 0 \end{pmatrix}\right] \quad \text{and} \quad H = \left(H_1 \vdots 0\right).$$

Assuming the structural equation (2) to be identifiable, the k-class estimator of δ in (2) is defined by

$$\hat{\delta}_k = \left[A'(I - k\bar{P}_X)A\right]^{-1} A'(I - k\bar{P}_X)Y \tag{5}$$

with $\bar{P}_X = I - X(X'X)^{-1}X'$ and k as the characterizing scalar.
The family (5) encompasses three popular estimators, *viz.*, ordinary least squares estimators(OLS), two stage least squares (TSLS) and limited information maximum likelihood (LIML) estimators. The OLS and TSLS estimators are specified by $k = 0$ and $k = 1$ respectively while LIML estimator is specified by $k = \lambda$ where

$$\lambda = \min_{\beta} \frac{(y - Y_1\beta)'\bar{P}_{X_1}(y - Y_1\beta)}{(y - Y_1\beta)'\bar{P}_X(y - Y_1\beta)} \tag{6}$$

with $\bar{P}_{X_1} = I - X_1(X_1'X_1)^{-1}X'_1$; see, *e.g.*, Kadane [3].

3 Asymptotic Properties of OLS Estimator

In order to examine the properties of estimators of structural coefficients in (2), we assume that the structural disturbances are normally distributed and the model does not contain any lagged endogenous variable. Further, it is assumed that all the exogenous variables in the model are asymptotically cooperative in the sense that $T^{-1}(X'X)$ tends to a finite nonsingular matrix as T tends to infinity. This rules out, for instance the presence of any trend variable in the model; see Krämer [4].

In order to present the large sample asymptotic results by following the approach of Nagar [5], let us introduce the following quantities:

$$d = H'\sigma_{(1)} \tag{7}$$

$$Q = \left(\frac{1}{T}\Pi' X' X \Pi\right)^{-1} \tag{8}$$

where $\sigma_{(1)}$ denotes the first column vector of the variance-covariance matrix Σ of disturbances.

Result I: The large sample asymptotic approximations, upto order $O(T^{-1})$, for the bias vector and variance covariance matrix of the OLS estimators $\hat{\delta}_0$ of δ are given by

$$\begin{aligned} B(\hat{\delta}_0) &= E(\hat{\delta}_0 - \delta) \\ &= Sd + \frac{1}{T}\left[\{(tr\ Q^{-1}S) - 1\}SQ^{-1}Sd + SQ^{-1}SQ^{-1}Sd\right] \end{aligned} \tag{9}$$

$$\begin{aligned} V(\hat{\delta}_0) &= E\left[\hat{\delta}_0 - E(\hat{\delta}_0)\right]\left[\hat{\delta}_0 - E(\hat{\delta}_0)\right]' \\ &= \frac{1}{T}\left[(\sigma_{11} - d'Sd)S - (d'SQ^{-1}Sd)SQ^{-1}S - SQ^{-1}Sdd'SQ^{-1}S\right] \end{aligned} \tag{10}$$

where

$$\begin{aligned} S &= \left(\frac{1}{T'}\Pi' X' X \Pi + H'\Sigma H\right)^{-1} \\ &= \left(Q^{-1} + H'\Sigma H\right)^{-1} \end{aligned} \tag{11}$$

and σ_{11} is the (1,1)th element of Σ.

From(9), it is observed that the OLS estimator is generally inconsistent and biased unlike the TSLS and LIML estimators which are consistent though generally biased.

If we examine the asymptotic variance covariance matrix of the three estimators, it is well known that the TSLS and LIML estimators have the same asymptotic variance covariance which is given by $T^{-1}\sigma_{11}Q$. Comparing it with (10) , we observe that

$$\begin{aligned} V(\hat{\delta}_1) - V(\hat{\delta}_0) &= V(\hat{\delta}_\lambda) - V(\hat{\delta}_0) \\ &= \frac{\sigma_{11}}{T}(Q - S) + \frac{1}{T}\left[(d'Sd)S\right. \\ &\qquad \left. + (d'SQ^{-1}Sd)SQ^{-1}S + SQ^{-1}Sdd'SQ^{-1}S\right] \end{aligned} \tag{12}$$

As the matrix

$$(Q - S) = \left(\frac{1}{T}\Pi' X' X \Pi\right)^{-1} - \left(\frac{1}{T}\Pi' X' X \Pi + H'\Sigma H\right)^{-1} \tag{13}$$

is at least positive semidefinite, it follows that the matrix expression (12) is also at least positive semidefinite. This implies that OLS estimator is generally asymptotically more efficient than TSLS and LIML estimators.
It may be remarked that if we employ small disturbance asymptotic theory, all the three estimators are consistent and share the same asymptotic distribution; see Kadane [3] and Srivastava and Giles [6].

4 Comparison of TSLS and LIML Estimators

Both the TSLS and LIML estimators are known to be consistent and to possess the same asymptotic distribution. In order to study the differences in their performance properties in finite samples, we need to examine higher order asymptotic approximations.
Let us first consider the case of TSLS estimator. The large sample asymptotic approximations for its bias vector to order $O(T^{-1})$ and its mean squared error matrix to order $O(T^{-2})$ have been obtained by Nagar [5]. His results are as follows:

$$\begin{aligned} B(\hat{\delta}_1) &= E(\hat{\delta}_1 - \delta) \\ &= \frac{1}{T}(L-1)Qd \end{aligned} \tag{14}$$

$$\begin{aligned} M(\hat{\delta}_1) &= E(\hat{\delta}_1 - \delta)(\hat{\delta}_1 - \delta)' \\ &= \frac{\sigma_{11}}{T}Q + \frac{1}{T^2}\Big[\{\sigma_{11}(trQH'\Sigma H) - 2(L-1)d'Qd\}Q \\ &\qquad + (L^2 - 3L + 4)Qdd'Q - (L-2)\sigma_{11}QH'\Sigma HQ\Big] \end{aligned} \tag{15}$$

where

$$L = (\Lambda - \ell - m) \tag{16}$$

denotes the degree of over-identification of the structural equation (2)

From (14) and (15) , we can obtain the variance covariance matrix upto order $O(T^{-1})$ as follows:

$$\begin{aligned} V(\hat{\delta}_1) &= E\Big[\hat{\delta}_1 - E(\hat{\delta}_1)\Big]\Big[\hat{\delta}_1 - E(\hat{\delta}_1)\Big]' \\ &= \frac{\sigma_{11}}{T}Q + \frac{1}{T^2}\Big[\{\sigma_{11}(trQH'\Sigma H) - 2(L-1)d'Qd\}Q \\ &\qquad - (L-3)Qdd'Q - (L-2)\sigma_{11}QH'\Sigma HQ\Big] \end{aligned} \tag{17}$$

Result II: For the LIML estimator $\hat{\delta}_\lambda$ of δ, the large sample asymptotic approximations for the bias vector to order $O(T^{-1})$ and the mean squared error matrix to order $O(T^{-2})$ are given by

$$B(\hat{\delta}_\lambda) = -\frac{1}{T}Qd, \tag{18}$$

and

$$M(\hat{\delta}_\lambda) = \frac{\sigma_{11}}{T}Q + \frac{1}{T^2}\Big[\{\sigma_{11}(tr\ QH'\Sigma H) + 2d'Qd\}Q - (L-4)Qdd'Q + (L+2)\sigma_{11}QH'\Sigma HQ\Big]. \tag{19}$$

From (18) and (19), the variance covariance matrix to order $O(T^{-2})$ is given by

$$V(\hat{\delta}_\lambda) = \frac{\sigma_{11}}{T}Q + \frac{1}{T^2}\Big[\{\sigma_{11}(tr\ QH'\Sigma H) + 2d'Qd\}Q - (L-3)Qdd'Q + (L+2)\sigma_{11}QH'\Sigma HQ\Big]. \tag{20}$$

It may be noticed that these results tally with the results obtained by Fuller [2] for a special case; see Lemma 1 and equation (12) of Fuller [2].

Comparing (14) and (18), we conclude that both the TSLS and LIML estimators are generally biased though consistent. As expected, both have identical biases to order $O(T^{-1})$ for exactly identified equation (2). If the equation is overidentified and the degree of overidentification exceeds one, the bias of TSLS estimator has a sign opposite to that of LIML estimator. Further, the magnitude of bias of TSLS estimator is larger in comparison to that of LIML estimator.

Next, we observe from (17) and (20) that

$$V(\hat{\delta}_\lambda) - V(\hat{\delta}_1) = \frac{2L}{T^2}\Big[(d'Qd)Q + 2\sigma_{11}QH'\Sigma HQ\Big] \tag{21}$$

which is always a positive definite matrix for $L > 0$. This implies that the TSLS estimator dominates the LIML estimator according to the criterion of asymptotic variance covariance matrix to order $O(T^{-2})$ provided that the structural equation is overidentified.

Similarly, if we compare the expressions (15) and (19) we get

$$M(\hat{\delta}_\lambda) - M(\hat{\delta}_1) = \frac{L}{T^2}\Big[2(d'Qd)Q - (L-2)Qdd'Q + 2\sigma_{11}QH'\Sigma HQ\Big] \tag{22}$$

Now we notice that the matrix $\Big[2(d'Qd)Q - (L-2)Qdd'Q\Big]$ is positive definite if and only if $L < 4$; see, *e.g.* Dube *et al.* [1](Lemma 1) . It thus follows

from (22) that the TSLS estimator dominates the LIML estimator according to the asymptotic mean squared error matrix criterion to the order of our approximation at least as long as the structural equation (2) is overidentified and the degree of overidentification is less than four.

It may be mentioned that the expression (18) is identical with the small disturbance asymptotic approximation of bias vector of LIML estimator; see Kadane [3](Theorem 1). Similarly, if we retain the terms upto order $O(T^{-2})$ only in the expression for the small disturbance asymptotic approximation of the mean squared error matrix obtained by Kadane [3](Theorem 2) , we get the same expression as (19). Finally, for the dominance of TSLS estimator over the LIML estimator with respect to the criterion of small disturbance asymptotic approximation for the mean squared error matrix, Kadane [3](Corollary 1, p.728) has found the condition that the degree of overidentification should be less than six provided $(T - \Lambda) > 2$.

5 Derivation of Results

First of all, we present the following expectation results.
Lemma: We have

$$E(UCU) = C'\Sigma \quad , \quad E(U'CU) = (trC)\Sigma$$
$$E(UCU') = (trC\Sigma)I \quad , \quad E(U'CU') = \Sigma C'$$
$$E(U'UCU'U) = T\Big[(T+1)\Sigma C\Sigma + (trC\Sigma)\Sigma\Big] \tag{23}$$

where C is a nonstochastic matrix of suitable order in each case.

For proof, see Kadane [3](Appendix) or Srivastava and Tiwari [7].

Now for the results related to OLS estimator, we observe that

$$\begin{aligned}(\hat{\delta}_0 - \delta) &= (A'A)^{-1}A'u \\ &= \Big[I + \frac{1}{T}S(H'U'X\Pi + \Pi'X'UH) + SH'\Delta H\Big]^{-1} \\ &\qquad \Big[SH'\sigma_{(1)} + S\Big(\frac{1}{T}\Pi'X'u + H'\epsilon\Big)\Big]\end{aligned} \tag{24}$$

where

$$S = \Big(\frac{1}{T}\Pi'X'X\Pi + H'\Sigma H\Big)^{-1}, \tag{25}$$

$$\Delta = \Big(\frac{1}{T}U'U - \Sigma\Big), \tag{26}$$

and

$$\epsilon = \left(\frac{1}{T}U'u - \sigma_{(1)}\right). \tag{27}$$

Observing that the elements of the matrix expression $\frac{1}{T}S(H'U'X\Pi + \Pi'X'UH) + SH'\Delta H$ are of order $O_p(T^{-1/2})$ and expanding the expression in the first square brackets on the right hand side of (24), we get

$$(\hat{\delta}_0 - \delta) = g_0 + g_{-1/2} + g_{-1} + O_p(T^{-3/2}) \tag{28}$$

where

$$g_0 = SH'\sigma_{(1)}, \tag{29}$$

$$\begin{aligned} g_{-1/2} = \frac{1}{T}S\Pi'X'u + SH'\epsilon - \frac{1}{T}S(H'U'X\Pi + \Pi'X'UH)Sd \\ - SH'\Delta HSd, \end{aligned} \tag{30}$$

and

$$g_{-1} = -S\left[\frac{1}{T}(H'U'X\Pi + \Pi'X'UH) + H'\Delta H\right]g_{-1/2}. \tag{31}$$

Here the suffixes of g indicate the order of magnitude in probability.
Thus the bias vector to order $O(T^{-1})$ is given by

$$B(\hat{\delta}_0) = g_0 + E(g_{-1/2} + g_{-1}). \tag{32}$$

It is straightforward to see that

$$E(g_{-1/2}) = 0. \tag{33}$$

Similarly, using the results in Lemma along with normality of disturbances, it is easy to see that

$$E(g_{-1}) = \left[tr(Q^{-1}S) - 1\right]SQ^{-1}Sd + SQ^{-1}SQ^{-1}Sd. \tag{34}$$

Substituting (33) and (34) in (32), we obtain the result (9)
Next, we consider the variance covariance matrix upto order $O(T^{-1})$:

$$\begin{aligned} V(\hat{\delta}_0) &= E\left[\hat{\delta}_0 - E(\hat{\delta}_0)\right]\left[\hat{\delta}_0 - E(\hat{\delta}_0)\right]' \\ &= E(g_{-1/2}g'_{-1/2}) - E(g_{-1/2})E(g'_{-1/2}) \end{aligned} \tag{35}$$

From the results mentioned in the Lemma, it can be shown that

$$\begin{aligned} E(g_{-1/2}g'_{-1/2}) = \frac{1}{T}\Big[(\sigma_{11} - d'Sd)S - (d'SQ^{-1}Sd)SQ^{-1}S \\ - SQ^{-1}Sdd'SQ^{-1}S\Big] \end{aligned} \tag{36}$$

Substituting (33) and (36) in (35), we obtain the result in (10)

For similar results in the context of LIML estimator, we first observe from (6) that λ can be expressed as

$$\begin{aligned}\lambda &= \min_{\beta} \frac{(y - Y_1\beta)' P_{X_1}^{-} (y - Y_1\beta)}{(y - Y_1\beta)' P_X^{-} (y - Y_1\beta)} \\ &= 1 + \min_{\delta} \frac{(y - A\delta)'(P_{X_1}^{-} - P_X^{-})(y - A\delta)}{(y - A\delta)' P_X^{-} (y - A\delta)} \\ &= 1 + \frac{(y - A\hat{\delta}_\lambda)'(P_{X_1}^{-} - P_X^{-})(y - A\hat{\delta}_\lambda)}{(y - A\hat{\delta}_\lambda)' P_X^{-} (y - A\hat{\delta}_\lambda)}.\end{aligned} \tag{37}$$

Now we can write

$$\begin{aligned}(y - A\hat{\delta}_\lambda) &= u - A(\hat{\delta}_\lambda - \delta) \\ &= u - (X\Pi + UH)\left[(\Pi' X' X \Pi)^{-1} \Pi' X' u + O_p(T^{-1})\right] \\ &= \left[I - X\Pi(\Pi' X' X \Pi)^{-1} \Pi' X'\right] u - UH(\Pi' X' X \Pi)^{-1} \Pi' X' u + \cdots\end{aligned} \tag{38}$$

so that

$$\frac{1}{T}(y - A\hat{\delta}_\lambda)'(P_{X_1}^{-} - P_X^{-})(y - A\hat{\delta}_\lambda) = \frac{1}{T} u' R U - \frac{2}{T^2} u' R U H Q \Pi' X' u + O_p(T^{-2}) \tag{39}$$

where $R = X(X'X)^{-1}X' - X\Pi(\Pi' X' X \Pi)^{-1}\Pi' X'$.

Similarly, we have

$$\frac{1}{T}\left[(y - A\hat{\delta}_\lambda)' P_X^{-} (y - A\hat{\delta}_\lambda)\right]^{-1} = \frac{1}{\sigma_{11}}\left[1 - \left(\frac{1}{\sigma_{11} T} u'u - 1\right) + \frac{2}{\sigma_{11} T} d' Q \Pi' X' u\right] + O_p(T^{-1}). \tag{40}$$

Employing (39) and (40) in (37), we find

$$\lambda = 1 + t_{-1} - t_{-3/2} + O_p(T^{-2}) \tag{41}$$

where

$$t_{-1} = \frac{1}{\sigma_{11} T} u' R u, \tag{42}$$

and

$$t_{-3/2} = \frac{2}{\sigma_{11}T^2}u'RUHQ\Pi'X'u + \frac{1}{\sigma_{11}T}u'Ru\Big(\frac{1}{\sigma_{11}T}u'u - 1\Big) - \frac{2}{\sigma_{11}^2T^2}u'Ru.d'Q\Pi'X'u \tag{43}$$

Here the suffixes of t indicate the orders of magnitude in probability. Next, using (4) and (41), we notice that

$$\frac{1}{T}A'(I - \lambda\bar{P_X})u = \frac{1}{T}\Pi'X'u + \Big(\frac{1}{T}H'U'P_Xu - t_{-1}d\Big) + (-t_{-1}H'\epsilon + t_{-3/2}d) + O_p(T^{-2}) \tag{44}$$

where $\epsilon = \Big(\frac{1}{T}U'u - \sigma_{(1)}\Big)$ is of order $O_p(T^{-1/2})$.

Similarly, observing that the elements of $\Big(\frac{1}{T}U'U - \Sigma\Big)$ are of order $O_p(T^{-1/2})$,we have

$$\frac{1}{T}U'(I - \lambda\bar{P_X})U = \frac{1}{T}U'P_XU - t_{-1}\Sigma + O_p(T^{-3/2}) \tag{45}$$

whence we can express

$$\Big[\frac{1}{T}A'(I - \lambda\bar{P_X})A\Big]^{-1} = Q - \frac{1}{T}Q(\Pi'X'UH + H'U'X\Pi)Q - QH'\Big(\frac{1}{T}U'P_XU - t_{-1}\Sigma\Big)HQ + \frac{1}{T^2}Q(\Pi'X'UH + H'UX\Pi) Q(\Pi'X'UH + H'UX\Pi)Q + O_p(T^{-3/2}). \tag{46}$$

Now using (44) and (46), we can express the estimation error of LIML estimator as follows:

$$(\hat{\delta}_\lambda - \delta) = \Big[A'(I - \lambda\bar{P_X})A\Big]^{-1}A'(I - \lambda\bar{P_X})u = e_{-1/2} + (e_{-1} + f_{-1}) + (e_{-3/2} + f_{-3/2}) + O_p(T^{-2}) \tag{47}$$

where

$$e_{-1/2} = \frac{1}{T}Q\Pi'X'u, \tag{48}$$

$$e_{-1} = \frac{1}{T}QH'U'Ru - \frac{1}{T^2}Q\Pi'X'UHQ\Pi'X'u, \tag{49}$$

$$f_{-1} = -t_{-1}Qd, \tag{50}$$

$$e_{-3/2} = \frac{1}{T^3}Q\Pi'X'UHQ\Pi'X'u - \frac{1}{T^2}QH'U'RUHQ\Pi'X'u - \frac{1}{T^2}Q\Pi'X'UHQH'U'Ru - \frac{1}{T^2}QH'U'X\Pi QH'U'Ru, \tag{51}$$

and

$$f_{-3/2} = \frac{1}{T}t_{-1}(QH'\Sigma HQ\Pi'X'u + Q\Pi'X'UHQd + QH'U'X\Pi Qd) - t_{-1}QH'\epsilon + t_{-3/2}Qd \tag{52}$$

It may be observed that the sum $(e_{-1/2} + e_{-1} + e_{-3/2})$ provides the expression for the estimation error of TSLS estimator to order $O(T^{-3/2})$. Thus the bias vector of LIML estimator to order $O(T^{-1})$ is given by

$$B(\hat{\delta}_\lambda) = E(e_{-1/2} + e_{-1}) - E(t_{-1})Qd \tag{53}$$

The first expectation on the right hand side of the above equation is essentially equal to the bias vector of TSLS estimator to order $O(T^{-1})$ and is given by (14).
As $E(t_{-1}) = (L/T)$, we have

$$B(\hat{\delta}_\lambda) = -\frac{1}{T}Qd \tag{54}$$

which is the result (18).

In a similar manner, the mean squared error matrix to order $O(T^{-2})$ is given by

$$M(\hat{\delta}_\lambda) = E(e_{-1/2}e'_{-1/2} + e_{-1}e'_{-1/2} + e_{-1/2}e'_{-1} + e_{-1}e'_{-1} + e_{-3/2}e'_{-1/2} + e_{-1/2}e'_{-3/2}) + E(f_{-1}e'_{-1/2} + e_{-1/2}f'_{-1}) + E(f_{-3/2}e'_{-1/2} + e_{-1/2}f'_{-3/2}) + E(e_{-1}f'_{-1} + f_{-1}e'_{-1} + f_{-1}f'_{-1}). \tag{55}$$

The first expectation on the right hand side of the above equation is the mean squared error matrix of TSLS estimator and is given by (15). For the remaining expectations, it can be verified that

$$E(f_{-1}e'_{-1/2}) = 0, \tag{56}$$

$$E(f_{-3/2}e'_{-1/2}) = \frac{L}{T^2}\Big[(d'Qd)Q + Qdd'Q + \sigma_{11}QH'\Sigma HQ\Big], \tag{57}$$

$$E(e_{-1}f'_{-1}) = \frac{L(L+1)}{T^2}Qdd'Q, \tag{58}$$

and

$$E(f_{-1}f'_{-1}) = \frac{L(L+2)}{T^2}Qdd'Q. \tag{59}$$

Using these alonwith (15) in (55), we obtain the result (19).

References

1. M. Dube, V.K. Srivastava, H. Toutenburg and P. Wijekoon, *Communications in Statistics* **20**, 2009 (1991).
2. W.A. Fuller, *Econometrica* **45**, 939 (1977).
3. J.B. Kadane, *Econometrica* **39**, 723 (1971).
4. W. Krämer, *Communications in Statistics, Theory and Methods* **14**, 1997 (1985).
5. A.L. Nagar, *Econometrica* **27**, 575 (1959).
6. V.K. Srivastava and D.E.A. Giles, *Journal of Quantitative Economics* **7**, 273 (1991).
7. V.K. Srivastava and R. Tiwari, *Scandinavian Journal of Statistics* **3**, 135 (1976).

RELATIVE STABILITY OF WEIGHTED MAXIMA OF BOUNDED I.I.D. RANDOM VARIABLES

R.J. TOMKINS
Department of Mathematics and Statistics, University of Regina
Regina, Sask. S4S 0A2, Canada
E-mail:jim.tomkins@uregina.ca

Let $\{X_n, n \geq 1\}$ be an *i.i.d.* sequence such that $P[X_1 > 0] > 0$ and $P[X_1 \leq x] = 1$ for all large x. For a positive real sequence $\{a_n\}$, define $Z_n = \max\{a_1X_1, \cdots, a_nX_n\}$ for $n \geq 1$. Necessary and sufficient conditions are given for $\{Z_n\}$ to be relatively stable; i.e., for $\frac{Z_n}{c_n} \to 1$ in probability for some real sequence $\{c_n\}$.

Let $X_1, X_2, \cdots$ be an *i.i.d.* sequence with distribution function (d.f.) F such that $F(0) < 1$ and $F(a) = 1$ for some $a > 0$. For a positive real sequence $\{a_n, n \geq 1\}$, define the random sequence

$$Z_n = \max\{a_1X_1, a_2X_2, \cdots, a_nX_n\}, \ n \geq 1. \tag{1}$$

This paper will present necessary and sufficient conditions for $\{Z_n, n \geq 1\}$ to be relatively stable; i.e., $Z_n/c_n \xrightarrow{p} 1$ for some real sequence $\{c_n\}$, where "$\xrightarrow{p}$" denotes convergence in probability, and will show that the norming sequence may be assumed to be $c_n = \max\{a_1, \cdots, a_n\}, n \geq 1$. As will be seen, these results may be viewed as part of the process of generalizing a well-known theorem of Gnedenko [1] to situations involving the maxima of an independent sequence.

Notice that $Z_n \leq Z_{n+1}$ almost surely (a.s.), $n \geq 1$, so that $\lim_{n\to\infty} Z_n$ exists a.s. Tomkins [2] proved that $\lim_{n\to\infty} Z_n = \infty$ a.s. if and only if (iff) $\sup_{n\geq 1} a_n = \infty$. Unless otherwise noted, the focus of this article is on the case where $\sup_{n\geq 1} a_n = \infty$ and X_1 is not degenerate.

Define the constant $x_0 = \inf\{x : F(x) = 1\}$. By hypothesis, $0 < x_0 < \infty$. The following five lemmas will lead to the main result: $\{Z_n\}$ is relatively stable iff $\max\{a_1, \cdots, a_{n+1}\} \sim \max\{a_1, \cdots, a_n\}$, where "$b_{n+1} \sim b_n$" means that $b_{n+1}/b_n \longrightarrow 1$ as $n \longrightarrow \infty$.

Lemma 1. Without loss of generality, it may be assumed that $X_n \geq 0$ a.s., $n \geq 1$, and hence that $Z_n \geq 0$ a.s., $n \geq 1$, if $Z_n/c_n \xrightarrow{p} 1$ and $c_n \to \infty$.

Proof: Define $X_n^+ = \max(X_n, 0), n \geq 1$. Note that $F(0) < 1$ by hypothesis, so $P[X_n > 0$ infinitely often (i.o.) $] = 1$ by the Borel Zero-one Law. But then $P[Z_n < 0$ i.o. $] = 0$. Therefore, $P[Z_n \neq Z_n'$ i.o.$] = 0$, where $Z_n' = \max\{a_1 X_1^+, \cdots, a_n X_n^+\}$. Hence $(Z_n' - Z_n)/c_n \longrightarrow 0$ a.s. for any sequence $\{c_n\}$ such that $c_n \longrightarrow \infty$. □

Lemma 2. Let $\{i_n\}$ be any integer sequence such that $1 \leq i_n \leq n$. Then $\liminf_{n\to\infty} \frac{c_n}{a_{i_n}} \geq x_0$ if $\frac{Z_n}{c_n} \xrightarrow{p} 1$.

Proof: Let $\varepsilon > 0$. Then

$$F\left((1+\varepsilon)c_n/a_{i_n}\right) \geq \prod_{i=1}^{n} F\left((1+\varepsilon)c_n/a_i\right) = P\left[Z_n \leq (1+\varepsilon)c_n\right] \longrightarrow 1,$$

since $Z_n/c_n \xrightarrow{p} 1$. Hence $F\left((1+\varepsilon)c_n/a_{i_n}\right) \longrightarrow 1$ as $n \to \infty$ for every $\varepsilon > 0$. If $\liminf_{n\to\infty} c_n/a_{i_n} < x_0$, then a number $\eta > 0$ exists such that $c_n/a_{i_n} < x_0 - \eta$ for an infinite number of values of n and, therefore,

$$\liminf_{n\to\infty} F((1+\varepsilon)c_n/a_{i_n}) \leq F((1+\varepsilon)(x_0-\eta)) < 1$$

if ε is chosen to be so small that $(1+\varepsilon)(x_0-\eta) < x_0$. This is a contradiction, so $\liminf_{n\to\infty} c_n/a_{i_n} \geq x_0$. □

Lemma 3. If $Z_n/c_n \xrightarrow{p} 1$ then $c_n \to \infty$ and $c_n \sim c_{n-1}$.

Proof: As noted earlier, $Z_n \to \infty$ a.s. since $\sup_{n\geq 1} a_n = \infty$. But $Z_n/c_n \xrightarrow{p} 1$, from which it follows easily that $c_n \to \infty$.

Now, since X_1 is non-degenerate, one may choose $\varepsilon_1 > 0$ so that $F((1-\varepsilon_1)^2 x_0) > 0$. For every $\varepsilon > 0$,

$$P[Z_n \leq (1-\varepsilon)c_n] = P[Z_{n-1} \leq (1-\varepsilon)c_n]F((1-\varepsilon)c_n/a_n). \tag{2}$$

By Lemma 2 (with $i_n = n$), $c_n/a_n > (1-\varepsilon)x_0$ for all large n (say, $n \geq N$). But then, if $\varepsilon \leq \varepsilon_1$ and $n \geq N$, $F((1-\varepsilon)c_n/a_n) \geq F((1-\varepsilon)^2 x_0) > 0$. Since $Z_n/c_n \xrightarrow{p} 1$, it follows from (2) that $P[Z_{n-1} < (1-\varepsilon)c_n] \to \infty$ for every $\varepsilon \leq \varepsilon_1$, from which fact it is clear that $P[Z_{n-1} \leq (1-\varepsilon)c_n] \to 0$ for every $\varepsilon > 0$. But $P[Z_{n-1} \leq (1+\varepsilon)c_n] \geq P[Z_n \leq (1+\varepsilon)c_n] \to 1$, $\varepsilon > 0$, so $Z_{n-1}/c_n \xrightarrow{p} 1$. It is easy to see that $c_n \sim c_{n-1}$, because $Z_{n-1}/c_{n-1} \xrightarrow{p} 1$ by hypothesis. □

Lemma 4. If $Z_n/c_n \xrightarrow{p} 1$ then $c_n \sim a_n^* x_0$ and $a_n^* \sim a_{n-1}^*$, where $a_n^* \equiv \max\{a_1, \cdots, a_n\}$.

Proof: For each $n \geq 1$, choose i_n, $1 \leq i_n \leq n$, such that $a_{i_n} = a_n^*$. By Lemma 2, $\liminf_{n\to\infty} c_n/a_n^* \geq x_0$.

Now, $Z_n \leq a_n^* x_0$ a.s., $n \geq 1$, so, for every $\varepsilon > 0$,

$$P[a_n^* x_0 \leq (1-\varepsilon)c_n] \leq P[Z_n \leq (1-\varepsilon)c_n] \to 0.$$

It follows that $P[a_n^* x_0 \leq (1-\varepsilon)c_n] = 0$, i.e., $c_n/a_n^* < (1-\varepsilon)^{-1}x_0$, for all large n. Hence $\limsup_{n\to\infty} c_n/a_n^* \leq x_0$. Thus $c_n \sim a_n^* x_0$, and it is an easy consequence of Lemma 3 that $a_n^* \sim a_{n-1}^*$. □

Lemma 5. Suppose that $0 < a_1 \leq a_2 \leq \cdots \uparrow \infty$. For $\delta \in (0,1)$ and all n so large that $a_n \geq \delta^{-1}a_1$, define

$$k_n = k_n(\delta) = \max\{j : a_j \leq \delta a_n\} \text{ and}$$
$$U_n = U_n(\delta) = \max\{X_j : k_n < j \leq n\}.$$

The following statements are equivalent:

(i) $\{Z_n, n \geq 1\}$ is relatively stable;

(ii) $a_n \sim a_{n-1}$;

(iii) $n - k_n(\delta) \to \infty$ as $n \to \infty$ for every $\delta \in (0,1)$;

(iv) $U_n(\delta) \xrightarrow{p} x_0$ as $n \to \infty$ for every $\delta \in (0,1)$.

Proof: That (i) implies (ii) is immediate from Lemma 4, since $a_n^* = a_n, n \geq 1$.

Suppose $a_n \sim a_{n-1}$. Then, for each integer $m \geq 1$, $\dfrac{a_n}{a_{n-m}} = \prod_{i=1}^{m} \dfrac{a_{n-i+1}}{a_{n-i}} \to 1$. If $n - k_n(\delta) = m$, then

$$\frac{a_n}{a_{n-m}} = \frac{a_n}{a_{k_n}} \geq \delta^{-1} > 1.$$

Thus, if $n - k_n(\delta) = m$ for an infinite number of values of n and some fixed $m \geq 1$ and $\delta \in (0,1)$, then it would follow that $\limsup_{n\to\infty} a_n/a_{n-m} \geq \delta^{-1} > 1$, a contradiction. Hence, $n - k_n(\delta) \to \infty$ for every $\delta \in (0,1)$; that is, (ii) implies (iii).

Note that $U_n(\delta) \leq x_0$ a.s. for every $n \geq 1$ and $0 < \delta < 1$, so

$$\begin{aligned} U_n(\delta) \xrightarrow{p} x_0 &\text{ iff } P[U_n(\delta) \leq (1-\varepsilon)x_0] \to 0 \text{ for every } \varepsilon > 0 \\ &\text{iff } F^{n-k_n}((1-\varepsilon)x_0) \to 0 \text{ for every } \varepsilon > 0 \\ &\text{iff } n - k_n(\delta) \to \infty \text{ for every } \delta \in (0,1). \end{aligned}$$

Therefore, (iii) and (iv) are equivalent.

Finally, suppose that $U_n(\delta) \xrightarrow{p} x_0$ for all $0 < \delta < 1$. In view of Lemma 1 and the definition of $k_n(\delta)$,

$$x_0 \geq \frac{Z_n}{a_n} \geq \frac{\max\{a_j X_j : k_n(\delta) < j \leq n\}}{a_n} > \delta U_n(\delta) \text{ a.s.}$$

for every $\delta \in (0,1)$. For any $\varepsilon > 0$, choose δ such that $1 - \varepsilon < \delta < 1$. Then

$$P[Z_n \leq (1-\varepsilon)a_n x_0] \leq P[U_n(\delta) \leq (1-\varepsilon)x_0/\delta] \to 0,$$

since $U_n(\delta) \xrightarrow{p} 1$. But $Z_n/a_n \leq x_0$ a.s., so $Z_n/a_n \xrightarrow{p} x_0$. Hence, (iv) implies (i). □

Here is the main result of the paper.

Theorem 1. Let $X_1, X_2, \cdots$ be non-degenerate *i.i.d.* random variables with d.f. F such that $F(0) < 1$ and $F(a) = 1$ for some $a > 0$. Let $\{a_n\}$ be a positive real sequence such that $\sup_{n \geq 1} a_n = \infty$, and define $\{Z_n\}$ by (1). Then $\frac{Z_n}{c_n} \xrightarrow{p} 1$ for some real sequence $\{c_n\}$ iff $a_n^* \sim a_{n-1}^*$, where $a_n^* = \max\{a_1, \cdots, a_n\}, n \geq 1$. Moreover, without loss of generality, $c_n = a_n^* x_0$.

Proof: In view of Lemma 4, it will suffice to show that $\{Z_n\}$ is relatively stable if $a_n^* \sim a_{n-1}^*$. To do so, define $m_1 = 1$ and, for $j \geq 1, m_{j+1} = \min\{n : a_n > a_{m_j}\}$. Then $a_{m_1} \leq a_{m_2} \leq \cdots \uparrow \infty$; $a_n^* = a_{m_j}^* = a_{m_j}$ if $m_j \leq n < m_{j+1}$; and $a_{m_j} = a_{m_j}^* \sim a_{m_j - 1}^* = a_{m_{j-1}}^* = a_{m_{j-1}}$ as $j \to \infty$, using the assumption $a_n^* \sim a_{n-1}^*$. It now follows immediately from Lemma 5 that

$$\max_{j \leq N} a_{m_j} X_{m_j} / (a_{m_N} x_0) \xrightarrow{p} 1 \text{ as } N \to \infty. \tag{3}$$

But, for each $n \geq 1$, there is a unique $N \geq 1$ such that $m_N \leq n < m_{N+1}$. Therefore,

$$\frac{\max_{j \leq N} a_{m_j} X_{m_j}}{a_{m_N}} \leq \frac{Z_n}{a_{m_N}} = \frac{Z_n}{a_n^*} \leq x_0 \text{ a.s.} \tag{4}$$

It is an easy consequence of (3) and (4) that $Z_n/(a_n^* x_0) \xrightarrow{p} 1$. □

Remark 1. If X_1 were degenerate and finite, then $X_n = x_0$ a.s. Thus $Z_n = a_n^* x_0$ a.s. in this case. If $\sup_{n \geq 1} a_n < \infty$, then $Z_n \to \lambda$ a.s. for a constant λ iff $\sup_{n \geq m} a_n = \sup_{n \geq 1} a_n$ for every $m \geq 1$, in which case $\lambda = a_n^* x_0$ (see Tomkins [2]); clearly, $a_n^* \sim a_{n-1}^*$ and $Z_n/(a_n^* x_0) \to 1$ a.s. in this case.

Remark 2. If $x_0 = \infty$ and $a_n = 1$, $n \geq 1$, then the relative stability of $\max\{X_1, \cdots, X_n\}$ was completely solved by Gnedenko [1]. The author intends to address the relative stability of $\{Z_n\}$ when $x_0 = \infty$ in a future publication.

Remark 3. As noted above, the relative stability problem for the maximum sequence $M_n = \max\{X_1, \cdots, X_n\}$ has been fully resolved in the case where $X_1, X_2, \cdots$ are *i.i.d.* with $F(x) < 1$ for all x. A natural extension is to ask what happens to M_n as $n \to \infty$ where $X_1, X_2, \cdots$ are merely independent with $P[X_n \leq x] < 1$ for all $n \geq 1$ and all real x. Theorem 1 above may be interpreted as a first step towards the solution of this general problem.

Acknowledgment

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The author is grateful for the comments of two referees.

References

1. B. Gnedenko, *Ann. Math.* **44**, 423 (1943).
2. R. J. Tomkins, *Proceedings of the Conference on Extreme Value Theory and Applications* (Gaithersburg, MD, May 1993) **3**, 197 (1994).

TRANSIENT ANALYSIS OF SOME INFINITE SERVER QUEUES

GORDON E. WILLMOT AND STEVE DREKIC

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

E-mail: sdrekic@hopper.math.uwaterloo.ca

In this paper, it is demonstrated using generating function arguments that the transient distribution of the number of customers in the system in many $M^X/G/\infty$ queues may be computed in a straightforward manner for a wide variety of bulk arrival size distributions. Numerical examples illustrate the ease of implementation of the computational procedure.

In this paper, we consider an infinite server queueing system in which customers arrive according to a Poisson process at rate $\lambda > 0$. We assume that the bulk arrival size random variable X is distributed with $q_k = Pr\{X = k\}, k \geq 1$, and probability generating function (pgf) $Q(z) = \sum_{k=1}^{\infty} q_k z^k$. We assume that the individual service times are independent and identically distributed positive random variables with distribution function (df) $F(y) = 1 - \bar{F}(y) = Pr\{Y \leq y\}$.

This is the $M^X/G/\infty$ queue and the purpose of this paper is to present a computational procedure for determining the distribution of the number of customers in the system (equivalently, the number of busy servers) at time t for a particular class of service time distributions. The $M^X/G/\infty$ queue can be used to model any system involving delays with no congestion, and is thus quite general in scope. In particular, the $M^X/G/\infty$ queue has important applications in a variety of information transmission and storage systems. It has also been used to model the number of incurred but not reported (IBNR) claims in an insurance context (see Willmot and Lin [1]). Theoretical results concerning the transient analysis of this system are contained in the excellent reference by Chaudhry and Templeton [2], and much of this is attributed to the pioneering work of Shanbag [3], Abol'nikov [4], Reynolds [5], and Brown and Ross [6]. These references do not place much emphasis on the computational aspects of the problem. However, Willmot and Drekic [7] have recently introduced a straightforward approach to computing the transient distribution of the number in system in the $M^X/M/\infty$ queue for a wide variety of bulk arrival size distributions.

Let N_t represent the number of customers in the system at time t, with $p_n(t) = Pr\{N_t = n\}$, $n \geq 0$. Assuming $N_0 = 0$, it follows that N_t has a

compound Poisson distribution with pgf

$$P_t(z) = \sum_{n=0}^{\infty} p_n(t) z^n = e^{\lambda t[Q_t(z)-1]} \tag{1}$$

where

$$Q_t(z) = \sum_{k=0}^{\infty} q_k(t) z^k = \frac{1}{t} \int_0^t Q[F(x) + z\bar{F}(x)]dx. \tag{2}$$

The derivation of this result follows that of Brown and Ross [6], who actually consider a more general model. We reproduce this derivation here for completeness. The probability that $m \geq 1$ batches of customers arrive by time t is $(\lambda t)^m e^{-\lambda t}/m!$. The conditional distribution of the (unordered) arrival times of these m batches are independent and uniformly distributed over $(0,t)$ with constant probability density function $1/t$ (see Ross [8], section 5.3). If a batch arrives at time x where $0 < x < t$, a customer from this batch is still in the system with probability $\bar{F}(t-x)$ and has left the system with probability $F(t-x)$. Therefore, the number of customers remaining in the system at time t from a batch arriving at time x has conditional pgf

$$\int_0^t Q[F(t-x) + z\bar{F}(t-x)]\frac{dx}{t} = \frac{1}{t}\int_0^t Q[F(x) + z\bar{F}(x)]dx,$$

which is, with probabilities $\{q_m(t);\ m = 1,2,3,\ldots\}$, equation (2). The total number in the system at time t is obtained by summing over the m (independent) batches, i.e. the pgf is thus raised to the power m, so that

$$P_t(z) = e^{-\lambda t} + \sum_{m=1}^{\infty} \frac{(\lambda t)^m e^{-\lambda t}}{m!} \{Q_t(z)\}^m$$

which is equation (1).

In the case of the compound Poisson with pgf (1), the following recursive formula can be used to compute $\{p_n(t);\ n = 0,1,2,\ldots\}$ (see Klugman *et al.* [9], pp. 239-240 or Tijms [10], p. 37):

$$p_0(t) = \exp\Big\{ -\lambda t \sum_{k=1}^{\infty} q_k(t) \Big\}, \tag{3}$$

$$p_n(t) = \frac{\lambda t}{n} \sum_{k=1}^{n} k q_k(t) p_{n-k}(t), \quad n = 1,2,3,\ldots. \tag{4}$$

Also, moments of N_t are easily obtainable from (1) and (factorial) moments associated with (2) (see Klugman *et al.* [9]).

In this paper, our goal is to demonstrate that numerical evaluation of the probabilities $\{p_n(t);\ n = 0, 1, 2, \ldots\}$ using (3) and (4) is relatively straightforward. In what follows, we shall show that under certain conditions a recursive formula may be used to obtain $\{q_k(t);\ k = 1, 2, 3, \ldots\}$. This formula normally requires evaluation of $q_0(t) = Q_t(0)$. Simplification results if a closed-form expression for $Q_t(x)$ is available. We briefly review some situations where this is the case, and then proceed to the derivation of the recursive formula alluded to above. We then reconsider evaluation of $q_0(t)$ in light of the recursion, and follow this with three examples.

In general, it is difficult to obtain a simple form for the pgf (2). However, this may be done in some special cases. For example, suppose that (see Willmot [11])

$$q_k = \frac{\alpha\Gamma(k-\alpha)}{k!\Gamma(1-\alpha)}, \quad k = 1, 2, 3, \ldots,$$

where $0 < \alpha < 1$, so that

$$Q(z) = 1 - (1-z)^{\alpha}. \tag{5}$$

Then from (2), it follows that

$$\begin{aligned} Q_t(z) &= \frac{1}{t}\int_0^t \left\{1 - \left[1 - F(x) - z\bar{F}(x)\right]^{\alpha}\right\} dx \\ &= \frac{1}{t}\int_0^t \left[1 - (1-z)^{\alpha}\bar{F}(x)^{\alpha}\right] dx \\ &= 1 - \theta_t(1-z)^{\alpha} \end{aligned}$$

where

$$\theta_t = \frac{1}{t}\int_0^t \bar{F}(x)^{\alpha} dx.$$

Evidently, $0 \leq \theta_t \leq 1$, and from (5) one has $Q_t(z) = 1 - \theta_t + \theta_t Q(z)$, from which it follows that

$$q_0(t) = 1 - \theta_t$$

and

$$q_k(t) = \theta_t q_k, \quad k = 1, 2, 3, \ldots.$$

In the case of exponential service, Willmot and Drekic [7] (see p. 142 for details) have shown that if X has a zero-truncated geometric distribution (see Klugman *et al.* [9], p. 590), both (1) and (2) simplify to reveal that N_t has a compound negative binomial – geometric distribution with easily calculable

parameters. As another example where service times are exponentially distributed at rate $\mu > 0$, suppose that X has an extended truncated negative binomial (ETNB) distribution with pgf (see Willmot [11], p. 18)

$$Q(z) = \frac{[1-2\beta(z-1)]^{\frac{1}{2}} - (1+2\beta)^{\frac{1}{2}}}{1-(1+2\beta)^{\frac{1}{2}}}, \quad \beta > 0. \tag{6}$$

Subsequent substitution into (2) then yields

$$\begin{aligned} Q_t(z) &= \frac{1}{t}\int_0^t \frac{[1-2\beta e^{-\mu x}(z-1)]^{\frac{1}{2}} - (1+2\beta)^{\frac{1}{2}}}{1-(1+2\beta)^{\frac{1}{2}}}dx \\ &= \left(\frac{x}{t} + 2\frac{[1-2\beta e^{-\mu x}(z-1)]^{\frac{1}{2}} - \ln\{1+[1-2\beta e^{-\mu x}(z-1)]^{\frac{1}{2}}\}}{\mu t[(1+2\beta)^{\frac{1}{2}}-1]}\right)\Bigg|_{x=0}^{x=t} \\ &= 1 + \frac{2}{\mu t[(1+2\beta)^{\frac{1}{2}}-1]}\left([1-2\beta e^{-\mu t}(z-1)]^{\frac{1}{2}} - [1-2\beta(z-1)]^{\frac{1}{2}}\right. \\ &\quad \left. + \ln\left\{\frac{1+[1-2\beta(z-1)]^{\frac{1}{2}}}{1+[1-2\beta e^{-\mu t}(z-1)]^{\frac{1}{2}}}\right\}\right). \end{aligned} \tag{7}$$

Hence, from (1), we get

$$P_t(z) = \left(\frac{\{1+[1-2\beta(z-1)]^{\frac{1}{2}}\}e^{-[1-2\beta(z-1)]^{\frac{1}{2}}}}{\{1+[1-2\beta e^{-\mu t}(z-1)]^{\frac{1}{2}}\}e^{-[1-2\beta e^{-\mu t}(z-1)]^{\frac{1}{2}}}}\right)^{\frac{2\lambda}{\mu[(1+2\beta)^{\frac{1}{2}}-1]}}.$$

It is clear from this last example that it may be difficult to extract the coefficient of z^n in $Q_t(z)$ as defined by (2). (Of course, one could expand the integrand in powers of z^n and integrate term by term, but this is not very suitable, since the resulting expression is cumbersome).

However, if the service time distribution has a failure rate whose reciprocal is a linear function, a simple computational formula for the probabilities $\{q_k(t);\ k = 1,2,3,\ldots\}$ is obtainable. That is, let us assume that the failure rate of the service time distribution has the form

$$r(y) = \frac{F'(y)}{\bar{F}(y)} = \frac{1}{\sigma + \theta y}. \tag{8}$$

Note that differentiation of (2) yields

$$\begin{aligned} Q_t'(z) = \frac{d}{dz}Q_t(z) &= \frac{1}{t}\int_0^t Q'[F(x) + z\bar{F}(x)]\bar{F}(x)dx \\ &= \frac{1}{t}\int_0^t \frac{\bar{F}(x)}{(1-z)F'(x)}Q'[F(x)+z\bar{F}(x)](1-z)F'(x)dx \end{aligned}$$

$$= \frac{1}{t(1-z)} \int_0^t \frac{1}{r(x)} dQ[F(x) + z\bar{F}(x)]. \tag{9}$$

By substituting (8) into (9), multiplying both sides by $1 - z$, and applying integration by parts, we obtain that $Q_t(z)$ satisfies the differential equation

$$Q_t'(z)(1-z) = \left(\frac{\sigma}{t} + \theta\right) Q[F(t) + z\bar{F}(t)] - \frac{\sigma}{t} Q(z) - \theta Q_t(z). \tag{10}$$

Although we intend to use (10) to derive a computational formula for the probabilities $\{q_k(t);\ k = 1, 2, 3, \ldots\}$, we remark that moments also follow easily from (10). In particular, by differentiating both sides of (10) r times, $r \geq 0$, we immediately get

$$Q_t^{(r+1)}(z)(1-z) - rQ_t^{(r)}(z) = \left(\frac{\sigma}{t} + \theta\right) \bar{F}(t)^r Q^{(r)}[F(t) + z\bar{F}(t)] - \frac{\sigma}{t} Q^{(r)}(z) - \theta Q_t^{(r)}(z). \tag{11}$$

Evaluating (11) at $z = 1$ gives rise to the following expression:

$$Q_t^{(r)}(1) = \frac{Q^{(r)}(1)}{r - \theta} \left[\frac{\sigma}{t}\left(1 - \bar{F}(t)^r\right) - \theta \bar{F}(t)^r\right], \quad \theta \neq r. \tag{12}$$

Of course, it is always the case that moments may be evaluated directly from (2), giving

$$Q_t^{(r)}(1) = \frac{Q^{(r)}(1)}{t} \int_0^t \bar{F}(x)^r dx. \tag{13}$$

Equation (10) may be used to derive a recursive relationship between the coefficients $q_n(t)$ of z^n in the expansion of $Q_t(z)$. First of all, define $q_{n,t}^*$ by introducing the pgf

$$Q_t^*(z) = Q[F(t) + z\bar{F}(t)] = \sum_{n=0}^{\infty} q_{n,t}^* z^n, \quad t \geq 0. \tag{14}$$

The probabilities $\{q_{n,t}^*;\ n = 0, 1, 2, \ldots\}$ can be readily obtained for a wide variety of bulk arrival size distributions of practical interest. In particular, let us consider distributions $\{q_k;\ k = 1, 2, 3, \ldots\}$ whose pgf is of the form

$$Q(z) = \sum_{k=1}^{\infty} q(k, \tau) z^k = \frac{B[\tau(1-z)] - B(\tau)}{1 - B(\tau)} \tag{15}$$

where τ is a parameter and $B(\cdot)$ is a function not depending on τ. This includes many standard distributions such as the truncated negative binomial, truncated Poisson, truncated binomial, and logarithmic series (see Willmot [11]). Thus, $q_k = q(k, \tau)$, and from (14)

$$Q_t^*(z) = \frac{B[\tau\bar{F}(t)(1-z)] - B(\tau)}{1 - B(\tau)}.$$

We then obtain that

$$Q^{(n)}(z) = \frac{(-1)^n \tau^n B^{(n)}[\tau(1-z)]}{1 - B(\tau)} \tag{16}$$

and

$$Q_t^{*(n)}(z) = \frac{(-1)^n [\tau\bar{F}(t)]^n B^{(n)}[\tau\bar{F}(t)(1-z)]}{1 - B(\tau)}, \tag{17}$$

from which it follows by setting $z = 0$ and dividing by $n!$ that

$$q_{0,t}^* = \frac{B[\tau\bar{F}(t)] - B(\tau)}{1 - B(\tau)},$$

$$q_{n,t}^* = \frac{1 - B[\tau\bar{F}(t)]}{1 - B(\tau)} q[n, \tau\bar{F}(t)], \quad n = 1, 2, 3, \ldots.$$

If one now equates coefficients of z^n in (10), one obtains

$$q_{n+1}(t) = \frac{1}{n+1}\left[(n-\theta)q_n(t) + \left(\frac{\sigma}{t} + \theta\right) q_{n,t}^* - \frac{\sigma}{t} q_n\right], \; n = 0, 1, 2, \ldots. \tag{18}$$

This is a first order linear difference equation which is easy to solve analytically, but the resulting solution offers little insight. Thus, for $\theta \neq 0$, our intention is to use (18) to generate $q_n(t)$ numerically. Note that (18) is the only place where the starting point $q_0(t)$ is needed (unless we have exponential service where $\theta = 0$, in which case considerable simplification results – see Example 1 below). Assuming $q_0(t)$ is available, the probabilities $\{q_n(t);\ n = 1, 2, 3, \ldots\}$ may be obtained recursively via (18). For a discussion of numerical issues related to recursions of this nature, see Press *et al.* [12], section 5.4.

In general, two approaches suggest themselves as a means of computing $q_0(t)$. First of all, if it is possible to obtain a simple analytical expression for $Q_t(z)$, as in the example which gave rise to (7), then one can substitute $z = 0$ easily in $Q_t(z)$ to get $q_0(t) = Q_t(0)$. Otherwise, a Taylor series expansion of $Q_t(z)$ about $z = 1$ yields

$$Q_t(z) = 1 + \sum_{k=1}^{\infty} \frac{Q_t^{(k)}(1)}{k!}(z-1)^k, \tag{19}$$

where $Q_t^{(k)}(1)$ is given by (12) or (13). Then, we set $z = 0$ in (19) and use (13) to obtain

$$q_0(t) = 1 + \frac{1}{t}\sum_{k=1}^{\infty}(-1)^k \frac{Q^{(k)}(1)}{k!} a_k(t) \tag{20}$$

where $a_k(t) = \int_0^t \bar{F}(x)^k dx$. We point out that: (i) $Q^{(k)}(1)$ may be computed quite readily via (16) for $Q(z)$ satisfying (15), and (ii) $a_k(t)$ may be evaluated explicitly for $F(x)$ satisfying (8). Moreover, since the infinite series in (20) is an alternating one, an effective and efficient means of computing this series is to employ *Euler summation.* For details regarding Euler summation and the control of the approximation error, the reader is referred to Abate *et al.* [13], p. 270 or Press *et al.* [12], section 5.1.

We now present three special cases:

Example 1: Exponential Service Time Distribution. If $F(y) = 1 - e^{-\mu y}$ for $y \geq 0$ where $\mu > 0$, one obtains $r(y) = \mu$ and so (18) is applicable with $\sigma = \mu^{-1}$ and $\theta = 0$. Substituting these parameters into (18) immediately yields

$$(n+1)q_{n+1}(t) - nq_n(t) = \frac{1}{\mu t}\left(q_{n,t}^* - q_n\right). \tag{21}$$

Summing both sides of (21) from $n = 0$ to $n = k - 1$ easily gives rise to the following explicit expression for $q_k(t)$, namely

$$q_k(t) = \frac{\sum_{n=0}^{k-1}(q_{n,t}^* - q_n)}{k\mu t}, \quad k = 1, 2, 3, \ldots, \tag{22}$$

in agreement with the result found by Willmot and Drekic [7], p. 139, eq. (8).

Example 2: Pareto Service Time Distribution. If $F(y) = 1 - \left(\frac{\mu}{\mu+y}\right)^{\alpha}$ for $y \geq 0$ where $\alpha,\ \mu > 0$, it is easy to verify that

$$r(y) = \frac{\alpha}{\mu + y},$$

and so (18) is applicable with $\sigma = \mu/\alpha$ and $\theta = 1/\alpha$. To compute $q_0(t)$, we employ (20) with

$$a_k(t) = \begin{cases} \frac{\mu}{1-\alpha k}\left[\left(1 + \frac{t}{\mu}\right)^{-\alpha k+1} - 1\right] & : \alpha k \neq 1 \\ \mu \ln\left(1 + \frac{t}{\mu}\right) & : \alpha k = 1 \end{cases}. \tag{23}$$

We remark that the Pareto distribution may be a good alternative to the exponential if it is felt that a more heavily tailed service time distribution is better suited to the model (for example, see Harris [14]).

Example 3: Complementary Power Function Service Time Distribution. The power function df (see Evans *et al.* [15], p. 161) is $H(v) = Pr\{V \leq v\} = \left(\frac{v}{\omega}\right)^{\beta}$ for $0 \leq v \leq \omega$ where $\beta,\ \omega > 0$. Suppose that the service time random variable Y is given by $Y = \omega - V$ with df $F(y) = 1 - H(\omega - y)$, that is $F(y) = 1 - \left(1 - \frac{y}{\omega}\right)^{\beta}$ for $0 \leq y \leq \omega$. This subclass of the beta distribution includes the continuous uniform as the special case $\beta = 1$. It is easy to verify that

$$r(y) = \frac{\beta}{\omega - y}, \quad 0 \leq y \leq \omega,$$

and so (18) is applicable with $\sigma = \omega/\beta$ and $\theta = -1/\beta$ if $0 \leq t \leq \omega$. Note that if $t > \omega$, (2) becomes

$$\begin{aligned} Q_t(z) &= \frac{1}{t}\left\{\int_0^{\omega} Q\left[1 - \left(1 - \frac{y}{\omega}\right)^{\beta} + z\left(1 - \frac{y}{\omega}\right)^{\beta}\right] dy + \int_{\omega}^{t} Q(1)dy\right\} \\ &= \frac{\omega}{t} Q_{\omega}(z) + \left(1 - \frac{\omega}{t}\right). \end{aligned} \tag{24}$$

Substitution of (24) into (1) then yields (for $t > \omega$)

$$P_t(z) = \exp\left\{\lambda t\left[\frac{\omega}{t} Q_{\omega}(z) + 1 - \frac{\omega}{t} - 1\right]\right\} = \exp\{\lambda\omega[Q_{\omega}(z) - 1]\} = P_{\omega}(z),$$

and so the results also extend to the case $t > \omega$. Finally, to compute $q_0(t)$, we employ (20) with

$$a_k(t) = \frac{\omega}{1 + \beta k}\left[1 - \left(1 - \frac{t}{\omega}\right)^{\beta k + 1}\right]. \tag{25}$$

To illustrate the effectiveness of the computational procedure, we consider a simple numerical example in which $\lambda = 2$ and X has a zero-truncated negative binomial distribution given by

$$q_n = \frac{f_n(m,p)}{1 - f_0(m,p)}, \quad n = 1, 2, 3, \ldots$$

where

$$f_n(m,p) = \binom{m+n-1}{n} p^m (1-p)^n, \quad n = 0, 1, 2, \ldots. \tag{26}$$

In this case, (15) is satisfied with $\tau = (1-p)/p$ and $B(x) = (1+x)^{-m}$. Consider $m = 4$ and $p = 0.8$, so that the mean bulk arrival size is $625/369 \simeq 1.69$ customers. Tables 1 and 2 display the results (to seven decimal places of accuracy) for $p_n(t), n = 0, 1, 2, \ldots, 10$, for values of $t = 0.125$, 0.25, 0.5, 1, 5, and 10 under the assumptions that: (i) service times are Pareto distributed with parameters $\alpha = 2$ and $\mu = 5$, and (ii) uniformly distributed on the interval $(0, 10)$ (i.e. the complementary power function distribution with parameters $\beta = 1$ and $\omega = 10$). The mean service time in both cases is 5. The sums of the first 11 probabilities corresponding to these values of t are also included in the tables for comparative purposes.

Table 1. Results for $p_n(t)$ under Pareto service with $\alpha = 2$ and $\mu = 5$.

n	$t = 0.125$	$t = 0.25$	$t = 0.5$	$t = 1$	$t = 5$	$t = 10$
0	0.7814843	0.6148611	0.3881397	0.1664161	0.0021185	0.0001672
1	0.1083984	0.1704123	0.2143892	0.1815371	0.0093877	0.0011223
2	0.0606528	0.1055363	0.1584359	0.1772388	0.0235195	0.0040202
3	0.0285572	0.0564078	0.1025238	0.1485030	0.0434800	0.0101899
4	0.0123726	0.0281093	0.0616352	0.1127000	0.0657831	0.0204642
5	0.0051287	0.0134367	0.0351694	0.0794752	0.0859879	0.0345967
6	0.0020698	0.0062313	0.0192524	0.0528923	0.1003555	0.0511161
7	0.0008195	0.0028187	0.0101812	0.0335707	0.1068854	0.0676914
8	0.0003194	0.0012478	0.0052278	0.0204739	0.1055072	0.0818113
9	0.0001228	0.0005420	0.0026166	0.0120659	0.0976337	0.0914660
10	0.0000466	0.0002315	0.0012805	0.0069013	0.0854428	0.0955872
Sum	0.9999721	0.9998348	0.9988517	0.9917743	0.7261013	0.4582325

Table 2. Results for $p_n(t)$ under Uniform service on $(0, 10)$.

n	$t = 0.125$	$t = 0.25$	$t = 0.5$	$t = 1$	$t = 5$	$t = 10$
0	0.7794793	0.6086560	0.3731071	0.1433320	0.0002386	0.0000073
1	0.1081531	0.1688919	0.2070083	0.1588770	0.0012633	0.0000592
2	0.0613089	0.1070310	0.1588470	0.1642742	0.0038446	0.0002609
3	0.0292273	0.0584697	0.1066481	0.1462919	0.0087131	0.0008191
4	0.0128160	0.0297673	0.0665646	0.1183933	0.0162544	0.0020490
5	0.0053756	0.0145380	0.0394700	0.0892471	0.0262992	0.0043339
6	0.0021951	0.0068900	0.0224708	0.0636007	0.0380937	0.0080387
7	0.0008794	0.0031859	0.0123660	0.0432810	0.0504483	0.0133999
8	0.0003469	0.0014420	0.0066108	0.0283298	0.0619999	0.0204284
9	0.0001350	0.0006405	0.0034462	0.0179332	0.0714954	0.0288568
10	0.0000519	0.0002798	0.0017572	0.0110248	0.0780155	0.0381517
Sum	0.9999685	0.9997921	0.9982961	0.9845850	0.3566660	0.1164049

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada.

References

1. G.E. Willmot and X.S. Lin, *Lundberg Approximations for Compound Distributions with Insurance Applications* (Springer-Verlag, New York, 2001).
2. M.L. Chaudhry and J.G.C. Templeton, *A First Course in Bulk Queues* (John Wiley & Sons, New York, 1983).
3. D.N. Shanbag, *Journal of Applied Probability* **3**, 274 (1966).
4. L.M. Abol'nikov, *Probl. Inform. Transm.* **4**, 82 (1968).
5. J.F. Reynolds, *Operations Research* **16**, 186 (1968).
6. M. Brown and S.M. Ross, *Journal of Applied Probability* 6604 1969.
7. G.E. Willmot and S. Drekic, *Operations Research Letters* **28**, 137 (2001).
8. S.M. Ross, *Introduction to Probability Models*, 7th Edition (Academic Press, San Diego, 2000).
9. S.A. Klugman, H.H. Panjer and G.E. Willmot, *Loss Models: From Data to Decisions* (John Wiley & Sons, New York, 1998).
10. H. Tijms, *Stochastic Modelling and Analysis: A Computational Approach* (John Wiley & Sons, Chichester, 1986).
11. G.E. Willmot, *ASTIN Bulletin* **18**, 17 (1988).
12. W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, *Numerical Recipes – The Art of Scientific Computing* (Cambridge University Press, New York, 1986).
13. J. Abate, G.L. Choudhury and W. Whitt, In *Computational Probability*, ed. W.K. Grassmann (Kluwer Academic Publishers, Boston, 2000).
14. C.M. Harris, *Operations Research* **16**, 307 (1968).
15. M. Evans, N. Hastings and B. Peacock, *Statistical Distributions*, 3rd Edition (John Wiley & Sons, New York, 2000).

EMPIRICAL LIKELIHOOD METHOD FOR FINITE POPULATIONS

CHANGBAO WU
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo ON N2L 3G1 Canada
E-mail: cbwu@uwaterloo.ca

This article provides an overview on recent developments of empirical likelihood methods in estimating the finite population means and totals, distribution function and quantiles, variance and other quadratic functions in the presence of auxiliary information. Major results are unified under the general framework of optimal estimation and model-calibration. Applications of the method to obtaining range-restricted weights in regression estimators and estimation under measurement error models are also presented.

1 Introduction

The empirical likelihood method was proposed by Owen [11,12] as a device for constructing confidence regions with independent observations. Owen proved that the empirical likelihood ratio statistic has an asymptotic χ^2 distribution and therefore is useful for interval estimation and hypothesis testing. Qin and Lawless [15,16] discovered that the empirical likelihood method is also a powerful tool for point estimation when side information can be incorporated into constrained maximization of the empirical likelihood function. The method soon became popular and major developments have been summarized in the recent book by Owen [13].

Historically, the first application of the concept behind empirical likelihood was suggested by Hartley and Rao [10] for finite populations. They assume that the study variable y takes only a finite set of scale points $y[1], y[2], \cdots, y[k]$. For a given sample $\{y_i, i \in s\}$, let n_j be the number of y_i's in s that take scale point $y[j]$. Under simple random sampling (SRS), $(n_1, \cdots, n_k)$ follows a multivariate hyper-geometric distribution. When the population size N is large, one can use the likelihood function from a multinomial distribution. Auxiliary information can be incorporated to find the constrained maximum likelihood estimators of the population proportions for each of the scale points. These estimated proportions can then be used to construct the so-called scale-load estimators for the finite population mean $\bar{Y} = N^{-1} \sum_{i=1}^{N} y_i$.

The first formal application of the empirical likelihood method in sur-

vey sampling was introduced by Chen and Qin [3] under SRS. Let p_i be the probability mass at y_i for $i \in s$ and $\bar{\boldsymbol{X}}$ be the known population means for the (vector-valued) auxiliary variable $\boldsymbol{x}$. The empirical maximum likelihood estimator of $\bar{Y}$ is defined as $\hat{\bar{Y}}_{EL} = \sum_{i \in s} \hat{p}_i y_i$ where the $\hat{p}_i$'s maximize the empirical likelihood function $L(\boldsymbol{p}) = \prod_{i \in s} p_i$ subject to

$$\sum_{i \in s} p_i = 1 \ (p_i > 0) \quad \text{and} \quad \sum_{i \in s} p_i \boldsymbol{x}_i = \bar{\boldsymbol{X}} \,. \tag{1}$$

Chen and Qin [3] showed that $\hat{\bar{Y}}_{EL}$ is asymptotically equivalent to the conventional regression estimator and therefore is very efficient.

The rest of this paper is organized as follows. In Section 2, we introduce the pseudo empirical likelihood approach under a general sampling design for the estimation of $\bar{Y}$ (Chen and Sitter [4]). The concept of optimal calibration estimation is presented along with the model-calibrated pseudo empirical likelihood (MCPE) approach of Wu and Sitter [21]. In Sections 3 and 4, the MCPE estimators are applied naturally to the estimation of distribution functions and quadratic finite population functions including the population variance. A simple solution to obtain range-restricted weights in regression estimators using the empirical likelihood method is presented in Section 5. In Section 6, we discuss how the empirical likelihood method can be applied to various measurement error problems. We conclude with some remarks in Section 7.

2 Pseudo Empirical Likelihood Approach

Sample data from a finite population obtained through an unequal probability sampling scheme are usually highly correlated with each other. What will be the "empirical likelihood function" to use under a general sampling design? Chen and Sitter [4] proposed a pseudo empirical likelihood function based on a two-stage argument which can be viewed as a non-parametric version of the estimation strategy discussed in Binder [1] and Godambe and Thompson [9]: if the data from the entire finite population, $\{y_1, y_2, \cdots, y_N\}$, is known, the correct empirical likelihood function would be $L(\boldsymbol{p}) = \prod_{i=1}^{N} p_i$. The corresponding empirical log-likelihood function $l(\boldsymbol{p}) = \sum_{i=1}^{N} \log(p_i)$ is a population total! The design-based unbiased (Horvitz-Thompson) estimator for this total is given by

$$\hat{l}(\boldsymbol{p}) = \sum_{i \in s} d_i \log(p_i) \,, \tag{2}$$

where $d_i = 1/\pi_i$ are the basic design weights and $\pi_i = P(i \in s)$ are the inclusion probabilities. The $\hat{l}(\boldsymbol{p})$ was referred to as pseudo empirical (log) likelihood

function. The pseudo empirical maximum likelihood estimator (PEML) of $\bar{Y}$ (Chen and Sitter [4]) was defined as $\hat{\bar{Y}}_{PE} = \sum_{i\in s} \hat{p}_i y_i$ where the $\hat{p}_i$'s maximize $\hat{l}(\boldsymbol{p})$ subject to constraints (1), assuming $\bar{\boldsymbol{X}}$ is known.

The $\hat{p}_i$'s are given by $\hat{p}_i = d_i^*/[1 + \boldsymbol{\lambda}'(\boldsymbol{x}_i - \bar{\boldsymbol{X}})]$, where $d_i^* = d_i/\sum_{i\in s} d_i$ and the vector Lagrange multiplier, $\boldsymbol{\lambda}$, is the solution to $g(\boldsymbol{\lambda}) = \sum_{i\in s}[d_i^*(\boldsymbol{x}_i - \bar{\boldsymbol{X}})]/[1 + \boldsymbol{\lambda}'(\boldsymbol{x}_i - \bar{\boldsymbol{X}})] = 0$. A modified Newton-Raphson algorithm has been developed by Chen *et al.* [6] for finding this solution. The algorithm is guaranteed to converge if a solution exists.

A crucial component in the (pseudo) empirical likelihood estimation is the use of constraints (1) in the maximization process. There are two issues related to this and any other estimation procedures: efficiency and consistency. Efficiency is measured by the overall performance of the estimator in terms of bias and variance or mean square error; consistency refers to some internal conditions and requirements imposed by the surveyor. The second constraint in (1) is a commonly used consistency requirement called benchmark constraints. Benchmark constraints are often imposed in practice for two reasons: the surveyor believes that the weights which give perfect estimates for the auxiliary variables should also give a good estimate for the study variable; the auxiliary information is only available at the aggregate level, i.e. only $\bar{\boldsymbol{X}}$ is known. On the other hand, if complete auxiliary information $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ is known, a compelling question to ask would be "what is the best constraint to use in the (pseudo) empirical likelihood estimation?"

To put this more formally, let $u_i = u(\boldsymbol{x}_i)$, $i = 1, \cdots, N$, where $u(\cdot)$ is a known function. We use u_i as a calibration variable and replace the second constraint in (1) by

$$\sum_{i\in s} p_i u_i = \frac{1}{N}\sum_{i=1}^{N} u_i \,. \tag{3}$$

The question becomes "what kind of choice of u_i will make $\hat{\bar{Y}}_{PE}$ most efficient"? It is very unfortunate that in survey sampling uniformly minimum variance (unbiased) estimators do not exist. Indeed the only choice of u_i that results in a $\hat{\bar{Y}}_{PE}$ with minimum variance is $u_i \equiv y_i$ and this of course cannot be used.

The model-assisted optimal estimators using the criterion of minimum expected design variance under a superpopulation model have been discussed by several authors. See, for example, the work by Godambe [7], Godambe and Thompson [8], and Cassel *et al.* [2]. Suppose that $y_1, y_2, \cdots, y_N$ is a random

sample from a superpopulation such that

$$E_\xi(y_i) = \mu_i\,,\quad V_\xi(y_i) = \sigma_i^2\,, i = 1, 2, \cdots, N\,,$$

and $y_1, y_2, \cdots, y_N$ are independent of each other. Here E_ξ and V_ξ denote the expectation and variance under the superpopulation model. The following result has been established by Wu [20].

Theorem. Under proper asymptotic settings and for any regular sampling designs, the use of $u_i = \mu_i$ as a calibration variable in (1) will result in an estimator $\hat{\bar{Y}}_{PE}$ with minimum expected asymptotic design variance $E_\xi[AV_p(\hat{\bar{Y}}_{PE})]$ among all possible choices of $(u_1, u_2, \cdots, u_N)$ such that $N^{-1}\sum_{i=1}^N (u_i - \bar{U})^2 \to c \neq 0$.

Here AV_p refers to the asymptotic design-based variance and $\bar{U} = N^{-1}\sum_{i=1}^N u_i$. See Wu [20] for a detailed discussion on the asymptotic framework and a definition of regular sampling designs. Note that $\hat{\bar{Y}}_{PE}$ is robust against model misspecification, since $\hat{\bar{Y}}_{PE}$ is asymptotically design unbiased irrespective of the model but will be particularly efficient if the model adequately depicts the finite population. The gain of efficiency depends on the correlation between the response variable and the covariates.

Assume complete auxiliary information is available, Wu and Sitter [21] proposed a model-calibration approach to implement this optimal estimator. They adapted a semi-parametric model

$$E_\xi(y_i|\boldsymbol{x}_i) = \mu_i = \mu(\boldsymbol{x}_i, \boldsymbol{\theta})\,,\quad V_\xi(y_i|\boldsymbol{x}_i) = \sigma_i^2\,,\quad i = 1, 2, \cdots, N\,.$$

The fitted value $\hat{\mu}_i = \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$ is used as the calibration variable in constraints (1), where $\hat{\boldsymbol{\theta}}$ is a design-based estimator for the model parameter $\boldsymbol{\theta}$. The resulting PEML estimator was termed a model-calibrated pseudo empirical maximum likelihood (MCPE) estimator, denoted by $\hat{\bar{Y}}_{MC}$. They also showed that replacing $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$ in $\mu_i = \mu(\boldsymbol{x}_i, \boldsymbol{\theta})$ does not change the resulting estimator asymptotically. In addition, with probability close to 1, the MCPE estimator exists and can be computed using a simple bi-section algorithm (Chen *et al.* [6]).

This optimal model-calibration approach clarified several fundamental issues in using auxiliary information from surveys:

(i) The effective use of auxiliary information depends on both the parameters to be estimated and the actual relationship between the response variable and the covariates. Blindly calibrating over auxiliary variables is usually not a good approach.

(ii) The benchmark constraints used in (1) are justifiable if the relationship between y and $\boldsymbol{x}$ is close to linear. In this case the resulting PEML estimator is asymptotically equivalent to the optimal MCPE estimator obtained using $\hat{\mu}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\theta}}$ as the calibration variable (Wu and Sitter [21]). So benchmarking implies efficient estimation.
(iii) If the relationship between y and $\boldsymbol{x}$ is linear, knowing $\bar{\boldsymbol{X}}$ is "sufficient" for efficient estimation of population mean $\bar{Y}$ or total Y. If the relationship is nonlinear, or the parameters of interest involve a nonlinear function, complete auxiliary information and/or more advanced modeling are essential for "optimal" estimation.
(iv) The optimal model-calibration approach provides a unified framework for the estimation of distribution function and quantiles, variance and other quadratic functions in the presence of auxiliary information. In particular, the intrinsically positive weights, $\hat{p}_i > 0$, associated with the pseudo empirical likelihood method turn out to be a very valuable asset, as evidenced in the following sections.

Under stratified random sampling, Zhong and Rao [22] used a different formulation of the empirical likelihood function by noting that observations from different strata are independent of each other. Each stratum is therefore assigned with an independent empirical distribution and the "overall" empirical (log) likelihood function is the sum of all these strata (log) likelihood functions. We will return to this formulation in Section 6.

3 Estimating the Distribution Function and Quantiles

The finite population distribution function $F_Y(t) = N^{-1}\sum_{i=1}^{N} I(y_i \leq t)$ is also a finite population mean defined over an indicator variable $z_i = I(y_i \leq t)$. Without using any auxiliary information, estimation of $F_Y(t)$ is a special case of estimating the population mean and is usually straightforward. In the presence of auxiliary information, special attention needs be given to the following:

(a) While benchmark constraints sometimes are justifiable for the estimation of $\bar{Y}$, this consistency requirement is not needed for the estimation of $F_Y(t)$. Efficiency will be the primary concern.
(b) It is the indicator variable $z_i = I(y_i \leq t)$ that we have to work with. There is also an issue of local efficiency (particular value of t) versus global efficiency (an arbitrary t) in estimating $F_Y(t)$.
(c) It is desirable that an estimator of $F_Y(t)$, say $\hat{F}(t)$, is itself a distribution function, so quantile estimates can be obtained through direct inversion of

$\hat{F}(t)$.

Many techniques for estimating $\bar{Y}$, when applied directly to the estimation of $F_Y(t)$, will produce unsatisfactory results. For instance, in the case of a scalar x variable, a regression-type estimator for $F_Y(t)$ will have the form $\hat{F}_{reg}(t) = \hat{F}_Y(t) + \{F_X(t) - \hat{F}_X(t)\}\hat{B}$, where $\hat{F}_Y(t)$ and $\hat{F}_X(t)$ are Horvitz-Thompson type estimators for $F_Y(t)$ and $F_X(t) = N^{-1}\sum_{i=1}^{N} I(x_i \leq t)$, and $\hat{B}$ is the estimated slope of regressing $I(y_i \leq t)$ on $I(x_i \leq t)$. $\hat{F}_{reg}(t)$ suffers from several drawbacks. The obvious one is that $\hat{F}_{reg}(t)$ is not a distribution function and it can take values outside of [0, 1]. The model-assisted difference estimator proposed by Rao *et al.* [18] and the regression-type estimators or the bias-adjusted estimators discussed in Rao [17] have similar problems.

The pseudo empirical likelihood method combined with the optimal model-calibration approach can be readily applied here. The MCPE estimator of $F_Y(t)$ defined by Chen and Wu [5] is given by $\hat{F}_{MC}(t) = \sum_{i\in s} \hat{p}_i I(y_i \leq t)$ where the $\hat{p}_i$'s maximize $\hat{l}(\boldsymbol{p})$ subject to

$$\sum_{i\in s} p_i = 1 \ (p_i > 0) \quad \text{and} \quad \sum_{i\in s} p_i g_i = \frac{1}{N}\sum_{i=1}^{N} g_i \,. \tag{4}$$

For a fixed t, the optimal choice of g_i is given by

$$g_i = E_\xi\{I(y_i \leq t|\boldsymbol{x}_i\} = P(y_i \leq t|\boldsymbol{x}_i)\,.$$

It is now clear that no single set of weights $\hat{p}_i$ is optimal for an arbitrary t. Chen and Wu [5] proposed to use a fixed $t = t_0$ in computing the g_i while the resulting weights $\hat{p}_i$ are used for any t. With this treatment the MCPE estimator $\hat{F}_{MC}(t)$ will be a genuine distribution function and is very efficient for t in the neighborhood of t_0. Three different ways in computing the g_i's were proposed by Chen and Wu [5].

(1) Compute g_i under a regression model

$$y_i = \mu(\boldsymbol{x}_i, \boldsymbol{\theta}) + v_i\varepsilon_i\,, \quad i = 1, 2, \cdots, N\,,$$

where $v_i = v(\boldsymbol{x}_i)$ is a known function , the ε_i are independent and identically distributed (iid) random variates with mean 0 and variance σ^2. If the model is linear, $\mu(\boldsymbol{x}_i, \boldsymbol{\theta}) = \boldsymbol{x}_i'\boldsymbol{\theta}$, but other non-linear regression models will also work. Under this model one can use $g_i = P(y_i \leq t_0|\boldsymbol{x}_i) = G\{[t_0 - \mu(\boldsymbol{x}_i, \boldsymbol{\theta})]/v_i\}$, where $G(\cdot)$ is the cumulative distribution function of the ε_i's. Finally, one can replace $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$, and estimate $G(\cdot)$ using the fitted residuals $\hat{\varepsilon}_i$ if $G(\cdot)$ is unspecified.

(2) Compute g_i using a logistic regression model

$$\log[g_i/(1-g_i)] = \boldsymbol{x}_i'\boldsymbol{\theta}\,, \quad i = 1, 2, \cdots, N\,,$$

with variance function $V(g) = g(1-g)$. The fitted values, $\hat{g}_i$, are then used.
(3) Use pseudo fitted values $\hat{g}_i = I[\mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}) \leq t_0]$, where $\mu(\boldsymbol{x}_i, \boldsymbol{\theta}) = E_\xi(y_i|\boldsymbol{x}_i)$.

A simulation study reported by Chen and Wu [5] showed that the resulting estimators perform well in all three cases, with cases (1) and (2) slightly better than (3).

Estimation for the population quantile $\xi_\alpha = F_Y^{-1}(\alpha) = \inf\{t : F_Y(t) \geq \alpha\}$ can be obtained through direct inversion of $\hat{F}_{MC}(t)$: $\hat{\xi}_\alpha = \hat{F}_{MC}^{-1}(\alpha)$, where $0 < \alpha < 1$. A Bahadur representation for the quantile process $\hat{\xi}_\alpha$ has been established by Chen and Wu [5] under certain sampling designs.

4 Estimation of Variance and Quadratic Functions

Estimation of variance and other second-order finite population quantities using auxiliary information has been addressed by many survey researchers. Various techniques, such as regression, ratio and calibration estimation, have been attempted. See Sitter and Wu [19] for a literature review. A common weakness of these approaches is the ad hoc argument of applying certain techniques, which were originally developed for estimating $\bar{Y}$, to estimate the variance without a common framework that unifies the two types of finite population parameters.

The model-calibrated pseudo empirical likelihood method can be extended to handle variances and other second-order finite population parameters through a batch approach (Sitter and Wu [19]). Let $\boldsymbol{y}$ be the (possibly vector-valued) study variable(s). For parameters in a quadratic form, $T = \sum_{i=1}^{N}\sum_{j=i+1}^{N}\phi(\boldsymbol{y}_i, \boldsymbol{y}_j)$, which includes the population variance, covariance, and variance of a linear estimator as special cases, a unified estimation strategy is as follows.

(1) View T as a total over a synthetic finite population, i.e. $T = \sum_{\alpha=1}^{N^*} t_\alpha$ where $\alpha = (ij) = 1, 2, \cdots, N^*$, $t_\alpha = \phi(\boldsymbol{y}_i, \boldsymbol{y}_j)$ for $\alpha = (ij)$, and $N^* = N(N-1)/2$ is the total number of pairs.
(2) The sample over the synthetic population consists of all the pairs from the original sample: $s^* = \{(ij) : i < j, i, j \in s\}$.
(3) The "first-order" inclusion probabilities under this setting are $\pi_{ij} = P(i, j \in s)$, and the "basic design weights" are $d_{ij} = 1/\pi_{ij}$.

(4) The pseudo empirical (log) likelihood function is modified to accommodate all the pairs (ij) using the d_{ij}'s.

$$\hat{l}(\boldsymbol{p}) = \sum_{i \in s} \sum_{j>i} d_{ij} \log(p_{ij}),$$

where the p_{ij} is the probability mass assigned for $(\boldsymbol{y}_i, \boldsymbol{y}_j)$.

The MCPE estimator of T is defined as

$$\hat{T}_{MC} = N^* \sum_{i \in s} \sum_{j>i} \hat{p}_{ij} \phi(\boldsymbol{y}_i, \boldsymbol{y}_j),$$

where the $\hat{p}_{ij}$'s maximize the modified $\hat{l}(\boldsymbol{p})$ subject to

$$\sum_{i \in s} \sum_{j>i} p_{ij} = 1 \ (p_{ij} > 0), \ \sum_{i \in s} \sum_{j>i} p_{ij} \phi(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{y}}_j) = \frac{1}{N^*} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \phi(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{y}}_j).$$

Here the $\hat{\boldsymbol{y}}_i$'s are the fitted values for the $\boldsymbol{y}_i$'s, as discussed in Section 2.

The above approach brings a unified framework to the estimation of linear and quadratic parameters using auxiliary information. The approach is also model-assisted in that the resulting estimator $\hat{T}_{MC}$ is approximately design unbiased irrespective of the working (superpopulation) model used and will be very efficient if the model is adequate. Also, since the weights $\hat{p}_{ij}$ are always positive, the method ensures positive estimation for some known positive parameters such as variances.

The optimality of the MCPE approach for quadratic finite population parameters has also been established by Wu [20]. The method is generally applicable and improvement over the naive Horvitz-Thompson estimator is guaranteed. Results of a simulation study reported in Sitter and Wu [19] showed that $\hat{T}_{MC}$ performs very well for samples of small and moderate size.

5 Range-restricted Weights in Regression Estimation

The pseudo empirical likelihood method, combined with a novel idea of Chen *et al.* [6], provides a simple solution to the range-restricted weights problem in regression estimation. The generalized regression estimator $\hat{\bar{Y}}_{GR}$ for the population mean $\bar{Y}$ is probably the most popular one used by survey practitioners. It is computationally simple, very efficient if the relationship between y and $\boldsymbol{x}$ is nearly linear, and requires only $\bar{\boldsymbol{X}}$ to be known to compute the estimate. If one rewrites $\hat{\bar{Y}}_{GR}$ in the form of a weighted average, $\hat{\bar{Y}}_{GR} = \sum_{i \in s} w_i y_i$, the so-called GREG weights w_i also satisfy the benchmark constraints, i.e. $\sum_{i \in s} w_i \boldsymbol{x}_i = \bar{\boldsymbol{X}}$.

The GREG weights w_i, however, possess a very undesirable property: they can be very small or very large, and sometimes can even be negative. This has long been recognized by survey statisticians. Several iterative algorithms have been proposed to adjust the GREG weights so that the adjusted weights will satisfy the range-restrictions: $\gamma_1 \leq w_i/d_i^* \leq \gamma_2$, where $0 < \gamma_1 < 1 < \gamma_2$ are pre-specified, and d_i^* is the standardized basic design weights. The basic design weights d_i can be interpreted as the number of units in the population represented by unit i in the sample. Range restrictions state that the adjusted weights obtained from incorporating auxiliary information are not allowed to deviate too far away from the basic design weights. In particular, the minimum restriction of positive weights should be imposed whenever is possible.

The PEML and MCPE estimators discussed in Section 2 are asymptotically equivalent to the GREG and the weights, $\hat{p}_i$, are always positive. The weights may, however, not satisfy a more restricted range specified by $0 < \gamma_1 < 1 < \gamma_2$. Chen *et al.* [6] proposed a simple solution to this. The idea is to relax the benchmark constraints a little bit while still make good use of auxiliary information. Taking the PEML estimator as an example, if we replace $\bar{\boldsymbol{X}}$ by $\hat{\bar{\boldsymbol{X}}}_{HT} = \sum_{i \in s} d_i^* \boldsymbol{x}_i$ in the constraints (1), the resulting PEML weights would be $\hat{p}_i = d_i^*$ which will automatically satisfy any range-restrictions. In general, if we replace $\bar{\boldsymbol{X}}$ by $\bar{\boldsymbol{X}} + \delta(\hat{\bar{\boldsymbol{X}}}_{HT} - \bar{\boldsymbol{X}})$, the smallest $\delta \in (0, 1)$ can be found through a simple bi-section algorithm such that the resulting PEML weights will satisfy the pre-specified range-restriction. The algorithm is simple and guaranteed to converge. The adjustment is "optimal" in the sense of minimum relaxation of the benchmark constraints.

6 Estimation Under Measurement Error Models

In many practical situations the cost to obtain exact measurements of a study variable can be high, but "inaccurate" measurements may be gathered quite easily. Let y_i be the exact measurement and z_i be the inaccurate measurement, i.e. measurement with error.

The empirical likelihood method provides useful tools for inference under measurement error models. The general framework is to treat the distribution of the study variable non-parametrically while modeling the measurement error parametrically or semi-parametrically. Depending on the structure of the sample data, different approaches can be applied.

Two general sampling schemes are often used with measurement error problems. One scheme is to take two (or more) independent samples, a rela-

tively small sample s_1 with exact measurement and a large sample s_2 measured with error. Another popular scheme is two-phase sampling where a large first phase sample s_2 is taken and inaccurate measurement z_i are obtained and then a much smaller second phase sample s_1 is drawn and the exact y_i's are measured. An extreme case for two-phase sampling is to take s_2 as the entire finite population. Let the parameter of interest be the finite population distribution function $F_Y(t)$ and assume all samples are obtained using simple random sampling.

With two (or more) independent samples, Zhong *et al.* [23] proposed to formulate an empirical (log) likelihood function similar to the one used under stratified sampling (Zhong and Rao [22]). Let p_i be the probability mass at y_i, $i \in s_1$, and q_i be the probability mass at z_i, $i \in s_2$. The empirical (log) likelihood function is defined as

$$l(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i \in s_1} \log(p_i) + \sum_{i \in s_2} \log(q_i)\,. \tag{5}$$

Extra model information regarding the measurement error can be used as constraints when one maximizes $l(\boldsymbol{p}, \boldsymbol{q})$. For instance, the inaccurate instrument and the accurate instrument may have a common average reading, one can then obtain $\hat{p}_i$ and $\hat{q}_i$ by maximizing $l(\boldsymbol{p}, \boldsymbol{q})$ subject to

$$\sum_{i \in s_1} p_i = 1\,, \quad \sum_{i \in s_2} q_i = 1\,, \quad \text{and} \quad \sum_{i \in s_1} p_i y_i = \sum_{i \in s_2} q_i z_i\,.$$

The $\hat{p}_i$'s and the exact measurements are used to construct estimators, i.e. $\hat{F}_Y(t) = \sum_{i \in s_1} \hat{p}_i I(y_i \le t)$.

Under a two-phase sample $s_1 \subset s_2$, measurement errors can be modeled more explicitly. There are two commonly used models for measurement errors: the regression calibration model and the classical measurement error model.

The regression calibration (RC) model treats z_i as a predictor of y_i:

$$y_i = \alpha + \beta z_i + \varepsilon_i\,, \quad i = 1, 2, \cdots, N\,,$$

where the ε_i's are *i.i.d.* random variates with $E_\xi(\varepsilon_i) = 0$ and $V_\xi(\varepsilon_i) = \sigma^2$. Further model information may suggest that $\alpha = 0$ or $\beta = 1$. Including both α and β in the model can accommodate systematic bias and/or departure accompanied with the error measurements. The role of z_i in the model is the same as an auxiliary variable discussed in Sections 2, 3 and 4. Methodologies developed in these sections can be used here with a minor modification. For instance, the second constraint in (4) should be replaced by $\sum_{i \in s_1} p_i g_i = n_2^{-1} \sum_{i \in s_2} g_i$, where n_2 is the first phase sample size. The model parameters α, β and σ^2 are estimated from the second phase sample data $\{(y_i, z_i) : i \in s_1\}$.

The classical measurement error (CM) model is a conditional model for z_i given y_i:

$$z_i = \alpha + \beta y_i + \varepsilon_i\,, \quad i = 1, 2, \cdots, N\,.$$

The ε_i's are assumed *i.i.d.* $N(0, \sigma^2)$. The RC model and the CM model look similar to each other but are fundamentally different. Under the CM model, the distribution of y, which is of primary interest, appears as a marginal distribution in the condition. A mixed likelihood function needs to be used to bring together the non-parametric likelihood function of y and the conditional parametric likelihood function of z given y.

Let $f(y, z)$ be the joint distribution function of (y, z), $f(y)$ be the marginal distribution of y, and $f(z|y)$ be the conditional distribution of z given y. The full likelihood function based on (y_i, z_i), $i \in s_1$ can be written as $\prod_{i\in s_1} f(y_i, z_i) = \prod_{i\in s_1} f(y_i) \prod_{i\in s_1} f(z_i|y_i)$. Denoting $f(y_i)$ by p_i and $f(z_i|y_i)$ by $\phi(z_i, y_i, \boldsymbol{\theta})$, the log-likelihood function is given by

$$l(\boldsymbol{p}, \boldsymbol{\theta}) = \sum_{i\in s_1} \log(p_i) + \sum_{i\in s_1} \log \phi(z_i, y_i, \boldsymbol{\theta})\,.$$

This log-likelihood function is based on the small second-phase sample s_1 only. Information contained in the large first-phase sample s_2 can be used for formulating constraints. Once again the $\hat{p}_i$'s and the y_i's $(i \in s_1)$ will be used to construct estimators. The $\hat{p}_i$ and $\hat{\boldsymbol{\theta}}$ are obtained through simultaneous maximization of $l(\boldsymbol{p}, \boldsymbol{\theta})$ subject to

$$\sum_{i\in s_1} p_i = 1 \ (p_i > 0) \ \text{ and } \ \sum_{i\in s_1} p_i E(z_i^r|y_i) = E(z^r)\,.$$

The second constraint comes from $E(z^r) = E[E(z^r|y)]$ for $r = 1, 2, 3, \cdots$, and reduces to $\sum_{i\in s_1} p_i = 1$ for $r = 0$. $E(z^r)$ should be replaced by the sample moments from the large sample s_2 and $E(z_i|y_i) = \alpha + \beta y_i$, $E(z_i^2|y_i) = \sigma^2 + (\alpha + \beta y_i)^2$, etc, are obtained from the conditional normal model.

7 Concluding Remarks

It has been shown by Owen [13] that the empirical likelihood method is a powerful non-parametric approach to inference with applications in many areas of statistics for infinite populations. This article presents an overview of various applications of the method to finite population problems. The optimal MCPE estimators under non-linear situations require complete auxiliary information for implementing the method. When such information is not available, a two-phase sampling scheme may be used where the large first-phase sample of $\boldsymbol{x}$

serves as "complete" auxiliary information. Variance estimation under this scenario has been investigated by Prasad and Thach [14]. The use of empirical likelihood methods for measurement error problems needs to be further investigated. Questions that need to be addressed include: (1) How to assess the method under different models; (2) What order of moment, r, to use in the CM model; and (3) How to use the inaccurate measurement z_i, $i \in s_2$ directly in the construction of estimators. The idea of using a mixed likelihood function with one component non-parametric and the other parametric or semi-parametric might be very useful for other finite population problems.

Acknowledgments

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The author thanks Professor Randy R. Sitter and Professor Jiahua Chen for helpful comments.

References

1. D. Binder, *Internat. Statist. Rev.* **51**, 279 (1983).
2. C.M. Cassel, C.E. Särndal and J.H. Wretman, *Biometrika* **63**, 615 (1976).
3. J. Chen and J. Qin, *Biometrika* **80**, 107 (1993).
4. J. Chen and R.R. Sitter, *Statistica Sinica* **9**, 385 (1999).
5. J. Chen and C. Wu, *Statistica Sinica* (Accepted, 2002).
6. J. Chen, R.R. Sitter and C. Wu, *Biometrika* (To appear , 2002).
7. V.P. Godambe, *J. Roy. Statist. Soc. Ser. B* **17**, 267 (1955).
8. V.P. Godambe and M.E. Thompson, *Ann. Statist.* **1**, 1212 (1973).
9. V.P. Godambe and M.E. Thompson, *Internat. Statist. Rev.* **54**, 127 (1986).
10. H.O. Hartley and J.N.K. Rao, *Biometrika* **55**, 547 (1968).
11. A.B. Owen, *Biometrika* **75**, 237 (1988).
12. A.B. Owen, *Ann. Statist.* **18**, 90 (1990).
13. A.B. Owen, *Empirical likelihood* (Chapman & Hall/CRC, 2001).
14. N.G.N. Prasad and T. Thach, *Variance estimation under two-phase sampling* (Working paper, Department of Mathematical Sciences, University of Alberta, 2001).
15. J. Qin and J.F. Lawless, *Ann. Statist.* **22**, 300 (1994).
16. J. Qin and J.F. Lawless, *Canad. J Statist.* **23**, 145 (1995).
17. J.N.K. Rao, *J. Official Statistics* **10**, 153 (1994).
18. J.N.K. Rao, J.G. Kovar and H.J. Mantel, *Biometrika* **77**, 365 (1990).
19. R.R. Sitter and C. Wu, *J. Amer. Statist. Assoc.* (To appear, 2002).

20. C. Wu, *Optimal calibration estimators in survey sampling* (Working paper 2002-01, Department of Statistics and Actuarial Science, University of Waterloo, 2002).
21. C. Wu and R.R. Sitter, *J. Amer. Statist. Assoc.* **96**, 185 (2001).
22. C.X.B. Zhong and J.N.K. Rao, *Biometrika* **87**, 929 (2000).
23. B. Zhong, J. Chen and J.N.K. Rao, *Canad. J. Statist.* **28**, 841 (2000).

SECOND ORDER ESTIMATING EQUATIONS FOR CLUSTERED LONGITUDINAL BINARY DATA WITH MISSING OBSERVATIONS

GRACE Y. YI AND RICHARD J. COOK

Department of Statistics and Actuarial Science, University of Waterloo
200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1
E-mail: yyi@icarus.math.uwaterloo.ca; rjcook@lanyard.math.uwaterloo.ca

For incomplete longitudinal data Robins *et al.* [15] developed inverse probability weighted generalized estimating equations for the marginal mean parameters. In many cases, however, the repeated measurements themselves may arise in clusters, which leads to both a cross-sectional and a longitudinal correlation structure. In some applications the correlation structure may become of interest itself. In this paper we describe second order inverse probability weighted generalized estimating equations for association parameters characterizing the dependence among observations within clusters. The inverse probability weights are estimated from conditional logistic models for the missing data process. The methods are applied to data from the Waterloo Smoking Prevention Project for illustrative purposes.

1 Introduction

Generalized estimating equations (GEE) have been widely used for the analysis of data from longitudinal studies and other settings featuring clustered data (Liang and Zeger [8], Crowder [2]). This marginal approach is widely viewed as attractive because it does not require complete specification of the joint distribution of the longitudinal responses but rather is based only on specification of the means and variances of the responses. Perhaps most frequently it is of primary interest to make inferences about the parameters in regression models for the marginal means, but there has been increasing interest in association parameters in recent years. When the association parameters are of central importance second order generalized estimating equations can be constructed to facilitate their estimation. Prentice [14] developed moment-based GEE methods which model the association between a pair of binary responses in terms of the correlation, whereas Lipsitz *et al.* [10], Liang *et al.* [9] and Fitzmaurice and Lipsitz [3] modeled the association in terms of the marginal odds ratio.

Missing data occur frequently in longitudinal studies. Although studies are frequently designed to collect data on every individual in the sample at each assessment, often not all responses are observed at all occasions. For example, some individuals may withdraw from the study after a number of visits and never return. This results in a so-called monotone missing data

pattern which has been the focus of much of the literature on missing data methodology. In other settings more general missing data patterns may exist. Such patterns arise when individuals who have one or more missing visits may return to the study leading to so-called intermittently missing data.

When the data are missing completely at random (MCAR), the GEE approach can produce consistent estimates for the parameters because missing data processes are not related to the processes of generating responses. In contrast, when the data are missing at random (MAR) or nonignorable, the estimating equations are not unbiased and they fail to provide consistent estimates. Robins *et al.* [15] proposed a modified GEE approach which leads to unbiased estimating equations for the estimation of parameters involved in the marginal means. Inverse probability weights are incorporated into the estimating equations to account for the effects of the missing data processes (*e.g.* Fitzmaurice *et al.* [4]).

In this paper we focus on methods for the analysis of incomplete binary responses which arise in clusters. Clustered longitudinal binary data feature both a cross-sectional and a longitudinal correlation structure and interest often lies in the strengths of both types of association. Examples include longitudinal community intervention studies and family studies which involve repeated assessments of individual members over time. Our problem is motivated by a school-based cluster-randomized longitudinal smoking study.

We describe second order generalized estimating equations with the inverse probability weights that are used to estimate the parameters associated with both longitudinal and cross-sectional effects. The mean response is characterized by a regression model, the variance is expressed as a function of the mean, and the form of the correlation structure is assumed. The inverse probability weights are estimated from conditional logistic regression models for the missing data process. In Section 2 we introduce notation and model assumptions. In Section 3 we describe estimation of interest parameters and robust variance estimation. Section 4 presents an application to a smoking prevention study in which the data are intermittently missing and involve both a longitudinal dependence in the repeated measurements and a cross-sectional dependence arising from between school variation. General remarks are made in Section 5.

2 Notation and Model Assumptions

2.1 The Response Process

Suppose there are I clusters and J_i subjects within cluster i, $i = 1, 2, \ldots, I$. Further suppose there are K visits planned and let Y_{ijk} denote the binary response for subject j in cluster i at the visit k, $k = 1, 2, ..., K$, $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2}, ..., Y_{ijK})'$, $j = 1, 2, ..., J_i$, and $\boldsymbol{Y}_i = (\boldsymbol{Y}'_{i1}, \boldsymbol{Y}'_{i2}, ..., \boldsymbol{Y}'_{iJ_i})'$, $i = 1, 2, ..., I$. Let lower case letters y_{ijk}, $\boldsymbol{y}_{ij}$, and $\boldsymbol{y}_i$ denote the realizations of Y_{ijk}, $\boldsymbol{Y}_{ij}$, and $\boldsymbol{Y}_i$, respectively. Let $\boldsymbol{x}_{ijk} = (1, x_{ijk,1}, ..., x_{ijk,p-1})'$ be the $p \times 1$ covariate vector for subject j in cluster i at time k, $k = 1, 2, ..., K$; the covariates may be time-dependent or fixed across the entire observation times. Let $\boldsymbol{x}_{ij} = (\boldsymbol{x}'_{ij1}, \boldsymbol{x}'_{ij2}, ..., \boldsymbol{x}'_{ijK})'$, and $\boldsymbol{x}_i = (\boldsymbol{x}'_{i1}, \boldsymbol{x}'_{i2}, ..., \boldsymbol{x}'_{iJ_i})'$, $j = 1, 2, ..., J_i, i = 1, 2, ..., I$.

Let

$$\mu_{ijk} = E(Y_{ijk}|\boldsymbol{x}_i) = P(Y_{ijk} = 1|\boldsymbol{x}_i),$$

and let $\boldsymbol{\mu}_{ij} = (\mu_{ij1}, \mu_{ij2}, ..., \mu_{ijK})'$, $j = 1, 2, \ldots, J_i$, and $\boldsymbol{\mu}_i = (\boldsymbol{\mu}'_{i1}, \boldsymbol{\mu}'_{i2}, ..., \boldsymbol{\mu}'_{iJ_i})'$, $i = 1, 2, \ldots, I$. We consider the logistic regression model for the mean response

$$\text{logit } \mu_{ijk} = \boldsymbol{x}'_{ijk}\boldsymbol{\beta} \tag{1}$$

for $i = 1, 2, ..., I$, $j = 1, 2, ..., J_i$, $k = 1, 2, ..., K$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})'$ is the vector of regression parameters. The variance for the response Y_{ijk} is specified as

$$v_{ijk} = \text{var}(Y_{ijk}|\boldsymbol{x}_i) = \mu_{ijk}(1 - \mu_{ijk}),$$

which depends on the regression parameter vector $\boldsymbol{\beta}$.

The most general representation of the correlation structure is $\text{corr}(Y_{ijk}, Y_{i'j'k'}) = \phi_{jkj'k'}$ if $i = i'$ and 0 otherwise. A more specific structure could be specified as

$$\phi_{jkj'k'} = \begin{cases} \rho_{kk'}, \text{ if } j = j', \ k \neq k', \\ \gamma_k, \quad \text{if } j \neq j', \ k = k', \\ \delta_{kk'}, \text{ if } j \neq j', \ k \neq k', \end{cases}$$

with $\boldsymbol{\rho} = (\rho_{kk'}, 1 \leq k < k' \leq K)', \boldsymbol{\gamma} = (\gamma_k, 1 \leq k \leq K)'$, $\boldsymbol{\delta} = (\delta_{kk'}, 1 \leq k < k' \leq K)'$, and denoted $\boldsymbol{\phi} = (\boldsymbol{\rho}', \boldsymbol{\gamma}', \boldsymbol{\delta}')'$, a $q \times 1$ vector containing all association parameters. That is, $\boldsymbol{\rho}$ reflects the correlation among repeated measurements within subjects, $\boldsymbol{\gamma}$ represents the cross-sectional association between concurrent responses from subjects within the same cluster at each visit, and $\boldsymbol{\delta}$ represents the association between responses from subjects within

the same cluster, but for responses taken at different time points. As may be inferred from the notation, in general the association parameters pertaining to responses at different time points may be functions of the span of time between visits. One may, for example, specify an autoregressive correlation structure for $\boldsymbol{\rho}$ by introducing the constraint $\rho_{kk'} = \rho^{|k-k'|}$. A simpler exchangeable structure results by letting $\rho_{kk'} = \rho$ for $k \neq k'$. Analogous specifications for $\delta_{kk'}$ can be made. The exchangeable structure seems most appropriate for the cross-sectional association within clusters and indeed it is frequently realistic to constrain the γ_k terms to be equal to a common parameter γ.

We remark that other parametric correlation structures may be adopted to take into account constraints on the second moments (*e.g.* Oman and Zucker [13]). The following constructions of the estimating equations for the mean and association parameters may be adapted to accommodate alternative correlation structures such as these.

2.2 The Missing Data Process

Let R_{ijk} be the indicator variable taking the value 1 if the response Y_{ijk} is observed and 0 otherwise, $k = 1, 2, ..., K$, let $\boldsymbol{R}_{ij} = (R_{ij1}, R_{ij2}, ..., R_{ijK})'$, $j = 1, 2, ..., J_i$, and $\boldsymbol{R}_i = (\boldsymbol{R}'_{i1}, \boldsymbol{R}'_{i2}, ..., \boldsymbol{R}'_{iJ_i})'$, $i = 1, 2, ..., I$. Let lower letters $r_{ijk}, \boldsymbol{r}_{ij}$, and $\boldsymbol{r}_i$ denote the realizations of $R_{ijk}, \boldsymbol{R}_{ij}$, and $\boldsymbol{R}_i$, respectively. Monotone missing data patterns have been the focus of much of the work in the analysis of longitudinal incomplete data. In practice however, subjects may miss one or more visits before returning for a subsequent visit creating what is termed intermittently missing data. We shall consider arbitrary missing data patterns where $R_{ijk} = 0$ does not necessarily imply $R_{ijk'} = 0$ for $k < k'$. We let $H_{ijk} = \{y_{ij1}, y_{ij2}, \ldots, y_{ij,k-1}\}$ denote the history of the responses from subject j in cluster i up to but not including visit k, $H^{(o)}_{ijk}$ denote the history of the observed components in H_{ijk}, and $H^{r}_{ijk} = \{r_{ij1}, r_{ij2}, \ldots, r_{ij,k-1}\}$ denote the history of the missing data indicators for subject j in cluster i up to but not including visit k, $k = 1, 2, \ldots, K$, $j = 1, 2, \ldots, J_i$, and $i = 1, 2, \ldots, I$.

Three types of missing mechanism have been distinguished (Laird [7]) based on how missing data processes depend on the responses. When the missing data process is independent of all responses (both observed and unobserved) given relevant covariates and auxiliary variables, the data are said to be missing completely at random (MCAR). When the missing data process is conditionally independent of the unobserved responses given the covariates, auxiliary variables, and observed responses, the missing data are said to be missing at random (MAR). Finally the missing mechanism is called nonignorable or informative where the missing data process depends on the unobserved re-

sponses. This last type of mechanism can be addressed only by sensitivity analyses (Rotnitzky *et al.* [16]), and so we shall focus on methods for dealing with missing at random mechanism in marginal models. Moreover, we assume $P(R_{ijk} = 1|H^r_{ijk}, \boldsymbol{y}_{ij}, \boldsymbol{x}_i) = P(R_{ijk} = 1|H^r_{ijk}, H^{(o)}_{ijk}, \boldsymbol{x}_i)$. This assumption states that the probability of observing subject j in cluster i at visit k, given H^r_{ijk}, the full set of responses $\boldsymbol{y}_{ij}$ for subject j, and covariates $\boldsymbol{x}_i$, does not depend on the unobserved past responses or on the future responses. This additional assumption is not necessary when we simply restrict attention to monotone missing data patterns but facilitates the derivations that follow. As noted in Robins *et al.*, [15] this assumption implies that the data are missing at random, although missing at random does not imply such an equation. We further assume that subjects are assessed at the first visit and so $R_{ij1} = 1$, $j = 1, 2, \ldots, J_i$, $i = 1, 2, \ldots, I$.

Suppose that the conditional probability $\lambda_{ijk} = P(R_{ijk} = 1|H^r_{ijk}, H^{(o)}_{ijk}, \boldsymbol{x}_i)$ for being observed at time k is known up to a vector of unknown parameters $\boldsymbol{\alpha}_k$, $k = 2, 3, ..., K$. That is, we assume that there exists a function $\lambda_{ijk}(\boldsymbol{\alpha}_k)$ of H^r_{ijk}, $H^{(o)}_{ijk}$ and $\boldsymbol{x}_i$ such that $\lambda_{ijk} = \lambda_{ijk}(\boldsymbol{\alpha}_k)$; typically a logistic link may relate a linear function of variables summarizing H^r_{ijk}, $H^{(o)}_{ijk}$, and $\boldsymbol{x}_i$ to the probability of being observed at visit k. Let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}'_2, \boldsymbol{\alpha}'_3, ..., \boldsymbol{\alpha}'_K)'$ and define $\pi_{ij}(\boldsymbol{\alpha}) = P(\boldsymbol{R}_{ij} = \boldsymbol{r}_{ij}|\boldsymbol{y}_{ij}, \boldsymbol{x}_i) = \prod_{k=2}^{K} [\lambda_{ijk}(\boldsymbol{\alpha}_k)^{r_{ijk}}(1-\lambda_{ijk}(\boldsymbol{\alpha}_k))^{1-r_{ijk}}]$. Since every subject is observed at $k = 1$, $\pi_{ij}(\boldsymbol{\alpha})$ is the conditional probability of missingness over the entire observation period for subject j in cluster i given the vector $\boldsymbol{y}_{ij}$ and the covariates $\boldsymbol{x}_i$.

We assume that within each cluster the indicator variables are conditionally independent given the whole set of observed responses within this cluster and cluster level covariates and that the conditional probability for missingness for one subject given the whole set of responses and covariates in cluster i does not depend on the responses of other subjects in the same cluster. These assumptions enable us to write

$$P(\boldsymbol{R}_i = \boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{x}_i) = \prod_{j=1}^{J_i} P(\boldsymbol{R}_{ij} = \boldsymbol{r}_{ij}|\boldsymbol{y}_i, \boldsymbol{x}_i) = \prod_{j=1}^{J_i} P(\boldsymbol{R}_{ij} = \boldsymbol{r}_{ij}|\boldsymbol{y}_{ij}, \boldsymbol{x}_i),$$

and the MAR assumption discussed earlier leads to the partial likelihood contribution from cluster i,

$$L_i(\boldsymbol{\alpha}) = \prod_{j=1}^{J_i} \prod_{k=2}^{K} \left[\lambda_{ijk}(\boldsymbol{\alpha})^{r_{ijk}} \cdot (1 - \lambda_{ijk}(\boldsymbol{\alpha}))^{1-r_{ijk}}\right],$$

and the overall partial likelihood is $L(\boldsymbol{\alpha}) = \prod_{i=1}^{I} L_i(\boldsymbol{\alpha})$. The contribution to the partial score vector from cluster i is

$$U_{0i}(\boldsymbol{\alpha}) = \sum_{j=1}^{J_i} \sum_{k=2}^{K} (r_{ijk} - \lambda_{ijk}(\boldsymbol{\alpha})) \frac{\partial \text{logit } \lambda_{ijk}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}, \tag{2}$$

and $U_0(\boldsymbol{\alpha}) = \sum_{i=1}^{I} U_{0i}(\boldsymbol{\alpha})$, which may be solved by letting $U_0(\boldsymbol{\alpha}) = \mathbf{0}$ to give the partial maximum likelihood estimator $\hat{\boldsymbol{\alpha}}$.

3 Inference Procedures

3.1 Estimating Equations for $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ with Complete Data

Our primary interest lies in estimating parameters associated with mean responses as well as correlation between responses. In this subsection we describe how to construct the estimating equations for both $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ when the data are complete. We let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\phi}')'$.

Let $D_i = \partial \boldsymbol{\mu}_i' / \partial \boldsymbol{\beta}$ be the $p \times J_i K$ derivative matrix of the mean vector $\boldsymbol{\mu}_i$ with respect to $\boldsymbol{\beta}$, and V_i be the covariance matrix for the response $\boldsymbol{Y}_i$ for cluster i, $i = 1, 2, ..., I$. Note that V_i is comprised of sub-matrices $V_{ijj'}$ of two types. The $K \times K$ matrix V_{ijj} is the covariance matrix for the repeated measurements for subject j in cluster i where its (k, k') entry is v_{ijk} when $k = k'$ and $\sqrt{v_{ijk} v_{ijk'}} \cdot \rho_{kk'}$ when $k \neq k'$. The $K \times K$ block matrix $V_{ijj'}$ $(j \neq j')$ contains the covariance terms between the responses from subject j and subject j' in cluster i, and have as (k, k') entry $\sqrt{v_{ijk} v_{ij'k}} \cdot \gamma_k$ when $k = k'$, and $\sqrt{v_{ijk} v_{ij'k'}} \cdot \delta_{kk'}$ when $k \neq k'$.

The generalized estimating equations for $\boldsymbol{\beta}$ are given by

$$U_1(\boldsymbol{\theta}) = \sum_{i=1}^{I} U_{1i}(\boldsymbol{\theta}) = \mathbf{0}, \tag{3}$$

where $U_{1i}(\boldsymbol{\theta}) = D_i V_i^{-1} \cdot (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$.

To estimate association parameters $\boldsymbol{\phi}$, we consider pairwise products among responses within clusters. For subject j in cluster i define $\boldsymbol{Z}_{i(j)} = (Y_{ij1}Y_{ij2}, Y_{ij1}Y_{ij3}, ..., Y_{ij,K-1}Y_{ijK})'$, a vector of $K(K-1)/2$ components consisting of the pairwise products over repeated measurements of subject j. For subjects j and $j' > j$ in cluster i define $\boldsymbol{Z}_{i(j,j')} = (Y_{ij1}Y_{ij'1}, ..., Y_{ij1}Y_{ij'K}$, $Y_{ij2}Y_{ij'1}, ..., Y_{ijK}Y_{ij'K})'$, a vector of K^2 components consisting of all pairwise products of responses between subjects j and j' within cluster i. Let $\boldsymbol{Z}_i$ be the vector consisting of all vectors $\boldsymbol{Z}_{i(j)}$ and $\boldsymbol{Z}_{i(j,j')}$ defined as $\boldsymbol{Z}_i = (\boldsymbol{Z}'_{i(1)}, ..., \boldsymbol{Z}'_{i(J_i)}, \boldsymbol{Z}'_{i(1,2)}, ..., \boldsymbol{Z}'_{i(J_i-1,J_i)})'$; this is a vector of Q_i components,

where $Q_i = KJ_i(KJ_i - 1)/2$. Let $\boldsymbol{\zeta}_i$ be the expectation vector of $\boldsymbol{Z}_i$ with components determined by the relation

$$E(Y_{ijk}Y_{ij'k'}) = \phi_{jkj'k'}\sqrt{\mu_{ijk}(1-\mu_{ijk})\mu_{ij'k'}(1-\mu_{ij'k'})} + \mu_{ijk}\mu_{ij'k'},$$

and $C_i = \partial\boldsymbol{\zeta}_i'/\partial\boldsymbol{\phi}$ be the $q \times Q_i$ derivative matrix of the mean vector $\boldsymbol{\zeta}_i$ with respect to $\boldsymbol{\phi}$. Lipsitz *et al.* [10] proposed a set of unbiased estimating equations to estimate association parameters. Since the variance covariance matrix of $\boldsymbol{Z}_i$ involves third and fourth moments of the responses, they suggested a "working" covariance matrix $W_i = \text{diag}(\zeta_{i\ell}(1-\zeta_{i\ell}))$, where $\zeta_{i\ell}$ is the ℓth element of $\boldsymbol{\zeta}_i$, which gives

$$U_2(\boldsymbol{\theta}) = \sum_{i=1}^{I} U_{2i}(\boldsymbol{\theta}) = \mathbf{0}, \tag{4}$$

where $U_{2i}(\boldsymbol{\theta}) = C_i W_i^{-1} \cdot (\boldsymbol{Z}_i - \boldsymbol{\zeta}_i)$.

The efficiency of estimation of the mean parameters $\boldsymbol{\beta}$ was discussed by Sutradhar and Das [17] and Sutradhar and Kumar [18]. Their findings suggested that independence working covariance matrix performs generally well when it is not possible to specify the true covariance matrix as is the case with the second order equations in (4).

If data are incomplete, unbiased estimating equations (Godambe and Kale [6]) of the form (3) and (4) may be constructed to accommodate a variable number of assessments per subject by making a suitable modification to the covariance matrix for (3). Because the estimating equations are unbiased the resulting estimators are consistent when the data are missing completely at random, but not otherwise.

3.2 Estimating Equations for $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ with Incomplete Data

To obtain unbiased estimating equations for incomplete data with observations missing at random, we modify the estimating equations by incorporating inverse probability weights. We begin with the case that the weights are specified by fixing $\boldsymbol{\alpha}$ at $\boldsymbol{\alpha}_o$. We do this to provide insight into the sources of variation in the estimators resulting from the estimating equations with estimated weights. Let $\Delta_i(\boldsymbol{\alpha}_o) = \text{diag}(\Delta_{ij}(\boldsymbol{\alpha}_o))$ be the $J_iK \times J_iK$ block diagonal matrix, where the jth diagonal matrix $\Delta_{ij}(\boldsymbol{\alpha}_o) = \text{diag}(I(R_{ijk} = 1)/\pi_{ij}(\boldsymbol{\alpha}_o), k = 1, 2, ..., K)$ is the $K \times K$ matrix with the diagonal elements being inverse probabilities $\pi_{ij}(\boldsymbol{\alpha}_o)^{-1}$ for observed data points, or zeroes if the corresponding observations are missing; $I(\cdot)$ is the indicator function. Recall

$\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\phi}')'$ denotes the parameter of interest. The generalized estimating equations for $\boldsymbol{\beta}$ are then given by

$$U_1(\boldsymbol{\theta}, \boldsymbol{\alpha}_o) = \sum_{i=1}^{I} U_{1i}(\boldsymbol{\theta}, \boldsymbol{\alpha}_o) = \mathbf{0}, \tag{5}$$

where $U_{1i}(\boldsymbol{\theta}, \boldsymbol{\alpha}_o) = D_i V_i^{-1} \cdot \Delta_i(\boldsymbol{\alpha}_o) \cdot (\mathbf{Y}_i - \boldsymbol{\mu}_i)$.

Let $\Delta^*_{i(j)}(\boldsymbol{\alpha}_o) = \text{diag}(I(R_{ijk} = 1, R_{ijk'} = 1)/\pi_{ij}(\boldsymbol{\alpha}_o), 1 \leq k < k' \leq K)$ be the $(K(K-1)/2) \times (K(K-1)/2)$ diagonal matrix corresponding to the elements in $\boldsymbol{Z}_{i(j)}$, and $\Delta^*_{i(j,j')}(\boldsymbol{\alpha}_o) = \text{diag}(I(R_{ijk} = 1, R_{ij'k'} = 1)/(\pi_{ij}(\boldsymbol{\alpha}_o)\pi_{ij'}(\boldsymbol{\alpha}_o)), k, k' = 1, 2, ..., K)$ be the $K^2 \times K^2$ diagonal matrix corresponding to the elements in $\boldsymbol{Z}_{i(j,j')}$. Consequently the $Q_i \times Q_i$ diagonal matrix corresponding to the elements in $\boldsymbol{Z}_i$ is written as

$$\Delta^*_i(\boldsymbol{\alpha}_o) = \begin{pmatrix} \text{diag}(\Delta^*_{i(j)}(\boldsymbol{\alpha}_o)) & \mathbf{0} \\ \mathbf{0} & \text{diag}(\Delta^*_{i(j,j')}(\boldsymbol{\alpha}_o)) \end{pmatrix},$$

and the inverse probability weighted estimating equations for $\boldsymbol{\phi}$ are given by

$$U_2(\boldsymbol{\theta}, \boldsymbol{\alpha}_o) = \sum_{i=1}^{I} U_{2i}(\boldsymbol{\theta}, \boldsymbol{\alpha}_o) = \mathbf{0}, \tag{6}$$

where $U_{2i}(\boldsymbol{\theta}, \boldsymbol{\alpha}_o) = C_i W_i^{-1} \cdot \Delta^*_i(\boldsymbol{\alpha}_o) \cdot (\boldsymbol{Z}_i - \boldsymbol{\zeta}_i)$.

When $\boldsymbol{\alpha}$ is unspecified it may be estimated by solving the sum of the score equations given in (2) and (5) and (6) can be modified by replacing $\boldsymbol{\alpha}_o$ with $\hat{\boldsymbol{\alpha}}$. A Fisher-scoring algorithm may be used to obtain $\hat{\boldsymbol{\theta}}$. Specifically, for $\boldsymbol{\phi} = \boldsymbol{\phi}^{(s)}$ at the sth iteration, recursively apply

$$\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)} + [M_1(\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\phi}^{(s)}, \hat{\boldsymbol{\alpha}})]^{-1} \cdot U_1(\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\phi}^{(s)}, \hat{\boldsymbol{\alpha}}),$$

where $M_1(\boldsymbol{\theta}, \boldsymbol{\alpha}) = -\sum_{i=1}^{I} D_i V_i^{-1} \cdot \Delta_i(\boldsymbol{\alpha}) \cdot D_i'$, until $\boldsymbol{\beta}^{(t)}$ converges to $\boldsymbol{\beta}^{(s)}$, say. Then given $\boldsymbol{\beta} = \boldsymbol{\beta}^{(s)}$, recursively apply

$$\boldsymbol{\phi}^{(t)} = \boldsymbol{\phi}^{(t-1)} + [M_2(\boldsymbol{\beta}^{(s)}, \boldsymbol{\phi}^{(t-1)}, \hat{\boldsymbol{\alpha}})]^{-1} \cdot U_2(\boldsymbol{\beta}^{(s)}, \boldsymbol{\phi}^{(t-1)}, \hat{\boldsymbol{\alpha}}),$$

where $M_2(\boldsymbol{\theta}, \boldsymbol{\alpha}) = -\sum_{i=1}^{I} C_i W_i^{-1} \cdot \Delta^*_i(\boldsymbol{\alpha}) \cdot C_i'$, until it converges to $\boldsymbol{\phi}^{(s+1)}$, say. These two steps may be cycled through until convergence is achieved for $\boldsymbol{\theta}$ at $\hat{\boldsymbol{\theta}}$.

3.3 Robust Variance Estimation

Let $U_i(\boldsymbol{\theta}, \boldsymbol{\alpha}) = (U_{1i}'(\boldsymbol{\theta}, \boldsymbol{\alpha}), U_{2i}'(\boldsymbol{\theta}, \boldsymbol{\alpha}))'$ and $U(\boldsymbol{\theta}, \boldsymbol{\alpha}) = I^{-1/2} \sum_{i=1}^{I} U_i(\boldsymbol{\theta}, \boldsymbol{\alpha})$. When $\boldsymbol{\alpha}$ is specified to be $\boldsymbol{\alpha}_o$, under standard regularity conditions for estimating functions $U(\boldsymbol{\theta}, \boldsymbol{\alpha}_o)$ is asymptotically normally distributed with mean zero and covariance $E(U_i(\boldsymbol{\theta}, \boldsymbol{\alpha}_o) U_i'(\boldsymbol{\theta}, \boldsymbol{\alpha}_o))$, and $I^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed with mean zero and covariance matrix given by $\Gamma_o^{-1} E(U_i(\boldsymbol{\theta}, \boldsymbol{\alpha}_o) U_i'(\boldsymbol{\theta}, \boldsymbol{\alpha}_o))[\Gamma_o^{-1}]'$, where $\Gamma_o = E(\partial U_i(\boldsymbol{\theta}, \boldsymbol{\alpha}_o)/\partial \boldsymbol{\theta}')$. When $\boldsymbol{\alpha}$ is unspecified, the variation in the estimator $\hat{\boldsymbol{\alpha}}$ must be taken into account, and under the regularity conditions, $U(\boldsymbol{\theta}, \hat{\boldsymbol{\alpha}})$ and $I^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ are asymptotically normal with mean zero and respective asymptotic variances Σ and $\Gamma^{-1}\Sigma[\Gamma^{-1}]'$, where $\Gamma = E\left[\partial U_i(\boldsymbol{\theta}, \boldsymbol{\alpha})/\partial \boldsymbol{\theta}'\right]$, $\Sigma = E\{U_i(\boldsymbol{\theta}, \boldsymbol{\alpha}) U_i'(\boldsymbol{\theta}, \boldsymbol{\alpha})\} - E(\partial U_i(\boldsymbol{\theta}, \boldsymbol{\alpha})/\partial \boldsymbol{\alpha}') \cdot \{\mathrm{var}(U_{0i}(\boldsymbol{\alpha}))\}^{-1} \cdot \{E(\partial U_i(\boldsymbol{\theta}, \boldsymbol{\alpha})/\partial \boldsymbol{\alpha}')\}'$. We can also write $\Sigma = \mathrm{var}\{\mathrm{Resid}(U_i, U_{0i})\}$, where $\mathrm{Resid}(A_i, B_i) = A_i - E(A_i B_i')\{E(B_i B_i')\}^{-1} B_i$ is the residual from the population least squares regression of A_i on B_i. The proof is sketched in the Appendix, which appears in Robins *et al.* [15].

The matrix Σ is consistently estimated by

$$\hat{\Sigma} = I^{-1} \sum_{i=1}^{I} [\widehat{\mathrm{Resid}}(\hat{U}_i, \hat{U}_{0i})][\widehat{\mathrm{Resid}}(\hat{U}_i, \hat{U}_{0i})]',$$

where $\widehat{\mathrm{Resid}}(\hat{U}_i, \hat{U}_{0i}) = \hat{U}_i - (\sum_{t=1}^{I} \hat{U}_t \hat{U}_{0t}')(\sum_{t=1}^{I} \hat{U}_{0t} \hat{U}_{0t}')^{-1} \hat{U}_{0i}$, $\hat{U}_i = U_i(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}})$, and $\hat{U}_{0i} = U_{0i}(\hat{\boldsymbol{\alpha}})$. And the matrix Γ is consistently estimated by

$$\hat{\Gamma} = I^{-1} \begin{pmatrix} M_1(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}) & \mathbf{0} \\ M_{21}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}) & M_2(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}) \end{pmatrix},$$

where $M_{21}(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \sum_{i=1}^{I} C_i W_i^{-1} \cdot \Delta_i^*(\boldsymbol{\alpha}) \cdot (\partial \boldsymbol{\zeta}/\partial \boldsymbol{\beta}')$.

4 Application to a Smoking Prevention Study

The Waterloo Smoking Prevention Project (WSPP) consists of a series of longitudinal cluster-randomized school-based smoking prevention studies coordinated at the University of Waterloo (Cameron *et al.* [1]). We report here on the results of some analyses of data from WSPP4, the fourth study in the series.

In WSPP4 100 schools from 7 school boards in Ontario were randomized to dispense either the regular health education program provided by the school or a more intensive anti-smoking program delivered either by a specially trained teacher or a public health nurse. The program was first delivered to

children in grade 6 at the start of the study, and so-called "booster sessions" were administered to these same students when they were in grades 7 and 8. The purpose of the booster sessions was to reinforce the material presented when they were in grade 6. Attitudes towards smoking as well as actual smoking behavior were assessed annually by questionaire. A response of interest is a binary indicator of whether the child was a smoker at each of the assessment periods. For the purpose of this analysis, we define a current smoker as a student who has indicated they are a regular or occasional smoker. If we wish to use data from grades 6, 7 and 8 from all students in participating schools we have at most $K = 3$ visits for each of J_i students in school i, $i = 1, 2, \ldots, I$. The cluster randomized nature of the design suggests the need to deal with the cross-sectional within school correlation. Hence we have both the longitudinal and cross-sectional correlation structure discussed in previous sections. For illustrative purposes we consider data from 2006 students in 45 schools from WSPP4. 13.86% of the data are incomplete where 4.74% of students have no observations in both grade 7 and grade 8 and 9.12% of students have no observations either in grade 7 or in grade 8.

Let $Y_{ijk} = 1$ if student j in school i is a smoker at assessment k, and let $Y_{ijk} = 0$ otherwise. Given the relatively small number of visits we consider exchangeable correlations by letting $\rho_{kk'} = \rho$ and $\delta_{kk'} = \delta$ and a common cross-sectional correlation parameter $\gamma_k = \gamma$. To avoid the need for introducing any constraints on the parameters in the algorithm we reparameterize from ϕ to τ where

$$\tau_\rho = \log((1+\rho)/(1-\rho)), \quad \tau_\gamma = \log((1+\gamma)/(1-\gamma)), \quad \tau_\delta = \log((1+\delta)/(1-\delta)),$$

and $\boldsymbol{\tau} = (\tau_\rho, \tau_\gamma, \tau_\delta)'$.

The logistic regression model for the mean response is specified as

$$\text{logit } \mu_{ijk} = \beta_0 + \beta_1 x_{ijk1} + \beta_2 x_{ijk2} + \beta_3 x_{ijk3} + \beta_4 x_{ijk4}, \tag{7}$$

where $x_{ijk1} = 1$ if school i was randomized to the treatment arm and zero otherwise, $x_{ijk2} = 1$ if student j in school i is male and zero otherwise, $x_{ijk3} = 1$ if $k = 2$ (i.e. the response is from a grade 7 assessment) and zero otherwise, and $x_{ijk4} = 1$ if $k = 3$ (i.e. the response is from a grade 8 assessment) and zero otherwise.

The logistic regression models for the missing data process are specified in equations (8) and (9) that follow,

$$\text{logit } \lambda_{ij2} = \alpha_{20} + \alpha_{21} y_{ij1}, \tag{8}$$

and

$$\text{logit}\,\lambda_{ij3} = \alpha_{30}+\alpha_{31}y_{ij1}+\alpha_{32}r_{ij2}+\alpha_{33}r_{ij2}y_{ij2}+\alpha_{34}r^*_{ij2}+\alpha_{35}r^*_{ij2}x_{ij2}+\alpha_{36}r^*_{ij2}x_{ij3}, \tag{9}$$

where $x_{ij2} = 1$ if the school received a medium social models risk score in grade 7 and zero otherwise, $x_{ij3} = 1$ if the school received a high social models risk score in grade 7 and zero otherwise, and $r^*_{ij2} = 1$ if the social models risk score was available in grade 7 and zero otherwise.

Table 1 contains the estimates and standard errors for the regression coefficients for three logistic models for the responses. Model 1 is the full model in which ρ, γ, and δ are estimated, and Models 2 and 3 involve the constraints $\delta = 0$ and $\gamma = \delta = 0$ respectively. For each model a weighted set of generalized estimating equations of the form of (5) and (6) was solved with $\boldsymbol{\alpha}_o$ replaced with $\hat{\boldsymbol{\alpha}}$ estimated from the sum of the score equations from (2), and an unweighted set of equations was solved according to (3) and (4).

By comparing the estimates of the treatment effect arising from the weighted and unweighted estimating equations for Model 1 one can see that for several of the regression parameters the weighting has little effect on the point estimates. The estimate of the treatment coefficient however is considerably smaller when weights are used and, while not statistically significant, is not inconsistent with a modest treatment benefit. As one would expect due to the sampling variation in the estimated weights, the standard errors are considerably greater in the weighted analyses than in the unweighted analyses. The estimates of the correlation parameters are somewhat larger in the weighted analyses than the unweighted analyses. The weighted and unweighted analyses of Models 2 and 3 yield broadly consistent relationships.

5 Discussion

Second order estimating equations are used to provide estimates of association parameters simultaneously with estimates of regression coefficients for marginal means and have been advocated for use on the grounds of improved efficiency for estimation of association parameters (Liang *et al.* [9]). Interest may lie in precise estimation of the correlation coefficients to facilitate sample size calculations for example, or perhaps simply to obtain understanding about the nature of the various types of associations. We formulate estimating equations for regression coefficients in logistic models for longitudinal binary data and related association parameters for the problem in which the longitudinal series occur in clusters. We modify these first and second order

Table 1. Estimates from several weighted and unweighted models

	Model 1			
	Weighted		Unweighted	
Covariate	est.	s.e.	est.	s.e.
Intercept (β_0)	-2.553	0.308	-2.826	0.237
Treatment (β_1)	-0.409	0.312	0.007	0.214
Gender (β_2)	-0.121	0.205	-0.034	0.075
Grade 7 (β_3)	0.489	0.103	0.540	0.085
Grade 8 (β_4)	1.769	0.146	1.465	0.116
Longitudinal (τ_ρ)	0.970	0.147	0.623	0.068
Cross-sectional (τ_γ)	0.244	0.070	0.075	0.035
Mixed (τ_δ)	0.232	0.067	0.050	0.034
	Model 2 ($\delta = 0$)			
	Weighted		Unweighted	
Covariate	est.	s.e.	est.	s.e.
Intercept (β_0)	-2.701	0.258	-2.819	0.253
Treatment (β_1)	-0.158	0.265	0.008	0.236
Gender (β_2)	-0.145	0.201	-0.025	0.076
Grade 7 (β_3)	0.597	0.122	0.542	0.091
Grade 8 (β_4)	1.728	0.185	1.458	0.121
Longitudinal (τ_ρ)	0.885	0.115	0.612	0.059
Cross-sectional (τ_γ)	0.220	0.058	0.070	0.028
	Model 3 ($\gamma = \delta = 0$)			
	Weighted		Unweighted	
Covariate	est.	s.e.	est.	s.e.
Intercept (β_0)	-2.512	0.241	-2.781	0.249
Treatment (β_1)	-0.182	0.232	0.029	0.232
Gender (β_2)	-0.093	0.155	-0.013	0.075
Grade 7 (β_3)	0.442	0.079	0.495	0.071
Grade 8 (β_4)	1.686	0.134	1.421	0.109
Longitudinal (τ_ρ)	0.725	0.073	0.581	0.048

equations by the introduction of weights to deal with the possibility that the missing data are missing for reasons related to the past observed responses and the history of the missing data process (i.e. missing at random). As noted earlier, the estimating equations are unbiased and hence generate consistent estimates of the mean and association parameters. We have shown empirically via simulation studies that accurate estimates of the mean and association parameters are obtained in many practical situations. For example, when the clustered binary data are simulated using the multivariate

Plackett distribution (Molenberghs and Lesaffre [11]) under a broad range of configurations with the association characterized through odds ratios, the relative differences between the true value and the mean estimate was always less than 2.9% for regression parameters and less than 4.7% for the association parameters. These simulation studies were computed based on estimating equations described in Yi and Cook [19] which are computationally efficient and designed specifically for associations parameterized through odds ratios. We anticipate that similar findings would be observed based on equations with associations parameterized through correlations.

When applied to data from the Waterloo Smoking Prevention Project (WSPP4) we find the weighted analyses provide mild suggestive evidence of a treatment effect which is not suggested in the unweighted analysis.

We describe these methods in the context of a study in which the subjects remain in the same cluster for all longitudinal assessments. This was quite reasonable since we examined responses from students when they were in grades 6, 7 and 8. These grades were chosen in part because the intervention was applied during this time period. However, students in participating elementary schools were tracked during high school for grades 9 to 12 to enable study of the the long-term effects of the intervention. When students move from elementary schools to secondary schools it is natural to consider that they have moved from one cluster of individuals to another. One may, for example, wish to accommodate a residual long-term correlation during secondary school grades for students sharing the same elementary schools. In addition, it would seem natural to introduce another correlation parameter for cross-sectional correlations for response from students within the same secondary schools. The methods described here can easily handle this complication by introducing a slightly more general correlation structure.

In some problems it may be of interest to examine how explanatory covariates modulate the cross-sectional or longitudinal associations and for such settings Liang *et al.* [9] described how to formulate such models. One may alternatively use odds ratios to model the association between the binary responses as discussed by many authors (*e.g.* Fitzmaurice *et al.* [3]), and introduce the dependence on the covariates by regression models.

The weighted estimating equations described in Section 3 are conditional on the specification of parameters that index the missing data processes. In Section 4 we propose logistic regression models to characterize the presence of responses via their past history of presence and observed outcomes, however unlike in the marginal models for the response, no cross-sectional correlation is addressed. One could include cluster effects for the drop-out process however, by modeling within cluster dependences with odds ratios and using the

Plackett distribution (Molenberghs and Lesaffre [11]). The estimation of the parameters involved in the missing data processes can be carried out using the GEE2 approach (Liang *et al.* [9]).

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada and the Canadian Institutes for Health Research. The authors would like to thank two anonymous referees whose comments greatly improved the form of this article. The authors thank Prof. K.S. Brown for providing the data from the Waterloo Smoking Prevention Project and Ms. Ker-Ai Lee for programming assistance. R.J. Cook is an Investigator for the Canadian Institutes for Health Research.

Appendix

Let $\boldsymbol{H}_i(\boldsymbol{\theta},\boldsymbol{\alpha}) = (\boldsymbol{U}_i'(\boldsymbol{\theta},\boldsymbol{\alpha}), \boldsymbol{U}_{0i}'(\boldsymbol{\alpha}))'$, then $E(\boldsymbol{H}_i(\boldsymbol{\theta},\boldsymbol{\alpha})) = \mathbf{0}$. If $\boldsymbol{H}_i(\boldsymbol{\theta},\boldsymbol{\alpha})$ further satisfies regularity conditions stated in Appendix A of Robins *et al.* [15], then by Theorem 3.4 of Newey and McFadden [12] with probability approaching 1, there is a unique solution, say $(\hat{\boldsymbol{\theta}}', \hat{\boldsymbol{\alpha}}')'$, to $\sum_{i=1}^{I} \boldsymbol{H}_i(\boldsymbol{\theta},\boldsymbol{\alpha}) = \mathbf{0}$, and it satisfies

$$I^{1/2}\begin{pmatrix}\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\\ \hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\end{pmatrix} = -E[\partial \boldsymbol{H}_i(\boldsymbol{\theta},\boldsymbol{\alpha})/\partial(\boldsymbol{\theta}',\boldsymbol{\alpha}')]^{-1}\cdot I^{-1/2}\sum_{i=1}^{I}\boldsymbol{H}_i(\boldsymbol{\theta},\boldsymbol{\alpha}) + o_p(\mathbf{1}).$$

Therefore, we obtain

$$\begin{aligned} I^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) &= -I^{-1/2}\{E(\partial \boldsymbol{U}_i(\boldsymbol{\theta},\boldsymbol{\alpha})/\partial\boldsymbol{\theta}')^{-1}\cdot\sum_{i=1}^{I}\boldsymbol{U}_i(\boldsymbol{\theta},\boldsymbol{\alpha}) - E(\partial \boldsymbol{U}_i(\boldsymbol{\theta},\boldsymbol{\alpha})/\partial\boldsymbol{\theta}')^{-1}\\ &\quad\cdot E(\partial \boldsymbol{U}_i(\boldsymbol{\theta},\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}')\cdot[E(\partial \boldsymbol{U}_{0i}(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}')]^{-1}\cdot\sum_{i=1}^{I}\boldsymbol{U}_{0i}(\boldsymbol{\alpha})\} + o_p(\mathbf{1})\\ &= -\boldsymbol{\Gamma}^{-1}I^{-1/2}\cdot\sum_{i=1}^{I}\mathrm{Resid}(U_i, U_{0i}) + o_p(\mathbf{1}), \end{aligned}$$

where $U_i = U_i(\boldsymbol{\theta},\boldsymbol{\alpha})$, $U_{0i} = U_{0i}(\boldsymbol{\alpha})$, $U_0 = I^{-1/2}\sum_i U_{0i}(\boldsymbol{\alpha})$. The central limit theorem then leads to the asymptotic distribution for $I^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$.

In deriving the last equation, we used the identities $E(\partial U_{0i}/\partial\boldsymbol{\alpha}') = -E(U_{0i}U_{0i}')$ and $E(\partial U_i/\partial\boldsymbol{\alpha}') = -E(U_iU_{0i}')$. The latter one is obtained by differentiating $E\{U_i(\boldsymbol{\theta},\boldsymbol{\alpha})\} = 0$ with respect to $\boldsymbol{\alpha}$ under the integral sign.

References

1. R. Cameron, K.S. Brown, J.A. Best, C.L. Pelkman, C.L. Madill, S.R. Manske, and M.E. Payne, *Am. J. Public Health* **89**, 1827 (1999).
2. M. Crowder, *Biometrika* **82**, 407 (1995).
3. G.M. Fitzmaurice and S.R. Lipsitz, *Appl. Statist.* **44**, 51 (1995).
4. G.M. Fitzmaurice, G.M. Molenberghs and S.R. Lipsitz, *J.R. Statist. Soc. B* **57**, 691 (1995).
5. G.M. Fitzmaurice, N.M. Laird and G.E.P. Zahner, *J. Am. Statist. Ass.* **91**, 99 (1996).
6. V.P. Godambe and B.K. Kale, in *Estimating Functions*, ed. V. P. Godambe (Oxford University Press, London, 1991).
7. N.M. Laird, *Statist. Med.* **7**, 305 (1988).
8. K.-Y. Liang and S.L. Zeger, *Biometrika* **73**, 13 (1986).
9. K.-Y. Liang, S.L. Zeger and B. Qaqish, *J.R. Statist. Soc. B* **54**, 3 (1992).
10. S.R. Lipsitz, N.M. Laird and D.P. Harrington, *Biometrika* **78**, 153 (1991).
11. G. Molenberghs and E. Lesaffre, *J. Am. Statist. Ass.* **89**, 633 (1994).
12. W. K. Newey and D. McFadden, in *Handbook of Econometrics*, Vol. 4, eds. D. McFadden and R. Engler (North-Holland, Amsterdam, 1993).
13. S.D. Oman and D.M. Zucker, *Biometrika* **88**, 287 (2001).
14. R.L. Prentice, *Biometrics* **44**, 1033 (1988).
15. J.M. Robins, A. Rotnitzky and L.P. Zhao, *J. Am. Statist. Ass.* **90**, 106 (1995).
16. A. Rotnitzky, J.M. Robins and D.O. Scharfstein, *J. Am. Statist. Ass.* **93**, 1321 (1998).
17. B.C. Sutradhar and K. Das, *Biometrika* **86**, 459 (1999).
18. B.C. Sutradhar and P. Kumar, *Stat. and Prob. Letters* **55**, 53 (2001).
19. G.Y. Yi and R.J. Cook, *Marginal methods for incomplete longitudinal data arising in clusters* (Submitted for publication, 2002).

RECENT ADVANCES IN STATISTICAL METHODS

This volume consists of research papers dealing with computational and methodological issues of statistical methods on the cutting edge of modern science. It touches on many applied fields such as Bayesian Methods, Biostatistics, Econometrics, Finite Population Sampling, Genomics, Linear and Nonlinear Models, Networks and Queues, Survival Analysis, Time Series, and many more.

P270 hc
ISBN 1-86094-333-0
9 781860 943331

Imperial College Press

www.icpress.co.uk