Studies in Theoretical and Applied Statistics
Selected Papers of the Statistical Societies

Tonio Di Battista
Elías Moreno
Walter Racugno *Editors*

Topics on Methodological and Applied Statistical Inference

Springer

Studies in Theoretical and Applied Statistics

Selected Papers of the Statistical Societies

More information about this series at http://www.springer.com/series/10107

Tonio Di Battista · Elías Moreno
Walter Racugno
Editors

Topics on Methodological and Applied Statistical Inference

Springer

Editors
Tonio Di Battista
DISFPEQ
"G. d'Annunzio" University
of Chieti-Pescara
Pescara
Italy

Walter Racugno
Department of Mathematics
University of Cagliari
Cagliari
Italy

Elías Moreno
Statistics and Operations Research
University of Granada
Granada
Spain

ISSN 2194-7767 ISSN 2194-7775 (electronic)
Studies in Theoretical and Applied Statistics
ISBN 978-3-319-44092-7 ISBN 978-3-319-44093-4 (eBook)
DOI 10.1007/978-3-319-44093-4

Library of Congress Control Number: 2016948792

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Dear reader,

On behalf of the four Scientific Statistical Societies—the SEIO, Sociedad de Estadística e Investigación Operativa (Spanish Society of Statistics and Operations Research); SFdS, Société Française de Statistique (French Statistical Society); SIS, Società Italiana di Statistica (Italian Statistical Society); and the SPE, Sociedade Portuguesa de Estatística (Portuguese Statistical Society)—we would like to inform you that this is a new book series of Springer entitled *Studies in Theoretical and Applied Statistics*, with two lines of books published in the series: *Advanced Studies and Selected Papers of the Statistical Societies*.

The first line of books offers constant up-to-date information on the most recent developments and methods in the fields of theoretical statistics, applied statistics, and demography. Books in this series are solicited in constant cooperation between the statistical societies and need to show a high-level authorship formed by a team preferably from different groups so as to integrate different research perspectives.

The second line of books presents a fully peer-reviewed selection of papers on specific relevant topics organized by the editors, also on the occasion of conferences, to show their research directions and developments in important topics, quickly and informally, but with a high level of quality. The explicit aim is to summarize and communicate current knowledge in an accessible way. This line of books will not include conference proceedings and will strive to become a premier communication medium in the scientific statistical community by receiving an Impact Factor, as have other book series such as *Lecture Notes in Mathematics*.

The volumes of selected papers from the statistical societies will cover a broad range of theoretical, methodological as well as application-oriented articles, surveys and discussions. A major goal is to show the intensive interplay between various, seemingly unrelated domains and to foster the cooperation between scientists in different fields by offering well-founded and innovative solutions to urgent practice-related problems.

On behalf of the founding statistical societies I wish to thank Springer, Heidelberg and in particular Dr. Martina Bihn for the help and constant cooperation in the organization of this new and innovative book series.

Rome, Italy — Maurizio Vichi

Preface

This volume contains a selection of the contributions presented in the 47th Scientific Meeting of the Italian Statistical Society, held at the University of Cagliari, Italy, June 2014.

The book represents a small but interesting sample of 19 out of 221 papers discussed in the meeting on a variety of methodological and applied statistical topics. Clustering, collaboration networks analysis, environmental analysis, logistic regression, mediation analysis, meta-analysis, outliers in time-series and regression, pseudolikelihood, sample design, weighted regression, are themes included in the book.

We hope that the overview papers, mainly presented by Italian authors, will help the reader to understand the state of art of the current international research.

Pescara, Italy — Tonio Di Battista
Granada, Spain — Elías Moreno
Cagliari, Italy — Walter Racugno

Contents

Introducing Prior Information into the Forward Search for Regression

Anthony C. Atkinson, Aldo Corbellini and Marco Riani

Abstract

The forward search provides a flexible and informative form of robust regression. We describe the introduction of prior information into the regression model used in the search through the device of fictitious observations. The extension to the forward search is not entirely straightforward, requiring weighted regression. Forward plots are used to exhibit the effect of correct and incorrect prior information on inferences.

1 Introduction

Methods of robust regression have been described in several books, for example [2,6,14]. The recent comparisons of [12] indicate the superior performance of the forward search (FS) in a wide range of conditions. However, none of these methods includes prior information; they can all be thought of as developments of least squares. The purpose of the present paper is to show how prior information can be

A.C. Atkinson (✉)
Department of Statistics, London School of Economics, London, UK
e-mail: a.c.atkinson@lse.ac.uk

A. Corbellini · M. Riani
Dipartimento di Economia, Università di Parma, Parma, Italy
e-mail: aldo.corbellini@unipr.it

M. Riani
e-mail: mriani@unipr.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_1

incorporated into FS for regression and to give some results indicating the comparative performance of this Bayesian method.

In order to detect outliers and departures from the fitted regression model in the absence of prior information, the FS uses least squares to fit the model to subsets of m observations, starting from an initial subset of m_0 observations. The subset is increased from size m to $m+1$ by forming the new subset from the observations with the $m+1$ smallest squared residuals. For each m ($m_0 \leq m \leq n-1$), we test for the presence of outliers, using the observation outside the subset with the smallest absolute deletion residual.

The specification of prior information and its incorporation into the FS is derived in Sect. 2. Section 3 presents the algebraic details of outlier detection with prior information. Forward plots in Sect. 4 show the dependence of the evolution of parameter estimates on prior values of the parameters. In the rest of the paper the emphasis is on forward plots of minimum deletion residuals which form the basis for outlier detection. These plots are presented in Sect. 4 for correctly specified priors and, in Sect. 4, for incorrect specifications. It is argued that use of analytically derivable frequentist envelopes is also suitable for Bayesian outlier detection when the priors are correctly specified. However, serious errors can occur with misspecified priors.

2 Prior Information in the Linear Model from Fictitious Observations

In the regression model without prior information $y = X\beta + \varepsilon$, y is the $n \times 1$ vector of responses, X is an $n \times p$ full-rank matrix of known constants, with ith row x_i^T, and β is a vector of p unknown parameters. The normal theory assumptions are that the errors ε_i are i.i.d. $N(0, \sigma^2)$.

In some of the applications in which we are interested, for example fraud detection [7], we have appreciable prior information about the values of the parameters. This can often conveniently be thought of as coming from n_0 fictitious observations y_0 with matrix of explanatory variables X_0. Then the data consist of the n_0 fictitious observations plus n actual observations. The search in this case now proceeds from $m = 0$, when the fictitious observations provide the parameter values for all n residuals from the data; the fictitious observations are always included in those used for fitting, their residuals being ignored in the selection of successive subsets.

There is one complication in combining this procedure with the forward search, which arises from the estimation of variance from subsets of observations. If we estimate σ^2 from all n observations, we obtain an unbiased estimate of σ^2 from the residual sum of squares. However, in the frequentist search we select the central m out of n observations to provide the mean square estimate $s^2(m)$, so that the variability is underestimated. To allow for estimation from this truncated distribution, let the variance of the symmetrically truncated normal distribution containing the central m/n portion of the full distribution be $\sigma_T^2(m)$. See [10] for a derivation from the general method of [15]. We take as our approximately unbiased estimate of variance

$s_T^2 = s^2(m)/\sigma_T^2 = s^2(m)/c(m,n)$. In the robustness literature $c(m,n)$ is called a consistency factor [5,13].

In the Bayesian procedure, the n_0 fictitious observations are treated as a sample with variance σ^2. However, the m observations from the actual data come from a truncated distribution with variance $c(m,n)\sigma^2$, which must be adjusted before the two samples are combined. This becomes a standard problem in weighted least squares (for example, [9, p. 230]). Let y^+ be the $(n_0 + m) \times 1$ vector of responses from the fictitious observations and the subset and let the covariance matrix of these observations be $\sigma^2 G$, with G a diagonal matrix. Then the first n_0 elements of the diagonal of G equal one and the last m elements have the value $c(m,n)$. In the least squares calculations we need only to multiply the elements of the sample values of y and X by $c(m,n)^{-1/2}$. The residual mean square error from this weighted regression provides the estimate $\hat{\sigma}^2(m)$.

The prior information can also be specified in terms of prior distributions of the parameters β and σ^2. The details and relationship with fictitious observations are given by [4] as part of a study of Bayesian methods for outlier detection and by [3] in the context of the forward search.

3 Algebra for the Bayesian Forward Search

Let $S^*(m)$ be the subset of size m found by FS, for which the matrix of regressors is $X(m)$. Weighted least squares on this subset of observations plus X_0 yields parameter estimates $\hat{\beta}(m)$ and $\hat{\sigma}^2(m)$, the latter on $n_0 + m - p$ degrees of freedom. Residuals can be calculated for all n observations including those not in $S^*(m)$. The n resulting least squares residuals are $e_i(m) = y_i - x_i^{\mathrm{T}}\hat{\beta}(m), \quad (i = 1, \ldots, n)$.

The search moves forward with the augmented subset $S^*(m+1)$ consisting of the observations with the $m+1$ smallest absolute values of $e_i(m)$. To start we take $m_0 = 0$, since the prior information specifies the values of β and σ^2.

To test for outliers the deletion residuals are calculated for the $n - m$ observations not in $S^*(m)$. These residuals are

$$r_i(m) = e_i(m)/[\hat{\sigma}^2(m)\{1 + h_i(m)\}]^{0.5}, \tag{1}$$

where the leverage $h_i(m) = x_i^{\mathrm{T}}\{X_0^{\mathrm{T}}X_0 + X(m)^{\mathrm{T}}X(m)/c(m,n)\}^{-1}x_i$. Let the observation nearest to those forming $S^*(m)$ be $i_{\min} = \arg\min_{i \notin S^*(m)} |r_i(m)|$. To test whether observation $i_{\min}$ is an outlier we use the absolute value of the minimum deletion residual

$$r_{\mathrm{imin}}(m) = e_{\mathrm{imin}}(m)/[\hat{\sigma}^2(m)\{1 + h_{\mathrm{imin}}(m)\}]^{0.5}, \tag{2}$$

as a test statistic. If the absolute value of (2) is too large, the observation $i_{\min}$ is considered to be an outlier, as well as all other observations not in $S^*(m)$.

4 Example 1: Correct Prior Information

To explore the properties of FS including prior information, we use simulation to provide forward plots of the distribution of quantities of interest during the search. These simulations are intended to complement the analysis of [3] based on the Windsor housing data introduced by [1]. In these data there are 546 observations on regression data with four explanatory variables and an intercept, so that $p = 5$. Because of the invariance of least squares results to the values of the parameters in the regression model, we simulated the responses as independent standard normal variables with all regression coefficients equal to zero. The explanatory variables were likewise independent standard normal, simulated once for each set of simulations, as were the fictitious observations providing the prior. We took $n = 500$ in all simulations reported here and repeated the simulations 10,000 times.

Figure 1 shows forward plots of the parameter estimates when there is relatively weak prior information ($n_0 = 30$). Because of the symmetry of our simulations in the coefficients β_j, the left-hand panel arbitrarily shows the evolution of $\hat{\beta}_3$. From the simulations all other linear parameters give indistinguishable plots. The plot is centred around the simulation value of zero with quantiles that decrease steadily and smoothly with m. The right-hand panel is more surprising: the estimate of σ^2 decreases rapidly from the prior value of one, reaching a minimum value of 0.73 before gradually returning to one. The effect is due to the value of the asymptotic correction factor $c(m, n)$ which is too large. Further correction is needed in finite samples. Reference [8] use simulation to make such corrections in robust regression, but not for the FS.

The differing widths of bands in the two panels serve as a reminder of the comparative variability of estimates of variance. Reference [3] give the plot for stronger prior information when $n_0 = 500$. With equal amounts of prior and sample information at the end of the search, the bands for $\hat{\beta}_3$ are appreciably more horizontal than those of Fig. 1. However, the larger effect of increased prior information is in estimation

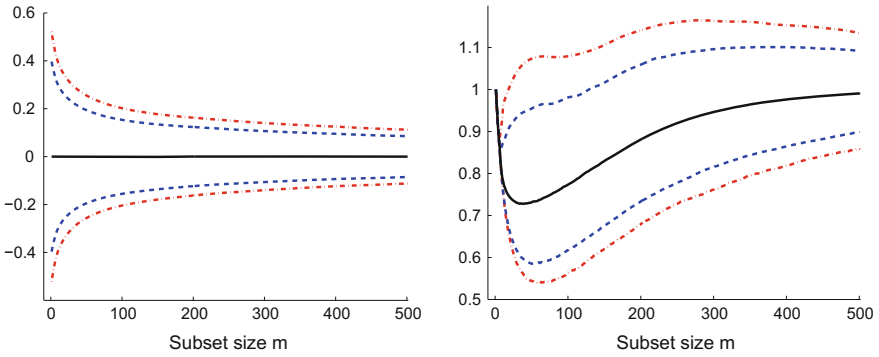


Fig. 1 Distribution of parameter estimates when $\beta_3 = 0$ and $\sigma^2 = 1$. *Left-hand panel* $\hat{\beta}_3$, *right-hand panel* $\hat{\sigma}^2$; weak prior information ($n_0 = 30$; $n = 500$). 1, 5, 50, 95 and 99 % empirical quantiles

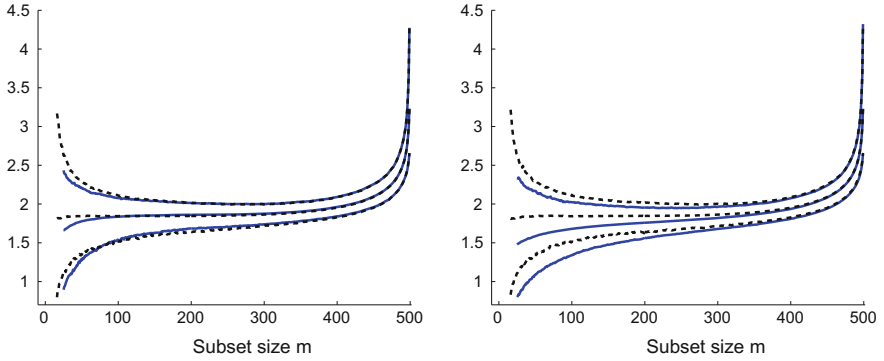


Fig. 2 The effect of correct prior information on forward plots of minimum deletion residuals. *Left-hand panel*, weak prior information ($n_0 = 30$; $n = 500$). *Right-hand panel*, strong prior information ($n_0 = 500$; $n = 500$), 10,000 simulations; 1, 50 and 99 % empirical quantiles. *Dashed lines*, without prior information; *heavy lines*, with prior information

of σ^2, which now has a minimum value of 0.97 and appreciably narrower bands for the quantiles.

The parameter estimates form an important component of the forward plots of minimum deletion residuals. The plots of these residuals, which are the focus of the rest of this paper, are the central tool for the detection of outliers in the FS. Outliers are detected when the curve for the sample values falls outside a specified envelope. The actual rule for detection of an outlier has to take account of the multiple testing inherent in the FS (once for each value of m). One rule, yielding powerful tests of the desired 1 % size, is given by [10] for multivariate data and by [11] for regression. The procedure has two stages, in the second of which envelopes are required for a series if values of n. The left-hand panel of Fig. 2 shows the envelopes for weak prior information ($n_0 = 30$), together with those from the FS in the absence of prior information. Unlike the Bayesian envelopes, those for the frequentist search are found by arguments based on the properties of order statistics. In this panel the frequentist and Bayesian envelopes agree for all except sample sizes around 100 or less. In the right-hand panel the prior information is stronger, with $n_0 = 500$. The upper envelopes for procedures with and without prior information agree for the second half of the search. For the 1 and 50 % quantiles the values of the statistics in the absence of prior information are higher than those in its presence, reflecting the increased prevalence of smaller estimates of σ^2 in the frequentist search. In general, the agreement in distribution of the statistics is not of central importance, since the envelopes apply to different situations. One important, although expected, outcome is the increase in power of the outlier tests that comes from including prior information, which is quantified by [3]. Also important is the agreement of frequentist and Bayesian envelopes towards the end of the search, which is where outlier detection usually occurs. This agreement allows us to use the frequentist envelopes when testing for outliers in the presence of prior information. Such envelopes can

be calculated analytically, avoiding the time consuming simulations that are needed when envelopes for different values of n are required.

5 Example 2: Incorrect Prior Information

In the housing data analysed by [3], there is evidence of incorrect specification of the prior values of some parameters. The effect of misspecification of σ^2 is easily described; estimates of β remain unbiased, although with a changed variance compared with those when the specification is correct. The estimate of σ^2 also behaves in a smooth fashion; initially close to the prior value it moves steadily towards the sample value.

The effect of misspecification of β is more complicated since both $\hat{\beta}$ and $\hat{\sigma}^2$ are affected. There are two effects. The effect on $\hat{\beta}$ is to yield an estimate that moves from the prior value to the sample value in a sigmoid manner. Because of the biased nature of $\hat{\beta}$, the residual sum of squares is too large and $\hat{\sigma}^2$ rapidly moves away from its correct prior value. As sample evidence increases the estimate gradually stabilises and then moves towards the sample value. There are then two conflicting effects on the deletion residuals; an increase due to incorrect values of β and a reduction in the residuals due to overestimation of σ^2. Plots illustrating these effects on the parameter estimates are given by [3]. Here we show the effect of misspecification of β on envelopes like those of Fig. 2.

Our interpretation of Fig. 2 was that the frequentist envelopes could be used for outlier identification with little change of size or loss of power in the outlier test compared with use of the envelopes for the correctly specified prior. We focus on this aspect in interpreting the envelopes from an incorrectly specified prior.

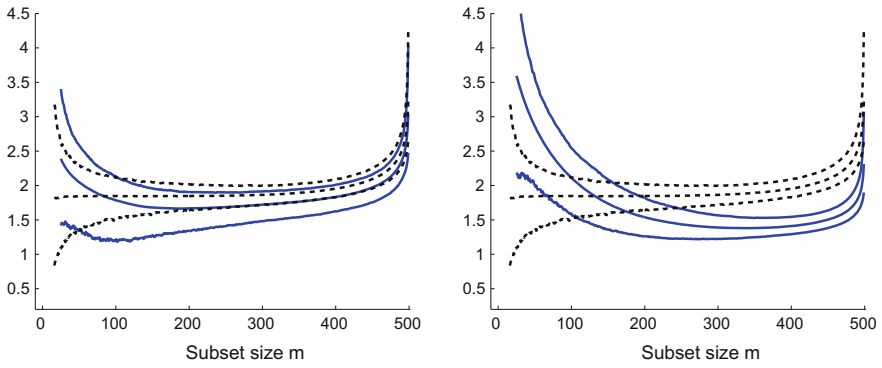


Fig. 3 The effect of incorrect prior information on forward plots of minimum deletion residuals; $\beta_0 = 1.5$. *Left-hand panel*, $n_0 = 6$, *right-hand panel*, $n_0 = 100$, 10,000 simulations; 1, 50 and 99 % empirical quantiles. *Dashed lines*, without prior information; *heavy lines*, with prior information

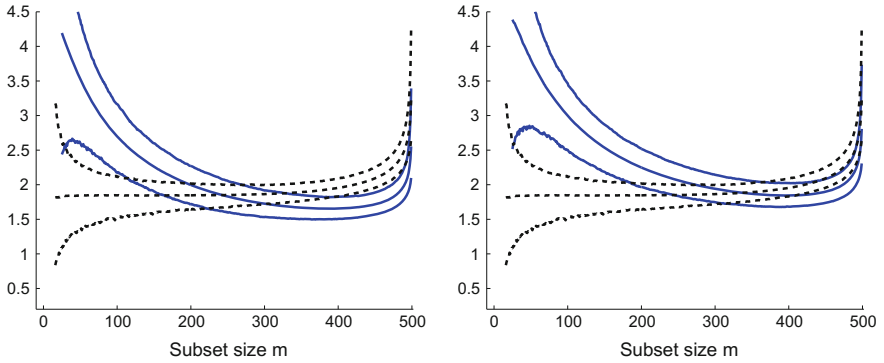


Fig. 4 The effect of increased incorrect prior information on forward plots of minimum deletion residuals; $\beta_0 = 1.5$. *Left-hand panel*, $n_0 = 250$, *right-hand panel*, $n_0 = 350$, 10,000 simulations; 1, 50 and 99 % empirical quantiles. *Dashed lines*, without prior information; *heavy lines*, with prior information

In the simulations all values of β were incremented by 1.5. In the left-hand panel of Fig. 3 we take $n_0 = 6$. Initially the envelopes lie above the frequentist bands, with a longer lower tail. Interest in outlier detection is in the latter half of the envelopes, for which the true envelopes lie below the frequentist ones; the residuals tend to be smaller and outliers would be less likely to be detected even at the very end of the search. In the right-hand panel, n_0 has been increased to 100. The result is to increase the size of the residuals at the beginning of the search. However, in the second half, the correct envelopes for this prior lie well below the frequentist envelopes; although outliers would be even less likely to be detected than before, the series of residuals lying well below the envelope would suggest a mismatch between prior and data.

Figure 4 shows two further forward plots of envelopes of minimum deletion residuals but now with greater prior information. In the left-hand panel $n_0 = 250$ and in the right-hand panel the value is 350. The trend follows that first seen in the right-hand panel of Fig. 3. In the first half of the search the envelopes continue to rise above the frequentist bands—very large residuals are likely at this early stage, which will provide a signal of prior misspecification. However, now, the envelopes for the right-hand halves of the searches are coming closer together. Particularly for $n_0 = 350$, there are unlikely to be a large number of residuals lying below the frequentist bands, although outliers will still have residuals that are less evident than they would be using the correct envelope.

This discussion suggests that forward plots of deletion residuals can provide one way of detecting a misspecification of the prior distribution. Similar runs of too small residuals can also be a sign of other model misspecification; they can occur, for example, in the frequentist analysis of data with beta distributed errors under

the assumption of normal errors. The analysis of the housing data presented by [3] provides examples of the effect of prior misspecification on forward plots of minimum deletion residuals.

References

1. Anglin, P., Gençay, R.: Semiparametric estimation of a hedonic price function. J. Appl. Econ. **11**, 633–648 (1996)
2. Atkinson, A.C., Riani, M.: Robust Diagnostic Regression Analysis. Springer, New York (2000)
3. Atkinson, A.C., Corbellini, A., Riani, M.: Robust Bayesian regression. Submitted (2016)
4. Chaloner, K., Brant, R.: A Bayesian approach to outlier detection and residual analysis. Biometrika **75**, 651–659 (1998)
5. Johansen, S., Nielsen, B.: Analysis of the Forward Search using some new results for martingales and empirical processes. Bernoulli **22** (2016, in press)
6. Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust Statistics: Theory and Methods. Wiley, Chichester (2006)
7. Perrotta, D., Torti, F.: Detecting price outliers in European trade data with the forward search. In: Palumbo, F., Lauro, C.N., Greenacre, M.J. (eds.) Data Analysis and Classification. Springer, Heidelberg (2010)
8. Pison, G., Van Aelst, S., Willems, G.: Small sample corrections for LTS and MCD. Metrika **55**, 111–123 (2002)
9. Rao, C.R.: Linear Statistical Inference and its Applications, 2nd edn. Wiley, New York (1973)
10. Riani, M., Atkinson, A.C., Cerioli, A.: Finding an unknown number of multivariate outliers. J. R. Stat. Soc., Ser. B **71**, 447–466 (2009)
11. Riani, M., Cerioli, A., Atkinson, A.C., Perrotta, D.: Monitoring robust regression. Electron. J. Stat. **8**, 646–677 (2014)
12. Riani, M., Atkinson, A.C., Perrotta, D.: A parametric framework for the comparison of methods of very robust regression. Stat. Sci. **29**, 128–143 (2014)
13. Riani, M., Cerioli, A., Torti, F.: On consistency factors and efficiency of robust S-estimators. TEST **23**, 356–387 (2014)
14. Rousseeuw, P.J., Leroy, A.M.: Robust Regression and Outlier Detection. Wiley, New York (1987)
15. Tallis, G.M.: Elliptical and radial truncation in normal samples. Ann. Math. Stat. **34**, 940–944 (1963)

A Finite Mixture Latent Trajectory Model for Hirings and Separations in the Labor Market

Silvia Bacci, Francesco Bartolucci, Claudia Pigini and Marcello Signorelli

Abstract

We propose a finite mixture latent trajectory model to study the behavior of firms in terms of open-ended employment contracts that are activated and terminated during a certain period. The model is based on the assumption that the population of firms is composed by unobservable clusters (or latent classes) with a homogeneous time trend in the number of hirings and separations. Our proposal also accounts for the presence of informative drop-out due to the exit of a firm from the market. Parameter estimation is based on the maximum likelihood method, which is efficiently performed through an EM algorithm. The model is applied to data coming from the Compulsory Communication dataset of the local labor office of the province of Perugia (Italy) for the period 2009–2012. The application reveals the presence of six latent classes of firms.

S. Bacci (✉) · F. Bartolucci · C. Pigini · M. Signorelli
Department of Economics, University of Perugia, Perugia, Italy
e-mail: silvia.bacci@unipg.it

F. Bartolucci
e-mail: francesco.bartolucci@unipg.it

C. Pigini
e-mail: pigini@stat.unipg.it

M. Signorelli
e-mail: marcello.signorelli@unipg.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_2

1 Introduction

Recent reforms of the Italian labor market [4] have shaped a prevailing dual system where, on the one side, workers with an open-ended contract benefit from a high degree of job security (especially in firms with more than 15 employees) and, on the other, temporary workers are exposed to a low degree of employment protection. Several policy interventions have been carried out with the purpose of improving the labor market performance and productivity outcomes. The effects of employment protection legislation in Italy have been investigated mainly with respect to firms' growth and to the incidence of small firms. The empirical evidence points toward a mild effect of these policies on firms' growth: Schivardi and Torrini [10] state that firms avoid the costs of highly protected employment by substituting permanent employees with temporary workers; Hijzen, Mondauto, and Scarpetta [4] find that employment protection has a sizable impact on the incidence of temporary employment. In this context, the analysis of open-ended employment turnover may shed some light on whether the use of highly protected contracts has declined especially in relation to the recent economic crisis.

In order to analyze the problem at issue, we use data from the Compulsory Communication (CC) database of the labor office of the province of Perugia (Italy) in the period 2009–2012, and we introduce a latent trajectory model based on a finite mixture of logit and log-linear regression models. A logit regression model is specified to account for the informative drop-out due to the exit of a firm from the market in a certain time window, mainly due to bankruptcy, closure of the activity, or termination. Besides, conditionally on the presence of a firm in the market, two log-linear regression models are defined for the number of open-ended hirings and separations observed at every time window. Finally, we assume that firms are clustered in a given number of latent classes that are homogeneous with respect to the behavior of firms in terms of open-ended hirings and separations, other than in terms of probability of exit from the market. Alternatively to the proposed approach, a more traditional one to deal with longitudinal data consists in adopting a generalized linear mixed model with continuous (usually normal) random effects. However, such a solution does not allow to classify firms in homogenous classes, other than having several problems related to the maximum likelihood estimation process and to the possible misspecification of the distribution of the random effects.

The paper is organized as follows. In Sect. 2 we describe the CC data coming from the local labor office of Perugia. In Sect. 3 we first illustrate the model assumptions and, then, we describe the main aspects related to the model estimation and to the selection of the number of latent classes. In Sect. 4 we apply the proposed model to the data at issue. Finally, we conclude the work with some remarks.

2 Data

The CC database is an Italian administrative longitudinal archive consisting of data collected by the Ministry of labor, health, and social policies through local labor offices. With the ministerial decrees n. 181 and n. 296, since 2008 Italian firms and Public Administrations (PAs) are required to transmit a telematic communication for each hiring, prolongation, transformation, or separation (i.e., firing, dismissal, retirement) to the qualified local labor office. In particular, we dispose of all communications from January 2009 to December 2012 sent by firms and PAs operating in the province of Perugia. The dataset, provided by the local labor office of Perugia, contains information on the single contracts as well as the workers concerned by each communication and the firms/PAs transmitting the record.

The single CC represents the unit of observation for a total of 937,123 records. In order to avoid a possible distortion due to new-born firms in the period 2009–2012, we consider only firms/PAs that sent at least one communication in the first quarter of 2009 and those communicating separations of contracts that started before 2009. Once these firms have been selected, we end up with 34,357 firms/PAs in our dataset. Note that if firms/PAs do not send any record between 2009 and 2012 they do not appear in the dataset. The number of firms and PAs entering the dataset in each quarter is reported in the first column of Table 1. In addition, firms exiting the market must be accounted for: relying on the information about the reasons of the communicated separations, if the firm communicates a separation for closing in a given quarter and no communications are recorded for the following quarters, we consider the firm closed from the quarter of its latest communication onward. The number of firms closing is 1,132.

In our analysis, we only consider open-ended contracts: for every firm we retrieve the number of open-ended contracts activated and terminated in each quarter. The total number of hirings and separations is reported in Table 1 for each quarter. The other available information at the firm level in the CC dataset concern the sector of the economic activity and the municipality in the province of Perugia where the

Table 1 CC data description, by quarter (q1–q4)

Quarter	Number of firms	Hirings	Separations	Quarter	Number of firms	Hirings	Separations
2009:q1	5,487	2,403	3,740	2011:q1	962	1,280	1,910
2009:q2	2,947	1,450	2,616	2011:q2	673	1,055	1,551
2009:q3	2,086	1,018	2,397	2011:q3	522	773	1,369
2009:q4	2,659	1,215	3,220	2011:q4	658	1,059	1,641
2010:q1	1,664	1,345	2,342	2012:q1	6,936	11,749	17,405
2010:q2	1,116	1,149	1,971	2012:q2	2,753	9,001	15,257
2010:q3	875	953	1,823	2012:q3	2,049	9,956	17,526
2010:q4	1,065	986	2,147	2012:q4	1,905	7,150	13,131

Table 2 Sectors of economic activity and municipalities

Sector	Number of firms	Municipality	Number of firms
Accommodation and food	2,770	Assisi	1,152
Activities of extraterritorial organizations	10	Bastia Umbra	944
Activities of households as employers	6,793	Castiglione del Lago	546
Administrative and support activities	1,057	Città di Castello	1,780
Agriculture, forestry and fishing	1,690	Corciano	819
Arts, sports, entertainment and recreation	705	Foligno	2,221
Constructions	4,144	Gualdo Tadino	552
Education	568	Gubbio	1,295
Electricity, gas, air conditioning supply	47	Magione	515
Financial and insurance activities	425	Marsciano	655
Health and social work activities	607	Perugia	7,795
Information and communication	958	Spoleto	1,763
Manufacturing products	4,723	Todi	781
Mining and quarrying products	46	Umbertide	708
Other personal service activities	1,829	Other	12,831
Professional, scientific, technical activities	1,388		
Public administration and defense	247		
Real estate activities	202		
Transport and storage	1,377		
Waste management	124		
Wholesale and retail trade	4,647		

firm/PA is operating. Sectors are identified by the ATECO (ATtività ECOnomiche) classification used by the Italian Institute of Statistic since 2008 (Table 2). The number of firms/PAs in each municipality is displayed in the second column of Table 2.

3 The Latent Trajectory Model

The application concerning the behavior of firms—we use hereafter the term "firm" to indicate both firms and PAs—in terms of open-ended hirings and separations during the period 2009–2012 relies on a finite mixture latent trajectory model, the assumptions of which are described in the following. Then, we give some details on parameter estimation based on the maximization of the model log-likelihood, and, finally, we deal with model selection.

3.1 Model Assumptions

We denote by i a generic firm, $i = 1, \ldots, n$, and by t a generic time window, $t = 1, \ldots, T$; in our application, we have $n = 34{,}357$ and $T = 16$. Moreover, let S_{it} be a binary random variable for the status of firm i at time t, with $S_{it} = 0$ when the firm is operating and $S_{it} = 1$ in case of firm's activity cessation in that quarter. For a firm i performing well we expect to observe all values of S_{it} equal to 0. Finally, we introduce the pair of random variables (Y_{1it}, Y_{2it}) for the number of open-ended employment contracts that firm i activated and terminated at time t. The observed number of hirings and separations is denoted by y_{1it} and y_{2it}, respectively, and it is available for $i = 1, \ldots, n$ and $t = 1, \ldots, T$ when $S_{it} = 0$, whereas when $S_{it} = 1$ no value is observed because the firm left the labor market.

To account for different behaviors in terms of open-ended hirings and separations during the time period from the first trimester 2009 to the last trimester 2012, we adopt a latent trajectory model [2,7,8] where firms are assumed to be clustered in a finite number of unobservable groups (or latent classes). Firms in each group are homogeneous in terms of their behavior and their status [6].

Let U_i be a latent variable that indicates the cluster of firm i. This variable has k support points, from 1 to k, and corresponding weights $\pi_u = p(U_i = u)$, $u = 1, \ldots, k$. Then, the proposed model is based on two main assumptions that are illustrated in the following.

First, we assume the following log-linear models for the number of hirings and separations:

$$Y_{hit}|U_i = u \sim \text{Poisson}(\lambda_{htu}), \quad \lambda_{htu} = \exp(\mathbf{x}_t'\boldsymbol{\beta}_{hu}), \quad h = 1, 2, \tag{1}$$

with $\boldsymbol{\beta}_{1u}$ and $\boldsymbol{\beta}_{2u}$ being vectors of regression coefficients driving the time trend of hirings and separations for each latent class u and $\mathbf{x}_t$ denoting a column vector containing the terms of an orthogonal polynomial of order r, which in our application is equal to 3.

Second, we account for the informative drop-out through a logit regression model, which is specified for the status of firm i at time t as follows:

$$\text{logit } p(S_{it} = 1 | S_{i,t-1} = 0, U_i = u) = \mathbf{x}_t' \boldsymbol{\gamma}_u, \tag{2}$$

where the vector of regression parameters $\boldsymbol{\gamma}_u$ is specific for each latent class u.

Note that the model described above may be extended to account for the presence of covariates, which may be included following different approaches. First, we can assume that time-constant covariates affect the probability of belonging to each latent class u, so that weights π_u are not constant across sample, but they depend on specific individual characteristics. Usually, the relation between weights and covariates is explained through a multinomial logit model. Second, linear predictors in (1) and (2) may be formulated through a combination of time-constant and time-varying covariates, in addition to the polynomial of order r.

3.2 Estimation

Parameters of the latent trajectory model described in the previous section are estimated by maximizing the log-likelihood function, which is expressed as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\mathbf{s}_i, \mathbf{y}_{1i,obs}, \mathbf{y}_{2i,obs}),$$

where $\boldsymbol{\theta}$ denotes the vector of model parameters, that is, $\boldsymbol{\beta}_{1u}, \boldsymbol{\beta}_{2u}, \boldsymbol{\gamma}_u, \pi_u$ for $u = 1, \ldots, k$, $\mathbf{s}_i = (s_{i1}, \ldots, s_{iT})'$ is a column vector describing the sequence of status observed for firm i along the time, and $\mathbf{y}_{hi,obs}$ $(h = 1, 2)$ is obtained from vector $\mathbf{y}_{hi} = (y_{hi1}, \ldots, y_{hiT})'$ omitting the missing values. Therefore, if $\mathbf{s}_i = \mathbf{0}$, then $\mathbf{y}_{hi,obs} \equiv \mathbf{y}_{hi}$, otherwise elements of $\mathbf{y}_{hi,obs}$ correspond to a subset of those of $\mathbf{y}_{hi}$.

The manifest distribution of the proposed model is obtained as

$$f(\mathbf{s}_i, \mathbf{y}_{1i,obs}, \mathbf{y}_{2i,obs}) = \sum_{u=1}^{k} \pi_u f(\mathbf{s}_i, \mathbf{y}_{1i,obs}, \mathbf{y}_{2i,obs} | U_i = u),$$

with the conditional distribution given the latent variable U_i defined as follows:

$$f(\mathbf{s}_i, \mathbf{y}_{1i,obs}, \mathbf{y}_{2i,obs} | U_i = u) = \prod_{t=1}^{T} p(s_{it} | U_i = u) \prod_{t=1:\, s_{it}=0}^{T} p(y_{1it} | U_i = u) p(y_{2it} | U_i = u),$$

for $u = 1, \ldots, k$, where $p(s_{it} | U_i = u)$ is defined in (2) and $p(y_{1it} | U_i = u)$ and $p(y_{2it} | U_i = u)$ are defined according to (1).

The maximization of function $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ may be efficiently performed through the Expectation–Maximization (EM) algorithm [3], along the usual lines based on alternating two steps until convergence

E-step: it consists in computing the expected value, given the observed data and the current values of parameters, of the complete data log-likelihood

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{u=1}^{k} z_{iu} \log \left[\pi_u f(\mathbf{s}_i, \mathbf{y}_{1i,obs}, \mathbf{y}_{2i,obs} | U_i = u) \right],$$

where z_{iu} is an indicator variable equal to 1 if firm i belongs to latent class u.

M-step: it consists in maximizing the above expected value with respect to $\boldsymbol{\theta}$ so as to update this parameter vector.

Finally, we remind that the EM algorithm needs to be initialized in a suitable way. Several strategies may be adopted for this aim on the basis of deterministic or random values for the parameters. We suggest to use both, so to effectively face the well-known problem of multimodality of the log-likelihood function that characterizes finite mixture models [6]. For instance, in our application we choose the starting values for π_u as $1/k$ for $u = 1, \ldots, k$, under the deterministic rule, and as random drawings from a uniform distribution between 0 and 1, under the random rule.

3.3 Model Selection

A crucial issue is the choice of the number k of latent classes. The prevailing approaches in the literature rely on information criteria, based on a penalization of the maximum log-likelihood, so to balance model fit and parsimony. Among these criteria, the most common are the Akaike Information Criterion (AIC; [1]) and the Bayesian Information Criterion (BIC; [11]), although several alternatives have been developed in the literature (for a review, see [6], Chap. 8). In particular, we suggest to use BIC, which is more parsimonious than AIC and, under certain regularity conditions, it is asymptotically consistent [5]. Moreover, several studies (see [9] that is focused on growth mixture models) found that BIC outperforms AIC and other criteria for model selection.

On the basis of BIC, the proper number of latent classes is the one corresponding to the minimum value of $BIC = -2\hat{\ell} + \log(n)$ #par, where $\hat{\ell}$ is the maximum log-likelihood of the model at issue. In practice, as the point of global minimum of above index may be complex to find, we suggest to fit the model for increasing values of k until the index begins to increase or, in presence of decreasing values, until the change in two consecutive values is sufficiently small (e.g., less than 1 %), and we take the previous value of k as the optimal one.

4 Results

In order to choose the number of latent classes we proceed as described above and fit the latent trajectory model for values of k from 1 to 9. The results of this preliminary fit are reported in Table 3. On the basis of these results, we choose $k = 6$ latent classes, as for values of k greater than 6 the reduction of BIC is less than 1 %.

As shown in Table 4, that describes the average number of hirings and separations for each latent class and the corresponding weight, most firms come from class 1 ($\hat{\pi}_1 = 0.524$), followed by class 3 ($\hat{\pi}_1 = 0.220$) and class 2 ($\hat{\pi}_1 = 0.198$), and do not exhibit relevant movements either in incoming or in outgoing. Indeed, the estimates of the average number of hirings and separations, obtained as $\bar{\lambda}_{hu} = \frac{1}{T}\sum_{t=1}^{T} \lambda_{1tu}$, $h = 1, 2$, are strongly less than 1. On the contrary, classes 5 and 6, that gather just the 1.4 % of total firms, show a different situation. Firms in class 5 hire 1.5 open-ended employees per quarter, whereas 2.4 employees per quarter stop their open-ended relation with the firm. As concerns firms in class 6, the average number of hirings and separations equal 6.95 and 9.89 per quarter, respectively. Besides, we observe that the separations tend to be higher than the hirings for all the classes.

With reference to the time trend of dropping out from the market, plot in Fig. 1 (top) shows that the probability of drop-out is increasing during year 2009, then it

Table 3 Model selection: number of mixture components (k), log-likelihood, number of free parameters (#par), BIC index, and difference between consecutive BIC indices (delta)

k	log-likelihood	#par	BIC	Δ
1	−476783.18	8	953649.90	–
2	−383685.95	17	767549.44	−0.1951
3	−361696.26	26	723664.05	−0.0572
4	−356020.51	35	712406.53	−0.0156
5	−348313.32	44	697086.15	−0.0215
6	−344502.01	53	689557.51	−0.0108
7	−341997.83	62	684643.13	−0.0071
8	−341091.09	71	682923.64	−0.0025
9	−339680.21	80	680195.87	−0.0040

Table 4 Estimated average number of hirings ($\hat{\bar{\lambda}}_{1u}$) and separations ($\hat{\bar{\lambda}}_{2u}$) and weights ($\hat{\pi}_u$) by latent class

	$u = 1$	$u = 2$	$u = 3$	$u = 4$	$u = 5$	$u = 6$
$\hat{\bar{\lambda}}_{1u}$	0.019	0.032	0.147	0.504	1.501	6.950
$\hat{\bar{\lambda}}_{2u}$	0.057	0.055	0.228	0.792	2.429	9.894
$\hat{\pi}_u$	0.524	0.198	0.220	0.044	0.013	0.001

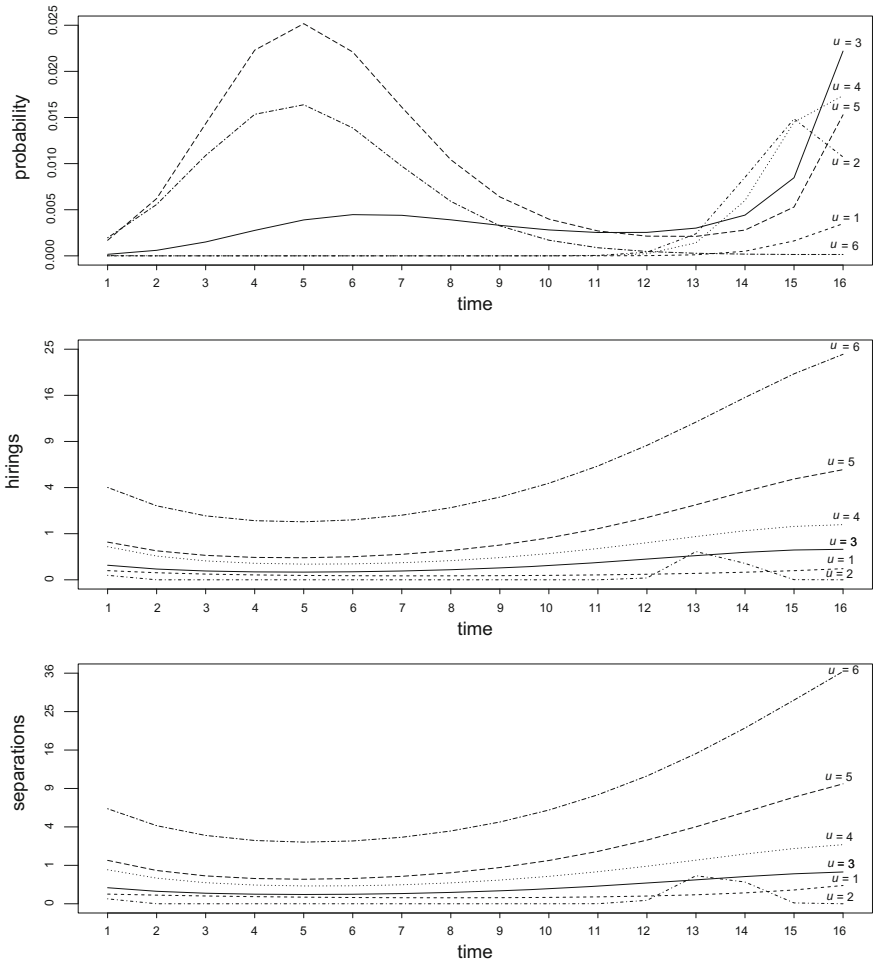


Fig. 1 Trend of the probability of leaving the market (*top*) and trends of the number of open-ended hirings (*middle*) and separations (*bottom*), by latent class

reduces and it again increases since the beginning of 2012. However, the estimated probabilities are always very small, being never higher than 2.5 %. Classes 5 and 6 are characterized by the highest probabilities of drop-out during the first two years, although firms in class 6 show the smallest probabilities of drop-out in the last year. On the contrary, class 3 shows an increase of these probabilities during year 2012, so that it has the highest probability of drop-out during the last observed quarter. Finally, firms in class 1 constantly preserve very low values.

As concerns the time trend of hirings and separations (Fig. 1 middle and bottom, respectively), both of them tend to increase along the time, although this phenomenon is evident only for classes 5 and 6. More in detail, the maxima values of hirings and

Table 5 Distribution of firms by economic sector and latent class (row frequencies)

	$u = 1$	$u = 2$	$u = 3$	$u = 4$	$u = 5$	$u = 6$
Accommodation and food	0.422	0.148	0.306	0.099	0.024	0.001
Extraterritorial organizations	0.700	0.200	0.100	0.000	0.000	0.000
Activities of households as employers	0.495	0.306	0.194	0.004	0.001	0.000
Administrative and support activities	0.553	0.141	0.192	0.085	0.025	0.006
Agriculture, forestry and fishing	0.785	0.106	0.094	0.012	0.004	0.000
Arts, sports, entertainment and recreation	0.697	0.123	0.119	0.038	0.013	0.010
Constructions	0.486	0.158	0.279	0.064	0.013	0.001
Education	0.585	0.040	0.134	0.090	0.148	0.004
Electricity, gas, air conditioning supply	0.617	0.149	0.128	0.064	0.043	0.000
Financial and insurance activities	0.642	0.172	0.142	0.028	0.017	0.000
Health and social work activities	0.604	0.191	0.147	0.033	0.017	0.008
Information and communication	0.652	0.172	0.152	0.020	0.003	0.001
Manufacturing products	0.483	0.155	0.278	0.070	0.014	0.001
Mining and quarrying products	0.435	0.152	0.304	0.087	0.022	0.000
Other personal service activities	0.627	0.215	0.140	0.011	0.005	0.002
Professional, scientific, technical activities	0.667	0.201	0.111	0.019	0.002	0.000
Public administration and defense	0.494	0.101	0.202	0.097	0.089	0.016
Real estate activities	0.704	0.166	0.121	0.000	0.010	0.000
Transport and storage	0.491	0.194	0.231	0.062	0.017	0.005
Waste management	0.517	0.125	0.275	0.050	0.033	0.000
Wholesale and retail trade	0.560	0.185	0.215	0.033	0.006	0.000

separations for firms from class 6 are achieved in the last quarter of 2012 and are equal to 23.9 and 36.5, respectively.

In order to further characterize the latent classes, we analyze the distribution of firms by economic sector (Table 5). Class 1 is characterized by a greater presence of extraterritorial organizations and of firms operating in the following sectors: agriculture, forestry, and fishing; arts, entertainment and recreation; electricity, gas, steam, and air conditioning supply; financial and insurance activities; health and social work

activities; information and communications; professional, scientific, and technical activities; and real estate activities. In class 2 there is a prevalence of activities characterized by households as employers, whereas in class 3 there is a greater presence of activities related to accommodation and food, construction, manufacturing products, mining and quarrying products, and waste management. Finally, both classes 5 and 6 show a prevalence of public administration and defense activities, other than education in case of class 5 and arts, entertainment, and recreation in case of class 6. Finally, no special difference comes out between municipalities (output here omitted).

5 Conclusions

The different trends of open-ended hirings and separations of a set of Italian firms in every quarter of the time period 2009–2012 has been analyzed through a finite mixture latent trajectory model. Six latent classes of firms were detected, which have specific trends for the probability of drop-out from the market and of hirings and separations. The results have a meaningful interpretation in the light of the recent economic downturn. In the period considered (2009–2012) the number of separations always exceeds the number of hirings of permanent employees in all clusters: such excess turnover describes the firms' tendency to diminish the labor cost by substituting permanent employees with temporary workers as well as by a reduction in the number of employees. However, the data contain only information of flows of employees so that the different levels of excess turnover may be tied only to the firms' size in each cluster. In addition, the profile of drop-out probability seems to capture the economic trend of the recent years, with a higher firm mortality rate in the moments of deepest recession (2009 and 2012).

Acknowledgments We acknowledge the financial support from the grant "Finite mixture and latent variable models for causal inference and analysis of socio-economic data" (FIRB - Futuro in ricerca - 2012) funded by the Italian Government (grant RBFR12SHVV). We also thank the Province of Perugia (Direction for "Work, Training, School and European Policies") for permitting to extract specific data from the "Compulsory Communication database of the Employment Service Centers".

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Caski, F. (eds.) Proceeding of the Second International Symposium on Information Theory, pp. 267–281. Akademiai Kiado, Budapest (1973)
2. Bollen, K.A., Curran, P.J.: Latent Curve Models: A Structural Equation Perspective. Wiley, Hoboken (2006)

3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. Ser. B. **39**, 1–38 (1977)
4. Hijzen, A., Mondauto, L., Scarpetta, S.: The perverse effects of job-security provisions on job security in Italy: Results from a regression discontinuity design. IZA Discussion Paper Number 7594 (2013). Available via DIALOG. http://ftp.iza.org/dp7594.pdf
5. Keribin, C.: Consistent estimation of the order of mixture models. Sankhya: Indian J. Stat. Ser. A **62**, 49–66 (2000)
6. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, Hoboken (2000)
7. Muthén, B.: Latent variable analysis: growth mixture modelling and related techniques for longitudinal data. In: Kaplan, D. (ed.) Handbook of Quantitative Methodology for the Social Sciences, pp. 345–368. Sage, Newbury Park (2004)
8. Muthén, B., Shedden, K.: Finite mixture modelling with mixture outcomes using the EM algorithm. Biometrics **55**, 463–469 (1999)
9. Nylund, K.L., Asparouhov, T., Muthén, B.O.: Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. Struct. Equ. Model. **14**, 535–569 (2007)
10. Schivardi, F., Torrini, R.: Identifying the effects of firing restrictions through size-contingent differences in regulation. Labour Econ. **15**, 482–511 (2008)
11. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 482–511 (1978)

Outliers in Time Series: An Empirical Likelihood Approach

Roberto Baragona and Domenico Cucina

Abstract

The empirical likelihood method is known to be a flexible and effective approach for testing hypotheses and building confidence regions in a nonparametric setting. This framework is adopted here for dealing with the outlier problem in time series where conventional distributional assumptions may be inappropriate in most cases. The procedure is illustrated by a simulation experiment. The results are also supported by the study of two well-known real-time series data: the fossil marine families extinction rates and the Nile river volume at Aswan 1871–1970.

1 Introduction

Outliers in time series may be defined as those observations that do not conform to the overall behavior of the data sequence. The time dependence that may be traced in a given time series is usually accounted for by a function that reproduces the correlation structure or more generally by a formal time series model. In the independent data framework usually outliers are searched for among either the largest or the smallest observations. In time series framework outliers are to be found instead among observations that show some unexpected departure from predicted value or

R. Baragona (✉)
Department of Economic and Political Sciences and Modern Languages, Lumsa University of Rome, Rome, Italy
e-mail: r.baragona@lumsa.it

D. Cucina
Department of Statistical Sciences, La Sapienza, University of Rome, Rome, Italy
e-mail: domenico.cucina@uniroma1.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_3

fail to fit the correlation structure deduced from the majority of the data. Such irregular behavior may be produced by outlying observations characterized by different shapes which reflect on time series statistics in some peculiar ways. Reference [14] distinguished outlying observations of four types that may distort linear model parameter estimates, i.e. additive (AO), innovation (IO), transient (TC) and permanent (LC) level change. In addition, outliers that may induce a variance change have been investigated therein as well. Other outlier types which have been considered in the literature are the so called patches, i.e. a sequence of consecutive outlying observations that do not show a steady pattern [2], and outliers in generalized autoregressive conditional heteroscedastic (GARCH) models which may impact either levels or volatility or both [1]. Further extensions refer to outliers in non linear and in vector time series (see, e.g., [7] for a review).

Statistical inference of outliers in time series usually relies on distributional assumptions for some appropriate data generating process. In this paper a distribution-free schema for building confidence regions for parameter estimates and conducting hypothesis testing in the context of time series data possibly affected by outlying observations is considered. The empirical likelihood (EL) methods [11] are adopted so that the familiar likelihood ratio statistic may be used which allows the statistical inference to be based essentially on the chi squared distribution. New developments that prove to be necessary in order to handle difficult situations are employed which came to be known as adjusted EL and balanced EL [6]. Attention is specially directed to outliers of the AO type and outliers which induce a permanent LC. A rather general framework is provided however, that allows several different other types to be handled along very similar guidelines. A simulation experiment is presented to illustrate the effectiveness of the method in case of small to moderate sample size. The results from the study of two real-time series data are also reported.

The plan of the paper is as follows. In Sect. 2 the framework in which outliers in time series are considered is explained. Specialization to particular cases is also dealt with in such a way that the developed methods may gain in generalization and are suitable for further development. In Sect. 3 inference methods are developed based on EL methods. In Sect. 4 the behavior of the statistics for inference in finite samples is outlined by means of a simulation experiment and the study of two real-time series data. Conclusions and possible suggestions for further research are provided in Sect. 5.

2 The Empirical Likelihood

EL methods have been introduced by [9–11] and have been used afterward for many applications, including time series analysis. Basically an unknown probability p_i is assigned to each observation in a sample $y = (y_1, y_2, \ldots, y_n)'$ to define an empirical probability distribution F specified by $(y_i, p_i),\ i = 1, \ldots, n$. This way the necessity to assume a family of probability distributions on which statistical inference may be based is avoided. The EL is defined instead as $L(F) = \prod_{i=1}^{n} p_i$ under the con-

straints $p_i \geq 0$, $\sum_{i=1}^{n} p_i = 1$. The probability distribution F may possibly depend on a parameters set θ so that one has to consider the maximum of $F(\theta)$ to obtain a well-defined probability distribution. If it is considered as a function of θ, $F(\theta)$ is called the profile EL.

The addition of the so-called estimating equations [11,12] to the constraint set is a further step that allows complicated models to be estimated and statistical inference to be based on EL ratio for building confidence regions and conducting tests of hypotheses. Let the data y be generated by a model which depends on a parameter vector θ of length q and assume that $r \geq q$ equations of the type

$$\mathrm{E}\{g(y, \theta)\} = 0, \quad g = (g_1, \ldots, g_r)', \tag{1}$$

exist that uniquely describe the relationships between the data and the model parameters. The functions $g_1, \ldots, g_r$ are called the estimating functions and Eq. (1) are called the estimating equations. The EL ratio may be written

$$\mathrm{ELR}(\theta) = \max_{p_1,\ldots,p_n} \left\{ \prod_{i=1}^{n} (np_i) \mid p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g(y_i, \theta) = 0 \right\}. \tag{2}$$

In Eq. (2) the EL has been divided by n^{-n} which may be shown to be the maximum EL that is obtained in correspondence of the exact solution of the system $\sum_{i=1}^{n} g(y_i, \theta) = 0$. If it is the case the probabilities p_i are all equal to $1/n$. If $r = q$ Eq. (1) are as many as the number of the unknown parameters. A model for which this circumstance occurs is often called a just identified model. In what follows such assumption will be held satisfied.

Let θ_0 denote in Eq. (2) the true parameter vector uniquely determined by the equation system $\sum_{i=1}^{n} g(y_i, \theta) = 0$. Assuming the $\{y_i\}$ to be independent identically distributed and under some conditions on g (in particular, the matrix $E\left(g(y, \theta_0)g(y, \theta_0)'\right)$ is positive definite,) [11] showed that $-2 \log ELR(\theta_0)$ converges in distribution to a χ^2 with q degrees of freedom in close agreement with the similar property which holds for ordinary parametric likelihood. So even in the absence of any assumption on the probability distribution of the data, confidence regions and tests of hypotheses may be computed all the same. Let $\mathrm{H}_0 : \theta \in \Theta_0$ be the q-dimensional null hypothesis, then the following limit in distribution holds:

$$-2\log \sup\{ELR(\theta),\ \theta \in \Theta_0\} \to \chi^2(q).$$

The case of dependent data generated by the autoregressive (AR) model has been investigated by [4] who showed that the limit in distribution still holds provided that all roots of the AR polynomial lie outside the unite circle.

3 Empirical Likelihood for Inference of Outliers in Time Series

It seems convenient in the present EL context to consider the following general time series model with outliers:

$$y_t = f(x_t, \theta) + \varepsilon_t, \tag{3}$$

where x_t summarizes all explanatory variables possibly including one or more dummies which account for outliers which occur at known time instants, and ε_t is a zero mean random error for which no distributional assumptions are made. The vector parameter θ includes both p model parameters and s outlier sizes. So the length of θ is $q = p + s$. The following procedure may be used to inscribe the inference problems related to model in Eq. (3) in the EL framework. Let $e_t = y_t - f(x_t, \theta)$. The least squares estimate $\hat{\theta}$ is obtained by solving the normal equations

$$\frac{\partial}{\partial \theta_k} \sum_{t=1}^{n} e_t^2 = 2 \sum_{t=1}^{n} (y_t - f(x_t, \theta)) \left\{ -\frac{\partial}{\partial \theta_k} f(x_t, \theta) \right\} = 0, \quad k = 1, \ldots, q. \tag{4}$$

Equation (4) are our estimating equations.

The linear autoregressive (AR) models may provide an example which show very well how this approach may be used for modeling outliers in time series data. Let the basic outlier model be

$$y_t = h(t) + z_t, \tag{5}$$

where $\{y_t\}$ is the observed time series, $h(t)$ a deterministic or stochastic function that represent outliers, and z_t the unobserved outlier free time series. Let z_t follow the stationary AR(p) process

$$\Phi(B) z_t = \varepsilon_t, \tag{6}$$

where $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 \ldots - \phi_p B^p$ is the AR polynomial and $\{\varepsilon_t\}$ are independent identically distributed random variables with mean zero and variance σ_ε^2. An AO is defined by letting in (5) $h(t) = \omega_t$ at some point t. An IO is obtained by letting $h(t) = \Phi(B)^{-1} \omega_t$. As far as level changes are concerned, the LC is defined by assuming in (5) $h(t) = \frac{1}{1-B} \omega_t$, the TC by letting $h(t) = \frac{1}{1-\delta B} \omega_t$ for some $\delta \in (0, 1)$. Let Eq. (5) be rewritten as

$$y_t = \sum_{j=1}^{p} (y_{t-j} - c_{t-j}\omega)\phi_j + c_t \omega + \varepsilon_t, \tag{7}$$

where c_t is a deterministic binary sequence, and $\{\varepsilon_t\}$ has been defined in Eq. (6). Let the outlier be located at time v and be ω its size. According to the outlier type, the sequence $\{c_t\}$ is defined as follows:

AO $c_t = 1$ if $t = v$ and $c_t = 0$ elsewhere.
IO $c_t = \psi_{t-v}$, where $\psi_j = 0$ if $j < 0$, $\psi_j = 1$ if $j = 0$, and ψ_j are the weights of the polynomial $\Psi(B) = \Phi(B)^{-1}$ if $j > 0$.
TC $c_t = \delta^{t-v}$ if $t \geq v$ and $c_t = 0$ otherwise.
LC $c_t = 1$ if $t \geq v$ while $c_t = 0$ if $t < v$.

The impact of either outlier type on the observed time series y_t has been discussed extensively by [14]. Considering ω an additional parameter makes model (7) a nonlinear time series model of the form $y_t = f(x_t, \theta) + \varepsilon_t$, where $x_t = (y_{t-1}, y_{t-2}, \ldots, y_{t-p}, c_t)'$ and the parameter vector is $\theta = (\phi_1, \ldots, \phi_p, \omega)'$.

Two cases will be considered here in some details, i.e., the AO and the LC outlier type. In both cases an AR model of order p will be assumed in the presence of a single outlier of size ω which occurs at time $t = v$. The dummy variable c_t may be built along the guidelines detailed above. Using the definitions of the explanatory variables and model parameters given before, the model in Eq. (3) reads in more compact form $y_t = x_t'\theta + \varepsilon_t$. The estimating functions $g_k(x_t, \theta)$, $k = 1, \ldots, q$, where $q = p + s$ and $s = 1$, are each of the terms in the sums in Eq. (4), i.e.,

$$g_k(x_t, \theta) = (y_{t-k} - c_{t-k}\omega)\, e_t, \quad k = 1, \ldots, p$$

$$g_{p+1}(x_t, \theta) = \left(\sum_{j=1}^{p} c_{t-j}\phi_j - c_t \right) e_t.$$

For each θ, the EL ratio function $\mathrm{ELR}(\theta)$ is well defined only if the convex hull of $\{g(x_t, \theta), t = 1, \ldots, n\}$ contains the $(p + 1)$-dimensional vector 0. Now a difficulty may arise which may well be exemplified by an AR(1) model with an AO. In this peculiar case the second line in the last constraint of Eq. (2) becomes

$$p_v g_2(x_v, \theta) + p_{v+1} g_2(x_{v+1}, \theta) = 0.$$

If the estimating functions have the same sign, the unique solution is $p_v = 0$, $p_{v+1} = 0$ and $\mathrm{ELR}(\theta)$ goes to infinity. Two kinds of EL adjustments have been suggested to address the convex hull constraint, i.e., the adjusted EL (AEL) and the balanced EL (BEL). An AEL has been proposed by [3] which consists of adding an artificial observation and then calculating the EL statistic based on the augmented data set. In the present example, this amounts to set $g_2(x_{n+1}, \theta) = -a_n \bar{g}$, where $\bar{g} = \frac{1}{n}\sum_{i=1}^{n} g(x_i, \theta)$. Reference [6] proposed a BEL where two balancing points are added to the data set, i.e., $g_2(x_{n+1}, \theta) = \delta$ and $g_2(x_{n+2}, \theta) = 2\bar{g} - \delta$. Such features will be used in the simulation experiment in next Sect. 4. The investigation on the BEL method for inference about a parameter vector θ seems very important for improving the method performance. An appropriate choice of location for the new extra points is made in order to guarantee that correct coverage levels be obtained.

4 A Simulation Experiment and Real Time Series Study

The first example run in the simulation experiment is concerned with an AO in an AR(1) model. 250 standard normal random numbers have been generated and used for building an AR(1) time series with parameter $\phi = 0.7$. The first 50 values have been discarded and an AO of size $\omega = 5$ has been added at time $v = 100$. The 90 % confidence region for the ELR test compared to the likelihood test in normality hypothesis, for one artificial time series, is displayed in left panel of Fig. 1. The confidence region computed under hypothesis of normality is narrower than that computed by the ELR statistic due to the strong distributional assumption. However as far as the AR parameter is concerned difference is negligible. Note that the BEL

had to be employed necessarily for the EL method to work properly, in accordance with the argument developed in the preceding Sect. 3. For nominal $1-\alpha$ confidence level the observed coverage, averaged for 1000 replications, based on EL ($1-\alpha_{EL}$) and normal-based confidence regions ($1-\alpha_N$) are displayed in columns 2 and 3 of Table 1. The coverage for the EL and that under normality assumptions may be considered quite satisfactory.

The second example is concerned with an LC in the same AR(1) model with standard normal innovations. In this case an outlier of size $\omega = 5$ has been added starting from time $v = 100$ on. No adjustment proved to be necessary in order to satisfy the convex hull condition. The 90 % confidence region for the ELR test compared to the likelihood test in normality hypothesis is displayed in right panel of Fig. 1 for one artificial time series. The confidence regions are quite similar in spite of the fact that much less information has been employed for building the ELR test. For nominal $1-\alpha$ confidence level the observed coverage, averaged for 1000 replications, based on EL ($1-\alpha_{EL}$) and normal-based confidence regions ($1-\alpha_N$) are displayed in columns 4 and 5 of Table 1. Results are quite satisfying overall, and with the only exception of 90 % confidence probability the EL coverage probabilities are slightly more accurate than their normal-based counterpart.

We used for computations a desktop equipped with a Intel i5 CORE processor (3.0 GHz) and 8 GB RAM running under the Windows 8.1 operating system. The algorithms were programmed in the MATLAB programming language. 1000 replicates took at most 120 seconds overall.

We also illustrate the construction of EL confidence regions through two empirical data set.

The first data set consists of the fossil marine families extinction rates collected by [13] restricted to the window of geologic time from 253 to 11.3 million years ago. This time series (39 observations) has been studied by [8] who fitted several autoregressive (AR) models. Their analysis suggests the occurrence of an outlier at $t = 30$. In view of the small sample size we adapted a first order AR model to the logarithm of the data and assumed an AO of unknown size at $t = 30$. The least squares

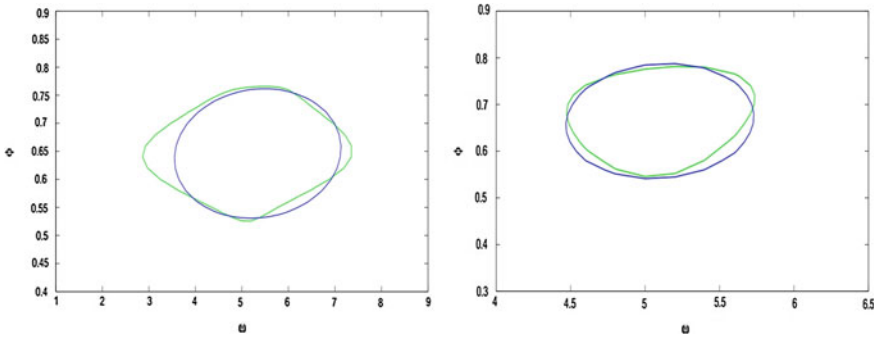


Fig. 1 AO (*left hand panel*) and LC (*right hand panel*) simulated in an AR(1) $\phi = 0.7$ $n = 200$ $\omega = 5$ $v = 100$. Confidence regions at 90 %, *green* = ELR and *blue* = normal ellipsoid

Table 1 Mean coverage across 1000 replications for an Additive Outlier and a Level Change in an AR(1) model

$1-\alpha$	Additive outlier		Level change	
	$1-\alpha_{EL}$	$1-\alpha_N$	$1-\alpha_{EL}$	$1-\alpha_N$
0.95	0.944	0.961	0.946	0.933
0.90	0.899	0.905	0.889	0.891
0.80	0.810	0.803	0.794	0.786
0.70	0.708	0.701	0.699	0.698
0.60	0.612	0.609	0.607	0.610
0.50	0.504	0.500	0.502	0.510

estimates of the autoregressive parameter and AO size are $\hat{\phi} = 0.459$ and $\hat{\omega} = 48.0$, respectively. Figure 2 (left panel) shows the 90 % EL and normal-based confidence regions for the 2-dimensional parameter $\theta = (\phi, \omega)'$. The normal ellipsoidal region is not too much larger than the EL one. Moreover, this example shows that the shape of the EL confidence regions are not constrained to be elliptical but may be markedly asymmetric.

The second data set is the time series ($n = 100$ observations) of the annual volume of discharge from the Nile River at Aswan (10^8 m^3) for the years from 1871 to 1970. The data have been taken from [5]. His study supports the occurrence of a level change at $t = 1898$. An AR(1) model has been fitted by least squares to the logarithm transform of the data, this time assuming a change in the level at $t = 1898$ while constraining the AR coefficient to remain unchanged. The estimates have been obtained $\hat{\phi} = 0.405$ for the AR coefficient and $\hat{\omega} = -0.068$ for the level change size. The 90 % EL and normal-based confidence regions for $\theta = (\phi, \omega)'$ are reported in right panel of Fig. 2. The two confidence regions nearly overlap for large values of ω while the EL confidence region is asymmetric for small values of ω. Such behavior, that has been already observed in the preceding example, originates from the fact that the elliptical shape depends on the normality assumption while the EL confidence regions shape depends on the data only.

5 Conclusions

Empirical likelihood methods have been considered for estimating parameters and outlier size in time series models and building confidence regions for the estimates. The balanced empirical likelihood has been used to obtain more accurate coverage and larger power in hypotheses testing, and to compute outlier size estimates in cases where plain empirical likelihood fails to provide feasible solutions. The procedure is illustrated by two simulated examples concerned with an additive outlier and a level change in a first -order autoregressive model. In addition, two real-world time

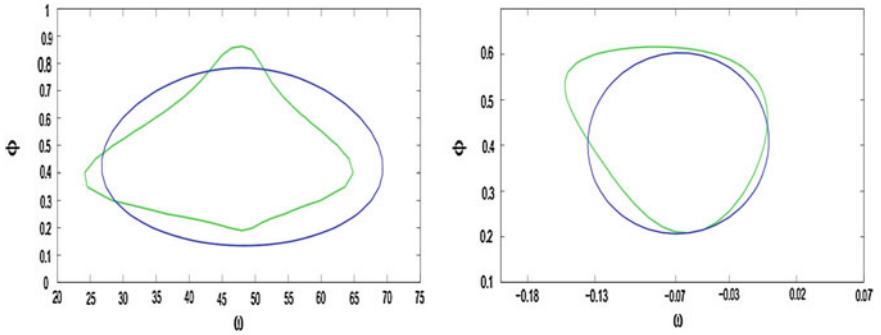


Fig. 2 Confidence regions at 90 % for the empirical likelihood (*green line*) and normal-based (*blue line*) estimates of the AR(1) parameter ϕ and outlier size ω in the presence of an AO in the extinction rate series (*left panel*) or LC in Nile river volume series (*right panel*)

series data have been studied and similar results obtained. Further interesting topics, e.g., other outlier types, including multiple outliers, and outlier identification, and estimation in a wider class of time series models, such as the general autoregressive moving average and the nonlinear models, are left for future research.

Acknowledgments This research was supported by the grant C26A1145RM of the Università di Roma La Sapienza, and the national research PRIN2011 "Forecasting economic and financial time series: understanding the complexity and modeling structural change", funded by Ministero dell'Istruzione dell'Università e della Ricerca.

References

1. Balke, N.S., Fomby, T.B.: Large shocks, small shocks, and economic fluctuations: outliers in macroeconomic time series. J. Appl. Econ. **9**, 181–200 (1994)
2. Bruce, A.G., Martin, R.D.: Leave-k-out diagnostics for time series. J. R. Stat. Soc. Ser. B. **51**, 363–424 (1989)
3. Chen, J., Variyath, A.M., Abraham, B.: Adjusted empirical likelihood and its properties. J. Comput. Graph. Stat. **3**, 426–443 (2008)
4. Chuang, C.S., Chan, N.H.: Empirical likelihood for autoregressive models, with applications to unstable time series. Stat. Sin. **12**, 387–407 (2002)
5. Cobb, G.W.: The problem of the Nile: conditional solution to a changepoint problem. Biometrika **65**, 243–251 (1978)
6. Emerson, S., Owen, A.B.: Calibration of the empirical likelihood method for a vector mean. Electron. J. Stat. **3**, 1161–1192 (2009)
7. Galeano, P., Peña, D.: Finding outliers in linear and nonlinear time series. In: Becker, C., Fried, R., Kuhnt, S. (eds.) Robustness and Complex Data Structures, pp. 243–260. Springer, Heidelberg (2013)
8. Kitchell, J.A., Peña, D.: Periodicity of extinctions in the geologic past: deterministic versus stochastic explanations. Science **226**, 689–692 (1984)

9. Owen, A.B.: Empirical likelihood ratio confidence intervals for a single functional. Biometrika **75**, 237–249 (1988)
10. Owen, A.B.: Empirical likelihood for linear models. Ann. Stat. **19**, 1725–1747 (1991)
11. Owen, A.B.: Empirical Likelihood. Chapman & Hall/CRC, Boca Raton (2001)
12. Qin, J., Lawless, J.: Empirical likelihood and general estimating equations. Ann. Stat. **22**, 300–325 (1994)
13. Sepkoski Jr., J.J.: A compendium of fossil marine families. Contrib. Biol. Geol. **51**, 1–125 (1982)
14. Tsay, R.S.: Outliers, level shifts and variance changes in time series. Int. J. Forecast. **7**, 1–20 (1988)

Advanced Methods to Design Samples for Land Use/Land Cover Surveys

Roberto Benedetti, Federica Piersimoni and Paolo Postiglione

Abstract

The particular characteristics of geographically distributed data should be taken into account in designing land use/land cover survey. The traditional sampling designs might not address the specificity of this survey. In fact, in the presence of spatial homogeneity of the phenomenon to be sampled, it is desirable to make use of this information in the sampling design. This paper discusses several methods for sampling spatial units that have been recently introduced in literature. The main assumption is to consider the geographical space as a finite population. The methodological framework is of design-based typology. The techniques outlined are: the GRTS, the cube, the SPCS, the LPMs, and the PPDs. These methods will be verified on data deriving from LUCAS 2012.

R. Benedetti (✉)
Department of Economic Studies, University of Chieti-Pescara, Chieti, Italy
e-mail: benedett@unich.it

F. Piersimoni
Agricultural Statistics Service, ISTAT, Roma, Italy
e-mail: piersimo@istat.it

P. Postiglione
Department of Economic Studies, University of Chieti-Pescara, Chieti, Italy
e-mail: postigli@unich.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_4

1 Introduction

Geographically distributed observations present particularities that should be appropriately considered when designing a survey [7,10,18]. Traditional sampling designs may be inappropriate when investigating geocoded data, because they might not capture the spatial information present in the units to be sampled. This spatial effect represents valuable information that can lead to considerable improvement in the efficiency of estimates. For these reasons, during the last decades, the definition of methods for sampling spatial units has become so popular, and many contributions have been introduced in the literature [12,13,16].

In this paper, our aim is the description and the evaluation of probability methods for spatially balanced samples. These samples have the property to be well spread over the spatial population of interest. Here, the methodological framework adopted is of design-based typology.

The spatially balanced concept is mainly based on intuitive considerations, and its impact on the efficiency of the estimates is not yet extensively analyzed. Besides, the well-spread property is not uniquely defined, and so the methods that have been proposed in the literature are based on several and personal interpretations of this concept.

In design-based sampling theory, if we assume that there is not a measurement error, the surveyed observations cannot be considered dependent. Conversely, a typical characteristic of spatial data is the dependence. Within a model-based or a model-assisted framework, a model for spatial dependence can be obviously used in defining a method for spatial sampling.

In the past, some survey scientists tried to develop methods following the intuition to spread the selected units over the space, because closer observations will provide overlapping information as an immediate consequence of the dependence [4,15]. This approach leads to the definition of an optimal sample that is the best representative of the whole population.

This sample selection cannot be evidently accepted if we follow the design-based sampling framework, since they do not respect the randomization principle. Following this approach, to consider this inherent characteristic of geographically observations, we should use the more appropriate concept of spatial homogeneity that can be measured in terms of the local variance of the variable of interest.

However, in order to select a well-spread sample, it is possible to stratify the units on the basis of their location, defining appropriate first-order inclusion probabilities. This selection strategy represents only an intuitive solution, and it has the major shortcoming that there is any impact on the second-order inclusion probabilities. Furthermore, it is not very clear how to obtain a good partition of the area under investigation.

To overcome these drawbacks, the survey practitioners usually divide the area in as many strata as possible, and select one or two units per stratum. Unfortunately, this simple plan is subjective and questionable, and so it is needed to move some steps further to define some other appropriate sampling designs.

Another objective of this paper is the application of spatially balanced samples

to land use/land cover (LULC) surveys. Land is very important for most biological and human activities on the earth. Land is the one of the main economic resource for agriculture, forestry, industries, and transport. Land can be divided into two interconnected concepts. The first is land cover that concerns the biophysical coverage of land (e.g., crops, grass, broad-leaved forest, or build-up area). The second is land use that identifies the socioeconomic use of land (e.g., agriculture, forestry, recreation or residential use).

The layout of this paper is the following. Section 2 is devoted to the description of some spatial sampling designs that have been recently proposed in literature. Section 3 contains our empirical application and discusses the main results. The case study is based on data derived from LUCAS 2012 survey. Finally, Sect. 4 concludes the paper.

2 Methodologies

The spatial distribution represents important information in designing an efficient survey or monitoring programs. Statistical units selected from a territory usually present spatial positive homogeneity. In fact, nearby locations tend to have more similar values for measured attributes than distant pairs.

Sample locations adjacent to other sample locations generally add less extra information about the target area; thus, the aim of a sampling design should be the selection of units that are well-spread across the territory under investigation. Trying to address this research question, many contributions concerning spatial sampling methods, which take into account spatial effects, have been introduced in literature.

The starting point of this debate about spatial sampling can be found in [1] that suggested a draw-by-draw scheme, called the dependent areal units sequential technique (DUST). Though DUST was inspired by purely model-based assumptions on the dependence of the stochastic process that generates the data, the properties of DUST can be also analyzed within a design-based framework, because it respects the randomization principle. The DUST begins with a unit randomly selected, say k, at every step $t < n$. Then, the algorithm updates the selection probabilities of any other unit l of the population according to the rule $\pi_l^{(t)} = \pi_l^{(t-1)}(1 - \lambda^{d_{kl}})$, where λ is a tuning parameter useful to control the distribution of the sample over the study region, and d_{kl} is the distance between unit k and l.

A step ahead in the research is represented by the introduction of the Generalized Random Tessellation Stratified (GRTS, [16]) that is now the most used plan for sampling spatial units for the monitoring of natural and environmental resources. For example, the GRTS design is a commonly used by the Environmental Protection Agency of the United States for aquatic resource monitoring.

The GRTS can be considered a very useful approach for selecting spatial samples. The GRTS is a form of spatially balanced sampling, where each point has a nonzero

probability of being included in the sample, and pairwise probabilities of including both points i and j are nonzero; thus supporting design-based inferences to the entire area. It integrates the benefits of being a probability sample with the characteristics of being spatially balanced.

The underlying idea is the extension of the use of systematic sampling to two or more dimensions [16]. This plan systematically draws the units, transforming the two-dimensional population into one dimension population. Besides, this design seeks to preserve some multidimensional order. The aim is that no points in the target population are too far from a sampled point, and few sampled points are close together.

The main idea is to apply recursive partitioning to create a spatial address. At each step of the procedure, the first-order inclusion probability for each cell is computed as the sum or integral of the first-order inclusion probability of all population elements within the cell. The first-order inclusion probability need not be constant, and very general variable probability designs can be adapted. The recursive procedure is executed until every cell has total first-order inclusion probability less than one, and then hierarchical randomization is applied. To each cell is assigned a length equal to its first-order inclusion probability, and then the lengths are linked together, forming a line with length equal to the total sample size. A systematic sample is selected along the line. This one-to-one recursive map guarantees that every point on the line agrees to some population elements.

The GRTS is essentially based on the Voronoi polygons that can be used to define an index of spatial balance. This index might be very helpful for comparing how well two or more algorithms spread a set of points over the study region.

For a generic sample $s = \{s_1, s_2, \ldots, s_n\} \in \Omega$, where Ω is the set of all possible samples, the Voronoi polygon for the sample unit $s_i = 1$ includes all population units closer to si than to any other sample unit $s_j = 1$. Now, define with v_i the sum of the first-order inclusion probabilities of all units in the i-th Voronoi polygon. Then, for any sample unit, we will have an expected value $E(v_i) = 1$; while all the v_{ks} should be close to 1 for a spatially balanced sample [16].

Thus, the variance of the v_i (i.e., $V(v_i)$) can be used as a measure of the spatial balance of a sample. Obviously, a lower value of $V(v_i)$ indicates a good spatially balanced sample. For details and some empirical illustrations about the GRTS and the spatial balance index, see [3].

A possible alternative scheme is represented by the balanced sampling and the cube technique [9]. The cube is a method for selecting balanced samples with equal or unequal first-order inclusion probabilities.

The idea underlying the plan is very simple. A researcher may request to check the quality of the selected sample by verifying how the plan works on some covariates $\mathbf{X}$ known for every unit of the population U. This method is based on the expectation that an error committed in estimating an auxiliary variable could be replicated in a similar manner on the survey variables.

This argument can be explained through the use of HT estimator of each auxiliary variable. In this case, the aim is to check if the HT estimator $\hat{t}_{HT,x_j}$ is close to the

known population total t_{x_j} for each of the available q covariates. Thus, a samples is said to be balanced on variables $\mathbf{X}$, if the following property is satisfied

$$\sum_{k \in s} w_k x_{kj} = \hat{t}_{HT,x_j} = t_{x_j} = \sum_{k \in U} x_{kj}, \forall j = 1, \ldots, q, \tag{1}$$

for all the $s \in \Omega$ such that $p(s) > 0$, and where x_{kj} is the value of j-th variable for the k-th unit. It is worth noticing that a sampling design satisfying balancing equations (1) does not necessarily exist. Therefore, the appropriate goal can be considered to find an approximate solution.

The cube method, proposed by [9], is composed of two phases: the *flight* and the *landing* phases. During the flight phase, the constraints, represented by the balancing equations (1), should be always exactly satisfied. The landing phase starts at the end of the flight phase, only if a sample is not obtained. In the landing phase, a sample is selected as close as possible to the constraint subspace. One possible method for the landing phase is to use an enumerative algorithm.

It is evident that the cube method has been introduced in a nonspatial context. However, the cube can be straightforwardly applied in a spatial setting by using the coordinates of the units as covariates and by imposing that any selected sample should respect for each coordinate the first p moments; assuming implicitly that the survey variable y follows a polynomial spatial trend of order p [6].

Note that these last sampling designs (i.e., GRTS and cube) do not explicitly use the concept of distance that is a key statistic to describe the spatial distribution of the sample units. As already noted before, since the homogeneity showed by geographically distributed data, the units that are close in space should rarely be selected in a sample, as they would provide similar information. The use of the distance in designing a sample might face better this need.

In fact, under the spatial homogeneity assumption, increasing the distance between two units i and j, the difference $|y_i - y_j|$ between the values of the survey variable is always increased. In this situation, it is evident that the estimates of the variance of the HT estimator will necessarily decrease, if we set high second-order inclusion probabilities to couples with very different y values that are far in the territory.

It is worth noting that the assumptions of stationarity and/or isotropy are crucial for defining spatially balanced samples based on distance measures, since the distances are sufficient for modeling spatial homogeneity only if the two previous assumptions are satisfied.

Many scientists try to address the problem of selecting spatially balanced samples using the distance between geographical units.

Reference [12] suggested a method called spatially correlated Poisson sampling (SCPS) by modifying the correlated Poisson sampling (CPS), introduced by [5].

The CPS is based on a list sequential of the visit of the units of the population. The units of the population are visited one by one in some order. The researcher must decide at the visit whether the unit should be sampled, since there is no possibility to subsequently revisit units. The method selects each unit k according to first-order inclusion probability. After each sampling decision, the first-order inclusion probabilities for the remaining units are updated according to a specific updating rule, to

generate correlations between the indicator variable of the unit visited, say I_k, and the indicator variables relative to all the other units of the population, say I_l, with $l \neq k$. Obviously, it is suitable to have negative correlations between the indicator variables of units that are closer to those already selected. In this case, it is very difficult that these closer units are included in the sample, thus defining good spatially balanced samples.

The algorithm can be described in the following way. If unit 1 is selected with probability $\pi_1^{(0)} = \pi_1$, we will set $I_1 = 1$ and $I_1 = 0$ otherwise. Starting with $\pi_k^{(0)} = \pi_k$, $k \geq 1$. At step t, the values of $I_1, I_2, \ldots, I_{t-1}$ are known, we will include the unit t with probability π_t^{t-1}. After each step, the inclusion probabilities for the remaining units, $k \geq t+1$, in the list are updated according to $\pi_k^{(t)} = \pi_k^{(t-1)} - (I_t - \pi_t^{(t-1)})w_{k-t}^{(t)}$, where $w_{k-t}^{(t)}$ are weights that depend on $I_1, I_2, \ldots, I_{t-1}$ but not on I_t [5].

The method is very flexible. In fact, as stated by [5], every without replacement design with predefined inclusion probabilities can be implemented through CPS.

Reference [12] followed the same logic of CPS to define SCPS method. The contribution of [12] can be essentially found in the definition of two different strategies for choosing the weights $w_{k-t}^{(t)}$: maximal weights and Gaussian preliminary weights.

The maximal weights strategy selects samples of fixed size, if the inclusion probabilities sum to an integer. After a decision on the unit t, the maximal weight strategy choose weights giving as much weight as possible to the closest unit (in distance) among the units $k = t+1, t+2, \ldots, N$; then as much weight as possible to the second closest unit, and so on with the restriction that the weights sum to 1. The maximal weight strategy always provides samples of fixed size if the inclusion probabilities sum to an integer.

The Gaussian preliminary weight strategy chooses weights with sum one that are controlled by a Gaussian distribution centered on the position of unit k. Note that it performs worse than the maximal weights method [12]. Essentially, the main idea of the SCPS method is a careful tuning of a procedure for selecting πps samples with fixed π_k, obtained by introducing the correlation between selection probabilities or by modifying the π_{kl}s (which remain unknown and cannot be fixed in advance).

A similar approach led [14] to introduce two alternative procedures to select samples with fixed π_k and correlated inclusion probabilities. These two methods are referred to as the local pivotal method 1 (LPM1), and the local pivotal method 2 (LPM2). These methods constitute an extension of the pivotal method introduced to select πps samples [8].

The LPM methods draw samples considering distances between units and in accordance with the updating rule of the Pivotal method, for two nearby units at each step. The LPM methods update the first-order probabilities π_k and π_l at each step. In order to select sample units, it is possible to choose between the LPM1, which is more spatially balanced, and the LPM2, which is less balanced, but computationally more feasible. The LPM1 randomly chooses the first unit k and then the closer unit l (if two or more units have the same distance to k, the method randomly chooses between them), under the hypothesis that k is the nearest neighbor of l. The LPM 2

is very similar to the LPM 1; the only difference is that the assumption of nearest neighboring between the two units is removed. In the case of availability of adequate auxiliary information, Local Pivotal methods allow to select samples that are well spread in the space.

However, the geographically distributed units are usually influenced by some spatial effects. In order to effectively use the distance in designing spatially balanced sampling, we need to suppose that the distance matrix summarizes all the features of the spatial distribution of the population and, as a consequence, of the sample. This general hypothesis within a design-based perspective implies that the problem of selecting spatially balanced samples can be led back to the definition of a design $p(s)$ with probability proportional to some synthetic index $M(D_s)$ of the matrix D_s of the distances observed within each possible sample s by using some MCMC algorithm to select such a sample [17].

This intuitive consideration constitutes the rationale on which is based the method developed by [2]. The algorithm starts at iteration $t = 0$, with an initial point $s(0)$, randomly selected from $\{0, 1\}^N$ according to a simple random sampling (SRS) with constant inclusion probabilities. In a generic iteration t the elements of $s(t)$ are updated in the subsequent steps

1. select at random two units included and not included in the sample in the previous iteration, say i and j. Formally, one among the units within the sample, for which $s_i^{(t)} = 1$, and another among the units outside the sample for which $s_i^{(t)} = 0$, respectively;
2. denote with ${}^*s_i^{(t)}$ the sample where the units in the position i and j exchange their status. Randomly decide whether to adopt ${}^*s_i^{(t)}$, that is

$$s^{(t+1)} = \begin{cases} {}^*s^{(t)}, & \text{with probability } p = min\{1, (M(D_{{}^*s^{(t+1)}})/M(D_{s^{(t+1)}})\} \\ s^{(t)}, & \text{otherwise} \end{cases} \quad (2)$$

3. repeat Steps 1. and 2. mq times.

The index $M(D_s)$ can be obviously defined in different ways. As evidenced by [2], *better* empirical results can be found with the use of

$$M(D_s) = \prod_{i;s_i=1} \prod_{j \neq i;s_j=1} d_{ij}^{\beta}, \quad (3)$$

where the exponent β gives the possibility of modeling the spread of the sample. Higher values of β produce samples with more spread units. The design obtained using the index (3) is defined as probability proportional to the product of the distances design (PPD).

The function (3) has an appealing interpretation, since it is congruent with the underlying assumption of proportionality of the distance between two units d_{ij} and their second-order inclusion probabilities π_{ij}. If such situation hold, the function (3)

could be viewed as an attempt to approximate the probability $p(s)$ through the product of the second-order inclusion probabilities of each couple $\{i, j\}$. In this case, we implicitly consider as realistic the additional hypothesis of conditional independence of the probabilities of order higher than two. Such definition of $p(s)$ is compatible with the conventional practice to try to control the π_i and the π_{ij}, since they directly influence the estimates and their variance without taking care of higher order probabilities.

Note that the underlying hypothesis in defining spatially balanced designs is that the spatial phenomenon under investigation shows a positive homogeneity. In the case of negative homogeneity (i.e., nearby locations tend to have more dissimilar values for measured attributes than distant pairs), the previous methods should be modified in order to provide spatial clustered samples, and not well-spread samples as in the case of positive homogeneity. Finally, if the phenomenon does not present any spatial homogeneity, the spatially balanced methods have similar efficiency of SRS.

For a detailed review of spatially balanced samples, the interested reader can refer to [3].

3 Empirical Evidence

In this Section, we will present an empirical application: the methods described in the previous Sections will be verified on LULC data derived from LUCAS 2012 survey. The aim is to compare the spatial sampling methods, highlighting advantages and drawbacks of each technique. In order to reach this objective, we built an artificial data set based on the LUCAS survey.

The Land Use/Cover Area frame Survey (LUCAS, [11]) is a project funded by EUROSTAT that has as main objective the production of European crop estimates. Furthermore, the LUCAS survey also delivers land use data, and is a valuable tool for environmental monitoring.

The sample frame is performed at the country level, because it is impossible to produce a regular grid over the complete European territory for statistical purposes. LUCAS is a spatial reference frame survey of point typology, and is based on the official digital geographic data of the administrative boundaries and coastlines of Europe. Very recently, EUROSTAT realized the LUCAS 2012 survey in the European Union. LUCAS 2012 covers all 27 EU countries.

Our artificial data set has been built considering 2012 as reference year and Italy as country under investigation. A sample of 21,013 points was drawn from a population obtained overlaying a regular grid of points selected every 2 km to the Italian national boundaries map. This sample constitutes our reference population (or first-phase sample) that was used in order to verify the different spatially balanced samples methods that have been described in the previous Section.

We compared the different designs using the mean square errors (MSEs) of the 10,000 HT estimates of the population mean to the same error obtained when using

an SRS design that is used as benchmark design. Note that, in every simulation, the MSEs were always very close to the variance of each design because the HT estimator was unbiased.

We investigated the performances of the GRTS, the balanced sampling constrained to first and second-order moments of the coordinates (i.e., CUBE 1 and CUBE 2), SCPS, LPM 1 and LPM 2, and PPD 1, PPD 5, PPD 10 (i.e., with $\beta = 1, 5, 10$). In this application, we only present the results obtained for land use survey.

Note that for reasons of space, it has not been possible to give tables about land cover survey. However, the results obtained are very similar to those described for land use survey.

The main evidences are reported in Table 1. We describe the performance of different plans using three different sample sizes $n = 100, 300, 600$, in order to evaluate the effect of sample size on the different spatial designs. The sample sizes chosen approximately represents a range between 1 – 10 % of the correspondent population under investigation. This choice is in accordance with other existing surveys on land cover/land use (for example, LUCAS). Our case study concerns Emilia Romagna, a NUTS 2 Italian region (see Fig. 1).

The results showed that the spatially balanced designs are generally more efficient than SRS. In particular, the results are coherent across different land uses and sample sizes. Besides, it is worth noticing that the worst performance are obtained with the GRTS plan. In some case, the GRTS is similar to SRS, showing that its spread over space is too light to substantially reduce the sampling errors. Better performance are usually obtained with PPD plans, corroborating our idea that the distance between units is a peculiar information that must be considered when sampling spatial units. Note that PPD plans with higher values of β (i.e., with units more spread) showed

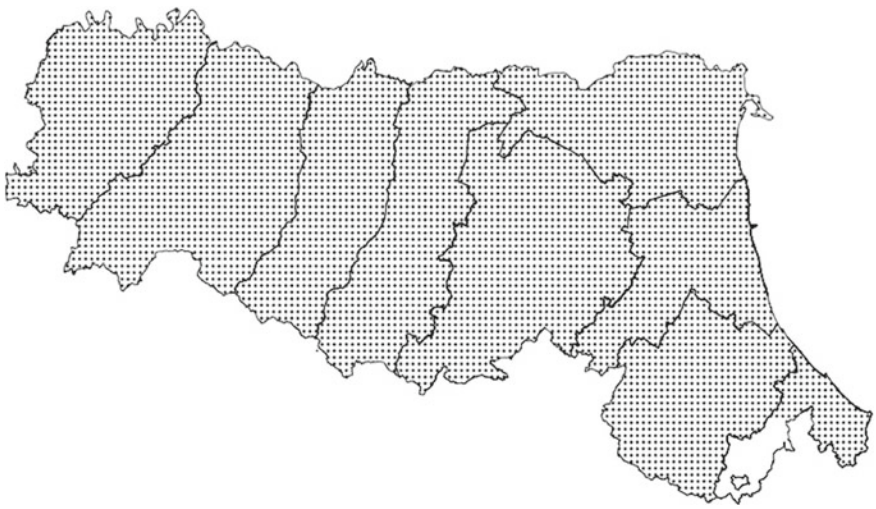


Fig. 1 Population grid overlaid the NUTS 3 provinces of Emilia Romagna

Table 1 Relative efficiency (MSE/MSE_{SRS}) and mean (μ) of the spatial balance indices of the area estimates for each land use and for each design estimated in 10,000 replicated samples in Emilia Romagna for $n = 100, 300, 600$

	n	Agriculture	Forestry	Urban	Unused	Other	μ
CUBE 1	100	0.892	0.860	1.000	0.991	1.005	0.329
CUBE 2	100	0.901	0.855	0.995	0.987	0.994	0.306
GRTS	100	0.859	0.774	0.956	0.938	0.938	0.288
LPM 1	100	0.850	0.753	0.941	0.917	0.921	0.110
LPM 2	100	0.847	0.770	0.943	0.927	0.921	0.068
SCPS	100	0.840	0.749	0.931	0.928	0.918	0.070
PPD 1	100	0.875	0.780	0.955	0.939	0.942	0.054
PPD 5	100	0.848	0.749	0.928	0.926	0.932	0.076
PPD 10	100	0.833	0.736	0.936	0.900	0.943	0.044
CUBE 1	300	0.908	0.869	1.001	0.971	0.992	0.311
CUBE 2	300	0.899	0.866	0.974	0.966	0.996	0.303
GRTS	300	0.829	0.733	0.906	0.877	0.871	0.296
LPM 1	300	0.786	0.705	0.877	0.846	0.837	0.103
LPM 2	300	0.788	0.716	0.866	0.846	0.837	0.068
SCPS	300	0.787	0.708	0.871	0.845	0.823	0.070
PPD 1	300	0.819	0.748	0.903	0.886	0.894	0.055
PPD 5	300	0.774	0.695	0.863	0.836	0.825	0.075
PPD 10	300	0.771	0.687	0.854	0.840	0.800	0.045
CUBE 1	600	0.893	0.858	1.000	0.965	1.005	0.305
CUBE 2	600	0.893	0.854	1.000	0.952	0.997	0.301
GRTS	600	0.780	0.699	0.881	0.834	0.818	0.298
LPM 1	600	0.730	0.651	0.841	0.767	0.756	0.109
LPM 2	600	0.728	0.648	0.825	0.781	0.779	0.074
SCPS	600	0.715	0.652	0.827	0.763	0.755	0.078
PPD 1	600	0.779	0.703	0.887	0.840	0.847	0.066
PPD 5	600	0.718	0.648	0.824	0.758	0.764	0.085
PPD 10	600	0.707	0.635	0.816	0.756	0.745	0.056

a gain of 36.5 % in relative efficiency for some land uses (for example, Forestry) if compared with SRS.

As already noted in Sect. 2, it is possible to use the spatial balance index to compare different spatial designs. In the last column of Table 1, we report the mean of spatial balance indices of 10,000 replicated samples for the different designs analyzed in this paper. The lower values of these means are acquired in correspondence of PPD plans: these results highlights that the more well-spread samples are obtained using PPD plans.

In conclusion, better performance in terms of relative efficiency is obtained using plans that are well-spread and that make effective use of the distance between spatial units (i.e., PPD designs).

4 Concluding Remarks

The particular characteristics of the geographically distributed populations should be considered when a sample is designed. In fact, many populations in environmental, agricultural, and forestry studies are distributed over space, and it is clear that spatial units cannot be sampled as if they were generated under the classical independent urn model. The main challenge for the researchers is how to include these spatial effects in the sampling designs to reduce the variance of the estimators. Unfortunately, the methods of spatial systematic sampling and spatial stratified sampling take advantage only partially from these peculiarities. For these reasons, in the past decades, many sampling designs that explicitly consider these spatial characteristics have been introduced in the literature.

The foremost research issue regards the capacity of a sample to be well-spread across the territory considering the occurrence of any spatial structure that is present in the geographically distributed data under investigation.

In this paper, we compared different spatial designs for land use/land cover surveys. The main results are that, if these spatial characteristics of the data exist, and the method uses this information, there can be a remarkable reduction of the sampling error if compared with SRS. Generally speaking, the results obtained indicate good performances for spatially balanced samples with particular reference to spatial designs based on the distance between geographical units. The computational effort of these methods is generally not prohibitive also in the case of large frame.

Several issues remain open for future research. In particular, it should be necessary to theoretically derive the second-order inclusion probabilities π_{kl}s that are often unknown.

References

1. Arbia, G.: The use of GIS in spatial statistical surveys. Int. Stat. Rev. **61**, 339–359 (1993)
2. Benedetti, R., Piersimoni, F.: A spatially balanced design with probabilities proportional to the within sample distance. Submitted (2014)
3. Benedetti, R., Piersimoni, F., Postiglione, P.: Sampling Spatial Units for Agricultural Surveys, Advances in Spatial Science Series. Springer, Heidelberg (2015)
4. Benedetti, R., Palma, D.: Optimal sampling designs for dependent spatial units. Environmetrics **6**, 101–114 (1995)
5. Bondesson, L., Thorburn, D.: A list sequential sampling method suitable for real-time sampling. Scand. J. Stat. **35**, 466–483 (2008)

6. Breidt, F.J., Chauvet, G.: Penalized balanced sampling. Biometrika **99**, 945–958 (2012)
7. Delmelle, E.M.: Spatial sampling. In: Fischer, M.M., Nijkamp, P. (eds.) Handbook of Regional Science, pp. 1385–1399. Springer, Berlin (2013)
8. Deville, J.C., Tillé, Y.: Unequal probability sampling without replacement through a splitting method. Biometrika **85**, 89–101 (1998)
9. Deville, J.C., Tillé, Y.: Efficient balanced sampling: the cube method. Biometrika **91**, 893–912 (2004)
10. Dickson, M.M., Benedetti, R., Giuliani, D., Espa, G.: The use of spatial sampling designs in business surveys. Open J. Stat. **4**, 345–354 (2014)
11. Gallego, J., Delincè, J.: The European land use and cover area-frame statistical survey. In: Benedetti, R., Bee, M., Espa, G., Piersimoni, F. (eds.) Agricultural Survey Methods, pp. 151–168. John Wiley & Sons Ltd, Chichester (2010)
12. Grafström, A.: Spatially correlated Poisson sampling. J. Stat. Plan. Inference **142**, 139–147 (2012)
13. Grafström, A., Schelin, L.: How to select representative samples. Scand. J. Stat. **41**, 277–290 (2014)
14. Grafström, A., Lundström, N.L.P., Schelin, L.: Spatially balanced sampling through the pivotal method. Biometrics **68**, 514–520 (2012)
15. Rogerson, P., Delmelle, E.: Optimal sampling design for variables with varying spatial importance. Geograph. Anal. **36**, 177–194 (2004)
16. Stevens Jr., D.L., Olsen, A.R.: Spatially balanced sampling of natural resources. J. Am. Stat. Assoc. **99**, 262–278 (2004)
17. Traat, I., Bondesson, L., Meister, K.: Sampling design and sample selection through distribution theory. J. Stat. Plan. Inference **123**, 395–413 (2004)
18. Wang, J.F., Stein, A., Gao, B.B., Ge, Y.: A review of spatial sampling. Spat. Stat. **2**, 1–14 (2012)

Heteroscedasticity, Multiple Populations and Outliers in Trade Data

Andrea Cerasa, Francesca Torti and Domenico Perrotta

Abstract

International trade data are often affected by multiple linear populations and heteroscedasticity. An immediate consequence is the false declaration of outliers. We propose the monitoring of the White test statistic through the Forward Search as a new robust tool to test the presence of heteroscedasticity. We briefly describe how the regression estimates change when considering a heteroscedastic regression model. We finally show that, if the data are analyzed on a monthly basis, the heteroscedastic problem can be often bypassed.

1 Introduction

The international trade between EU Member States and third countries produces a huge amount of data which are first collected by the Customs and then monthly aggregated by national statistical offices (e.g., the Italian ISTAT). The analysis of the resulting dataset through suitable statistical procedures is usually focused on the detection of anomalies of various kinds: recording errors, specific market price

A. Cerasa (✉) · F. Torti · D. Perrotta
Institute for the Protection and Security of the Citizen (IPSC), Global Security and Crisis Management Unit, European Commission, Joint Research Centre (JRC), Ispra, Italy
e-mail: andrea.cerasa@jrc.ec.europa.eu

F. Torti
e-mail: francesca.torti@jrc.ec.europa.eu

D. Perrotta
e-mail: domenico.perrotta@ec.europa.eu

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_5

dynamics, or more in general transactions which are not in line with the market. Monitoring such discrepancies is of primary interest for EU authorities as they could hide unfair or fraudulent commercial behaviors, such as trade-based money laundering, dumping, import duty evasion. All these can have a big negative impact on the EU economy, the EU budget, and on the internal competition of EU market. The statistical treatment of such data requires the availability of robust methods, in order to obtain results that are not affected by the irregularities that we are looking for.

One of the robust statistical techniques most successfully applied in this context is the Forward Search, "...a powerful general method for detecting unidentified subsets and masked outliers and for determining their effect on models fitted data" [6]. In [6], the Forward Search is applied on imports of a fishery product that is now available in the FSDA toolbox [7]. The Forward Search output of the regression of volume of trades on quantities traded clearly points out the presence of a pointwise outlier and of a group of transactions occurred at a price sensibly lower than the normal ("fair") price of the product. Unfortunately, international trade dataset are sometimes characterized by two general problems that dramatically affect the good properties of Forward Search and could lead to misleading conclusions: the presence of multiple linear structures and the heteroscedasticity. They can be considered as a direct consequence of the limits of products' classification, whose categories are often not enough precise to distinguish the different quality levels of the products in the same category. This topic has been extensively debated in several official documents regarding the possibility of using custom data for the calculation of Import/Export Price Indexes. See [5].

A typical case of multiple populations is presented in the left panel of Fig. 1, where the exchanged quantities of an ornamental fish and the corresponding volumes are plotted. It is possible to clearly detect at least three linear relations, representing three different prices. It is obvious that no robust method based on a linear relation between quantity and volume, like the Forward Search (right panel), would be able to disentangle the three relations and to provide a plausible result.

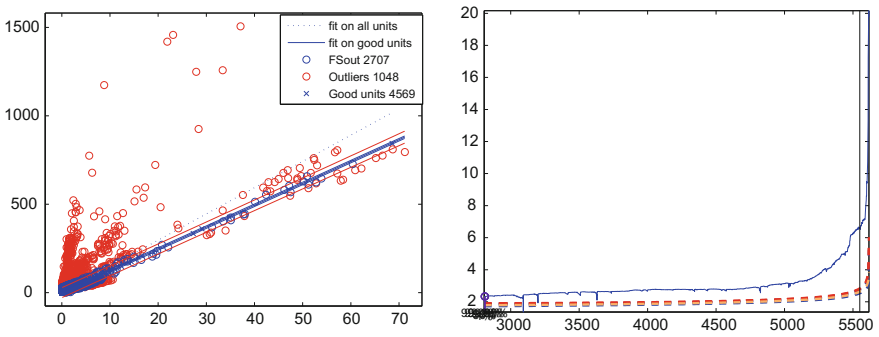


Fig. 1 EU imports of an ornamental fish. Data represent monthly aggregates for the period starting in January 2009 until December 2011. *Left panel* scatterplot of VALUE (in thousands of Euro) against QUANTITY (in tons) exchanged. *Right panel* forward plot of the Minimum Deletion Residual

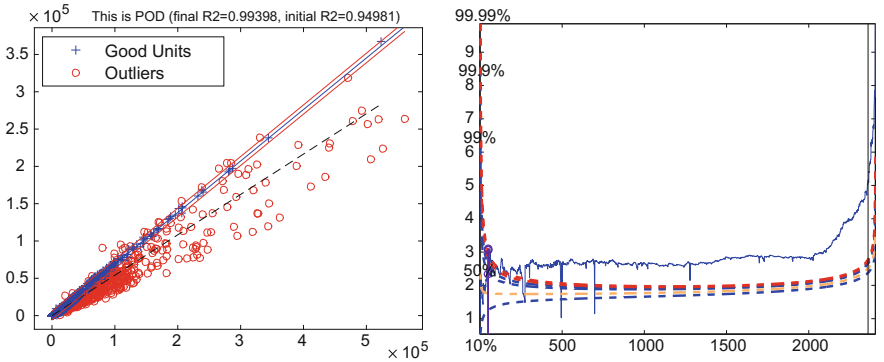


Fig. 2 EU imports of a mineral product. Data represent monthly aggregates for the period starting in January 2009 until December 2011. *Left panel* Forward Search fit on the scatterplot of VALUE (in thousands of Euro) against QUANTITY (in tons) exchanged; the *dashed line* is the fit on all data. *Right panel* forward plot of the Minimum Deletion Residual; the trajectory escapes from the bands since the very first steps

Left panel of Fig. 2 represents instead a case of heteroscedastic trade data; the good considered here is a mineral product. As the data show, the assumption of homoscedasticity for the OLS residuals is clearly violated. Again, the results and the conclusions based on the Forward Search regression (right panel) could be affected by the violation of such a fundamental assumption and should be evaluated carefully. Therefore, the availability of an instrument which is able to point out when we can trust the single normal model and therefore the Forward Search output and when instead it has been probably contaminated by one of the mentioned problems, is of primary importance. It could help us indeed to distinguish the real discrepancies in trade data from the "spurious" ones.

The paper is organized as follows. In the next section, the consequences of the application of the Forward Search on dataset characterized by multiple populations and heteroscedasticity are analyzed. Then, the use of the White test as a diagnostic tool for monitoring and evaluating the departure from the homoscedasticity assumption is presented. A first attempt to adopt an heteroscedastic model in trade data is given in Sect. 4. The consequences on the regression estimates of such model are briefly discussed. Finally, a monthly price approach is introduced as a possible alternative to address the problem without adopting a heteroscedastic model.

2 The Forward Search in Presence of Multiple Populations and Heteroscedasticity

The basic idea of the Forward Search [1] is to start from a small, robustly chosen, subset of the data and to fit subsets of increasing size, in such a way that outliers and subsets of data not following the general structure are clearly revealed by diagnostic

monitoring. If there is only one population the increasing fitting from a few observations to all observations will be stable. Otherwise, if in the data there are outliers or groups, there will be a point where the stable progression of fits is interrupted. In the classical regression model we have one univariate response Y and v explanatory variables

$$X_1, \ldots, X_v$$

satisfying

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_v x_{iv} \tag{1}$$

under the usual assumptions, in particular $E(\varepsilon_i) = 0$ and $E(\varepsilon_i^2) = \sigma^2$ being $\varepsilon_i = |E(y_i) - y_i|$. Each subsample is obtained by looking at the n squared regression residuals

$$e_i^2(m) = [y_i - \{\hat{\beta}_0(m) + \hat{\beta}_1(m)x_{i1} + \ldots + \hat{\beta}_v(m)x_{iv}\}]^2 \qquad i = 1, \ldots, n$$

computed from the OLS estimate of beta at step m. $S(m+1)$ is defined as the subset of observations corresponding to the $m+1$ smallest squared residuals $e_i^2(m)$. The search starts from an outlier-free subset of m_0 observations. Usually $m_0 = v + 1$, with $S(m_0)$ chosen through the least median of squares criterion of [8].

In our context, the response variable is the value of the transactions, whereas the explanatory one is the quantity of product exchanged. The output of the Forward Search applied to the ornamental fish trades of Fig. 1 and to the mineral product data of Fig. 2 are represented in the corresponding right panels. In the case of multiple population, the Forward Search automatically selects the linear structure with the highest leverage and declares outliers all the points of the other structures. However, this declaration may be misleading since the points detected could simply represent prices of different quality levels. Besides, in the case of heteroscedasticity the Forward Search tends to overdeclare the outliers. It seems indeed that most of them are actually coherent with the model, once heteroscedasticity is taken into account. Moreover, the line estimated by the method does not seem to represent the central tendency of the data.

Then the main consequence of applying the Forward Search in the presence of heteroscedasticity or multiple linear structures is an overdeclaration of outliers. This represents a serious problem in the anti-fraud context because each outlier should be analyzed in detail in order to determine whether it may hide or not a fraudulent behavior and to initiate a costly investigation. In other words, declaring false outliers means wasting economic resources and should be thus avoided.

3 Forward Plot of the White Test

In order to verify the presence of heteroscedasticity, we can use the well-known White test [4]. The White test is very popular also because it does not assume a specific form of heteroscedasticity. It is based on an auxiliary regression with squared residuals as dependent variable and independent variables given by the regressors of the initial

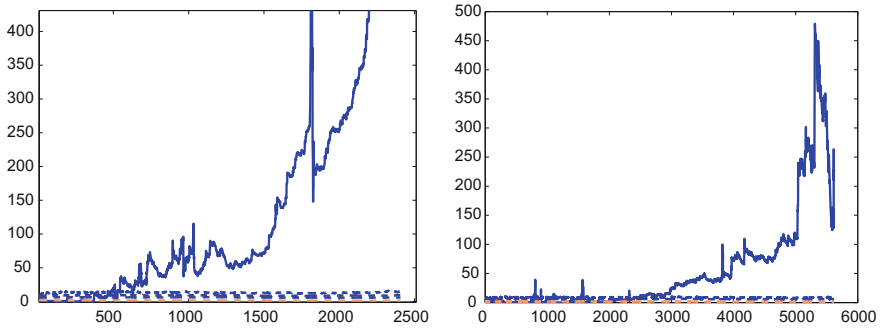


Fig. 3 Forward plot of the White test statistic for the mineral product (*left panel*) and the ornamental fishery (*right panel*) datasets

model, their squares and their cross products. To avoid the bias introduced by the possible presence of outliers, we have robustified the test by monitoring the statistic with the Forward Search. More precisely, the quantity monitored is the coefficient of determination of the following auxiliary regression, that we report for the simple case of one independent variable:

$$e_i^2(m) = \hat{\alpha}_0(m) + \hat{\alpha}_1(m)x_{i1} + \hat{\alpha}_2(m)x_{i1}^2 + u_i \qquad i = 1, \ldots, n \tag{2}$$

Figure 3 presents the forward plots of the White test statistic for the two datasets considered. For the mineral product (left panel) the plot shows that after about 500 steps the statistic exceeds the 90, 95 and 99 % confidence bands, obtained through montecarlo simulations, represented with dashed lines. This highlights very clearly the presence of heteroscedasticity in the dataset. As a consequence, the forward plot of the minimum deletion residual, traditionally monitored in the regression context, exceeds the bands from the very first steps (right panel of Fig. 2). In conclusion many outliers are wrongly declared. The right panel of Fig. 3 demonstrates that the forward plot for the White test can be also used to detect situations where data contain different linear structures, like for the ornamental fish dataset.

As a result, in our operational context we can use the proposed approach to highlight cases when the standard Forward Search outlier detection model is not reliable and more complex models have to be adopted. Possible models to account for multiple populations in this difficult context have been addressed elsewhere (see for example [3,6]). Here, we briefly outline a possible approach to heteroscedastic data (Sect. 4) and a practical approach to work around the problem, specifically applicable in our context (Sect. 5).

4 Forward Search Monitoring with Harvey's Heteroscedastic Model

A natural step forward for avoiding the problems exemplified by Fig. 2 is to replace in the Forward Search the traditional homoscedastic model with a hetheroscedastic one. A good candidate is the well-known multiplicative model of Harvey [4], which has been recently proposed in the context of international trade analysis by [2]. In this model, the error variance is modified according to the following equation:

$$\sigma_i^2 = \sigma^2 \exp(x_i'\alpha) \tag{3}$$

where α is a parameter to be estimated. When Harvey's model is used in combination with the Forward Search to analyze the mineral product data, the fit and the set of outliers detected change completely, as shown in Fig. 4. With the traditional model (Fig. 2) the part of the data retained by the Forward Search for the final fit is a strip located in the upper part of the scatterplot; the resulting prediction interval bands are essentially parallel and reflect the homoscedasticity assumption. With Harvey's model of Fig. 4, the fit is well in the center of the data and the outliers are separated from the good units by curvilinear bands, as it appears more clearly from the zoom in the right panel. However, the number of outliers is still evidently too large, meaning that the bands should open in a even more curvilinear way.

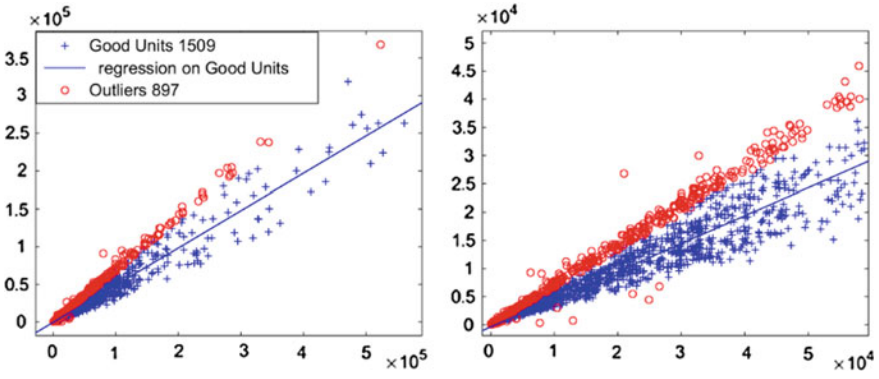


Fig. 4 Scatterplot of VALUE (in thousands of Euro) against QUANTITY (in tons) exchanged for the mineral product data. Linear model and outliers estimated by the Forward Search with Harvey's multiplicative model (*left panel*). On the right panel a zoom which emphasizes the curvilinear shape of the data along the regression fit

5 Monthly Fair Prices

In trade data the latent factor which often explains the presence of heteroscedasticity is the time factor. The price of a product is indeed subject to changes during time due to several reasons (inflation, technical improvements, seasonal effects). As a result, considering simultaneously data related to different time periods could produce multiple groups. On the other hand, heteroscedasticity could also be seen as a particular case of multiple populations. In fact, if the linear structures are so close to almost overlap, it would be difficult to distinguish their graphical representation from one originating from a single population with heteroscedastic error.

But if we analyze each single time period independently, then we should be able to isolate every linear structure, provided that the product is well and univocally defined and that there are no other latent factors. As an example, Fig. 5 shows in its left panel the scatterplot for the January data of the mineral product dataset. The data lie on an almost perfect line and there are no signs of heteroscedasticity. As a further confirmation, the right panel shows the corresponding forward White test plot. The trajectory confirms indeed the absence of heteroscedasticity: the curve exits from the 90 % band only in the last step of the Forward Search, which is indication of the presence of some outliers but not of heteroscedasticity. Moreover, now the Forward Search regression is able to capture the central tendency of the data and to give a reliable and robust estimate of the monthly price of the product. Scatterplots of other months show similar results.

Another advantage of the monthly approach is that it gives the possibility to study the trade price dynamics. This is relevant for detecting general patterns of economic interest such as unexpected peaks, seasonal components, market trends, and so on.

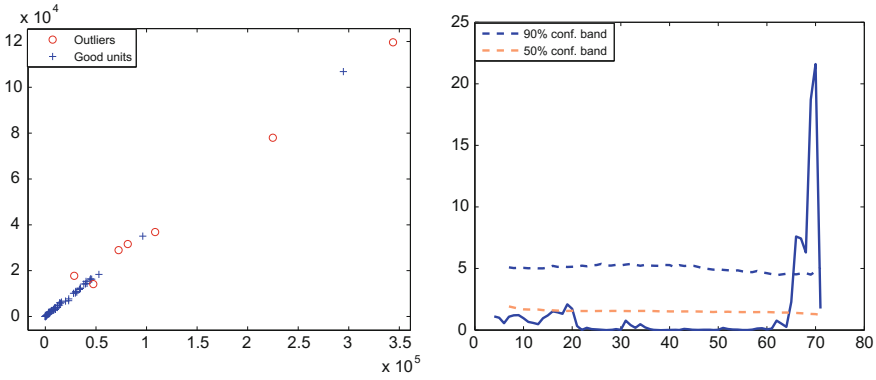


Fig. 5 EU imports of a mineral product, for January 2009 only. *Left panel* scatterplot of VALUE (in thousands of Euro) against QUANTITY (in tons) with outliers detected by the Forward Search. *Right panel* forward plot of the White test statistic

6 Discussion

We have briefly discussed some frequent issues in trade data analysis, which have to do with deviations from the standard model assumptions of single homoscedastic population and absence of outliers.

In practice, this issues complicates a lot the problem of estimating the fair import price using data in a reasonable time window (typically 3 years). For this reason, these issues can be avoided by considering the time component and fitting the data on a monthly basis. This approach is applicable on a considerable number of cases but not in general.

Therefore, the monitoring of the White test was used as an easy and automatic instrument to detect at least the presence of the above deviations in complete scans of trade data.

The more sophisticated monitoring of Harvey's multiplicative model was briefly demonstrated on a trade dataset. Although the results given by the classic model of Harvey are clearly closer to the user expectations, the distribution and number of outliers suggest that the model is not capturing perfectly the actual heteroscedastic component of these data. Improvements of the model in this context and demonstrations on different real datasets are proposed and recently discussed by [2]. Future work should include an extensive testing of such models on simulated data.

References

1. Atkinson, A.C., Riani, M.: Robust Diagnostic Regression Analysis. Springer, New York (2000)
2. Atkinson, A.C., Riani, M., Torti, F.: Robust Methods for Heteroskedastic Regression. Submitted (2014)
3. Cerioli, A., Perrotta, D.: Robust clustering around regression lines with high density regions. In: Advances in Data Analysis and Classification, vol. 8, pp. 5–26, Springer, Berlin (2014)
4. Greene, W.H.: Econometric Analysis, 7th edn. Prentice Hall, Upper Saddle River (2008)
5. ILO, IMF, OECD, Eurostat: United Nations Economic Commission for Europe, The World Bank: Export and Import Price Index Manual (2009)
6. Riani, M., Cerioli, A., Atkinson, A.C., Perrotta, D., Torti, F.: Fitting mixtures of regression lines with the forward search. In: Fogelman-Soulié, F., et al. (eds.) Mining Massive Data Sets for Security, pp. 271–286. IOS Press, Amsterdam (2008)
7. Riani, M., Perrotta, D., Torti, F.: FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. Chemom. Intell. Lab. Syst. **116**, 17–32 (2012)
8. Rousseeuw, P.J., Leroy, A.M.: Robust Regression and Outlier Detection. Wiley-Interscience, New York (1987)

How to Marry Robustness and Applied Statistics

Andrea Cerioli, Anthony C. Atkinson and Marco Riani

Abstract

A striking feature of most applied statistical analyses is the use of methods that are well known to be sensitive to outliers or to other departures from the postulated model. Since data contamination is often the rule, rather than the exception, we investigate the reasons for this contradictory (and perhaps unintended) choice. We also provide empirical evidence, in a real-world regression problem concerning international trade, of the advantages of a new approach to data analysis based on monitoring. Our approach enhances the applicability of robust techniques and the interpretation of their results, thus yielding a positive step towards a reconciliation between robustness and applied statistics.

1 Introduction

An early use of the term robustness is due to [5] in a study of the effect of non-normality on tests of equality of variances. He commented that means are robust to departures from normality, but that estimates of variances are not. The matter

A. Cerioli (✉) · M. Riani
Dipartimento di Economia, Università di Parma, Parma, Italy
e-mail: andrea.cerioli@unipr.it

M. Riani
e-mail: mriani@unipr.it

A.C. Atkinson
Department of Statistics, School of Economics and Political Science, London, UK
e-mail: a.c.atkinson@lse.ac.uk

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_6

is clearly important, since data frequently depart from the assumptions behind the models of mathematical statistics used to derive tests and other statistical procedures. Twenty years after Box, the outlines of the modern theory of robust statistics were becoming clearly established as the development of procedures that behaved well under small departures from standard assumptions, typically those of normality. This is a much narrower study than that implied by Box. One purpose of our paper is to show how the range of application of robust methods can be extended through the use of 'monitoring', exemplified in Sect. 3, where we study aspects of fitted models under a series of assumptions about the level of contamination in the data. An important byproduct is a simplification of the numerous choices required in the application of robust methods. We conclude with a discussion of problems that have mostly not been the subject of robust analysis. One example, treated in Sect. 4.1, is the identification of data that not only include outliers but in which groups of observations come from different models. But we start with a brief history of the development of robust methods.

Reference [38] gives a short history of robustness. The earliest book-length reference is [1] (the Princeton Robustness Study), at which time it was expected that all statistical analyses would, by default, be robust. Now, a further forty years on, there are at least 6 books about robust statistics with over 1,000 citations in Google Scholar. At the time of writing, the most highly cited is [23] (and its second edition [25]), the others, in citation order, are [1,18,21,28,36]. Unfortunately, this activity seems largely to be statisticians talking to other statisticians. Recent references with more applied emphasis, especially to health sciences, include [15,20] and the forthcoming book by [14]. However, the impact of robust methodologies in substantive domains still remains minor.

Although the term "robust" was first popularized by Box, the idea of considering the distribution of statistics under departures from the assumption of normality goes back at least to E.S. Pearson's review [29] of the second edition of Fisher's *Statistical Methods for Research Workers*, an interest that was to stay with Pearson until the end of his scientific life. The current understanding of robust was much more the creation of Tukey, starting with [39], and of [22]. Stigler writes:

"... by 1972 a number of the early workers in robust statistics expected that from the 1970s to 2000 we would see the same development with robust methods—extensions to linear models, time series, and multivariate models, and widespread adoption to the point where every statistical package would take the robust method as the default.... This was, and I [Stigler] will call it, a Grand Plan. But that plainly is not what has occurred".

Stigler then presents a lively and warm discussion of the early history of robust statistics. One reason for the lack of opening to the scientific world may be that robust statistics, as often understood and practised, has led to a new mathematical statistics, more complicated than the old, in which ever more refined solutions are presented to a few well-defined problems. We describe some of these complications in Sect. 2. From the standpoint of a user of statistical methods, the result of a robust analysis is to provide an alternative, for example for regression, to least squares. There are therefore two summaries of the data, rather than one. That this is not an especially

appealing development is evidenced by the failure of major commercial statistical packages to implement robust methods of data analysis except as special procedures within a well-segregated collection of routines. We appreciate that there are several robust libraries available in R, but would argue that, again, this package tends to be statisticians talking to statisticians.

Stigler suggests that the first signs of trouble with the Grand Plan were already evident in 1972 at the time of the publication of the Princeton Robustness Study. To quote him again:

"From the full set of 10,465 estimates of a location parameter they had considered, they reported in detail on the accuracy of 68 estimates that had received extensive study, focusing upon small samples and an inventively wide selection of 32 distributions, nearly all of which were symmetric scale mixtures of normal distributions".

This is far from the Grand Plan and, indeed, none of the authors of the conclusions in Chap. 7 of the Study made any grand claims for their work. Unlike the psalmist, they all display a compulsive refusal to lift their eyes to the hills, even for a moment; no Grand Plan is needed. But the year's work in Princeton by many intellectually impressive statisticians did not move far in solving the typical problems of data analysis mentioned in our first sentences. Indeed, your second author remembers the mounting despair with which a reading party organized by David Cox at Imperial College worked through the Study. We were quickly mired in the details of trying to remember what was, for example, an 'iteratively C-skipped trimean'. In Cox's recent book on applied statistics [13] the index contains just one reference to robustness. The relevant page carefully discusses the identification and treatment of outliers, stressing the comparative difficulty of the identification of multiple outliers and the importance of considering physical interpretation for any outliers found; points partially illustrated in the analysis of our example in Sect. 3. Likewise, the main reference to robustness in Huber's recent book on data analysis [24] downplays formal methods of robustness. In Sect. 5.3, 'Mathematical statistics and approximate models' Huber writes about the work of Fisher that, after Fisher "the robustness paradigm – explicitly permitting small deviations from the idealized model when optimizing – carried [the argument] only a few steps further". We hope to show that these works underestimate the contribution to intelligent data analysis that can be made by proper monitoring of the robust methods developed over 50 years since the study.

The most extreme forms of robustness usually considered are a very robust fit, asymptotically resistant to 50 % of aberrant observations, and maximum likelihood, including least squares, which have zero breakdown point. It is common [36,37] to suggest comparison of the residuals or Mahalanobis distances from such fits. In the approach illustrated in Sect. 3 we extend this idea, monitoring such quantities as residuals or distances, parameter estimates, test statistics and other quantities of interest as the robustness of the fit decreases. We thus obtain information on important changes in conclusions that come from differing assumptions about the degree of contamination in the data.

One consequence of our monitoring of robust procedures is that, by considering a variety of procedures for robust fitting, we are able to determine which, amongst

the many parameters of the algorithms, are those that are critical, distinguishing them from those that are only of secondary importance. The final goal is to provide insightful data analyses by following well-specified procedures that can be straightforwardly applied by non-specialists in robust statistics.

Our paper is structured as follows: in Sect. 2 we discuss the choice of an appropriate form of robust method, with an emphasis on regression, difficulties in numerical procedures and the interpretation of the results of a robust analysis. An important statistical drawback to downweighting methods, as opposed to trimming, is the breaking of the connection between each observational unit and quantities derived from the analysis, such as parameter estimates.

An example of monitoring is in Sect. 3 where we compare two methods of robust regression. One, S estimation, reveals that robust and non-robust fits to the data are very different; the other method, MM estimation, fails to do so, a finding in line with the conclusions of [33], who use monitoring to compare many different forms of robust regression.

As the quotation above from [24] indicates, standard robust methods have typically been developed under the assumption that there is a single model from which there are small departures, such as a slightly non-normal distribution of errors, perhaps together with gross outliers. This is only a slight part of the broad range of possible departures the data analyst may face. We indicate many such problems in Sect. 4, the theme of which is "robustness against what"? One important form of departure arises when the data are a mixture of observations from more than one model. For multivariate normal populations, this leads to problems of clustering. In Sect. 4.1 we continue the analysis of the regression data from Sect. 3, showing that they come from two different regression models. There are also a number of outliers. An important feature of robust clustering is that it is not necessary to cluster all observations. Our random start method based on the Forward Search does not require prior specification of the amount of trimming required, a feature it shares with the method of monitoring of Sect. 3. The subsequent section of the paper discusses some related issues that may contribute to discouraging the use of robust techniques, such as the difficulty in obtaining a reliable estimate of the number of outliers and the lack of knowledge about the empirical behaviour of the methods when the errors are very non-normal.

2 Which Method and How to Tune It?

A major disincentive to the routine use of standard robust methods is the number of decisions that have to be made before the analysis of the data begins. We now describe some of these.

1. The efficient application of robust methods depends on the proportion of outliers expected in the particular set of data being analysed. These determine the desired efficiency or, equivalently, breakdown point. Clearly, a very robust analysis can

always be used, but this results in an unnecessarily low efficiency for data that are virtually outlier free.
2. The next choice is the nature of robust estimator that is required. For regression [33] identify three classes of estimators:

   a. Hard (0,1) trimming such as Least Trimmed Squares - LTS: [17,35] or Least Median of Squares - LMS: [35] in which the amount of trimming is determined by the choice of the trimming parameter.
   b. Adaptive Hard Trimming. In the Forward Search (FS), the observations are again hard trimmed, but the amount of trimming is determined by the data, being found adaptively by the search. See [2,32] for regression and [4] for a general survey of the FS, with discussion.
   c. Soft trimming (downweighting). M estimation and derived methods, including weighted likelihood. The intention is that observations near the centre of the distribution essentially retain their value, but a suitable weight function ensures that increasingly remote observations have an effect on fitting that decreases with distance from the centre [25,27,28].

3. Within the soft trimming family, both the weight function, often called $\rho(\cdot)$, and the one or two parameters determining efficiency have to be chosen. Reference [33] use monitoring to compare three methods: S, MM and τ for four ρ functions: Tukey's bisquare, optimal, Hyperbolic and Hampel.

The calculations for robust estimation are also much more difficult than those of least squares. The functions to be maximized when using robust estimators are typically complicated, with many local maxima. In consequence, approximate methods are used. The standard approach uses randomly sampled subsets of p observations (elemental sets). We now list some of the choices that have to be made to provide a viable algorithm.

1. The number of subsamples to extract, to each of which the model is fitted exactly. These fitted values are used to evaluate the function to be maximized.
2. The maximum number of refining iterations (concentration steps), if any, within each subsample.
3. The tolerance for the convergence of the estimate of β in the refining steps.
4. The number of best subsets resulting from the refining steps to be brought to convergence.
5. The number of refining iterations for each best subset being brought to convergence.
6. The tolerance for the estimate of β in the refining steps for each subset being brought to convergence.
7. The tolerance for the estimate of scale in the best subsets.

In calculations for the example in Sect. 3 we follow the recommendations of the FSDA toolbox. Reference [19] show that inappropriate choices of some of these tuning constants may lead to inconsistency of the resulting algorithms.

Perhaps even more important than these technical matters, are the statistical problems related to, in particular, the downweighting of observations.

1. There is a loss of simplicity in the tests related to parameter estimates. In their Sect. 7.6, [25] point out that there may be alternative and equally plausible robust variants of the asymptotic standard errors of estimated regression coefficients. Reference [33] describe, and later exemplify, two such robust variants of the usual t-test, which sometimes differ in the conclusions they lead to. There is no guidance as to which is to be preferred in such circumstances.
2. Through the use of downweighting, the analyst loses the connection between each unit and the parameter estimates and other statistically important quantities. We note that this connection is maintained in the FS and other hard trimming methods.

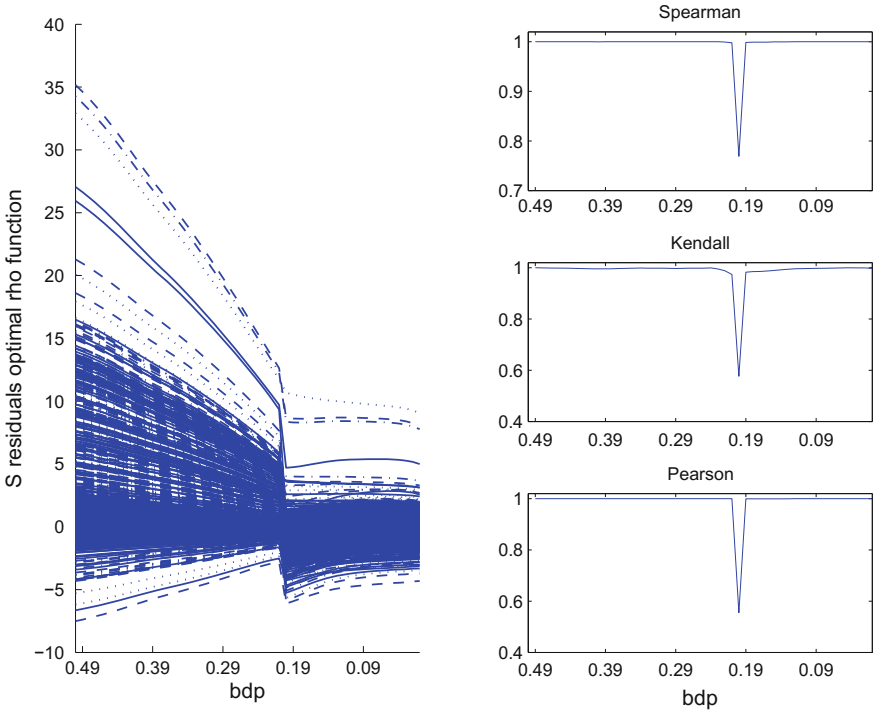


Fig. 1 Vegetable products data. S estimation, optimal ρ function. *Left-hand panel*, plot of scaled residuals. *Right-hand panel*, three measures of the correlations of adjacent residuals. The abrupt switch virtually to LS at 0.20 is evident in both panels

3 An Example of Monitoring

We illustrate the use of monitoring in the context of international trade, which is an important field of application for the EU economy. For instance, [8] describe the importance of careful statistical analysis of international trade data and some of the challenges emerging in such an exercise. The dataset that we consider contains the value and weight of $n = 1{,}558$ import transactions of vegetable products, such as oils and seeds, to one specific EU Member State from a non-EU country. To illustrate the usefulness of monitoring in understanding the properties of various robust estimators, we compare S and MM estimation. Typically we require 50 robust regression fits per analysis; a computational burden only made possible by the efficiency of the FSDA robust library [31] and by the recent technical advances of [34].

In monitoring S estimators we vary the bdp from 0.5 to 0.01. For MM estimates it is more convenient to monitor changes as the efficiency goes from 0.5 to 0.99. In both cases we look at plots of all n residuals as a function of efficiency or bdp. A useful diagnostic, summarizing the plot of residuals, is to plot correlation of the ranks between the residuals at adjacent monitoring values. We consider three standard measures of correlation:

1. Spearman. The correlations between the ranks of the two sets of observations.
2. Kendall. Concordance of the pairs of ranks.
3. Pearson. Product-moment correlation coefficient.

If there is a clear division of the solutions into a robust fit and a non-robust one, with a sharp break between them, this is clearly shown by the correlation plot. For more complicated examples the point of transition is not so clearly visible. But the structure of the residual plot is well summarized by looking at correlations.

Figure 1 shows the plot of residuals for S estimation. There is a clear break in the plot between bdp 0.21 and 0.20, as the robust fit changes to least squares. For the LS fit there seems to be an almost symmetrical distribution of residuals, with around half a dozen large positive outliers. The robust fit, for higher values of the bdp, exhibits a highly skewed structure for the residuals. The constancy of the ranking of the residuals over the two regimes is clearly shown in the right-hand panel of the plot; all three correlations are virtually one, except for the break point between the bdp of 0.21 and 0.20.

This figure is very different from that for MM estimation in Fig. 2. Here the pattern of residuals is constant for all efficiencies in the range studied and similar to that for the robust part of the S residuals in Fig. 1. The correlation plots show no change in the pattern.

These results show an appreciable difference between S estimation and MM, which is tuned to have a high efficiency for the parameters of the linear model. We now explore the parameter estimates of the linear model and their relationship with the data.

In these data there is a single explanatory variable. Figure 3 shows how the estimate of the slope changes with the bdp for S estimation and the efficiency for MM

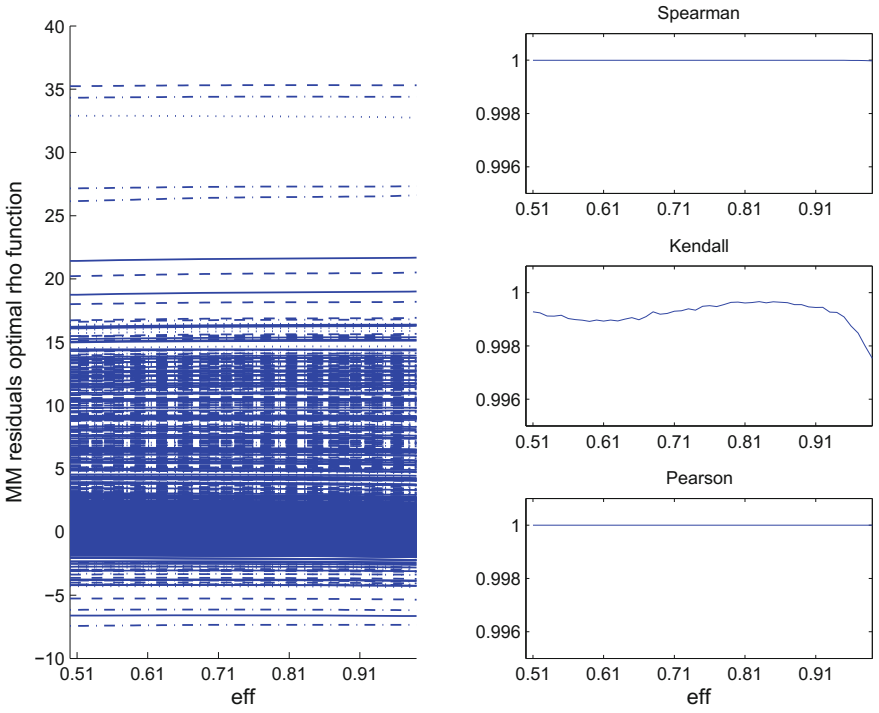


Fig. 2 Vegetable products data. MM estimation, optimal ρ function. *Left-hand panel*, plot of scaled residuals. *Right-hand panel*, three measures of the correlations of adjacent residuals. The skewed distribution of residuals remains constant over the considered range of efficiency. There is no change in the values of the correlation coefficients in the *right-hand panel* (note the vertical scale)

estimation. For S estimation the slope remains virtually constant, decreasing from 3.42 to 3.32, until, with a bdp of 0.20, it jumps up to 5.95. Thereafter it again decreases slightly, with a minimum value of 5.18. On the contrary, for MM estimation the slope decreases slowly from 3.40 to 3.32; the jump in values is missing.

The behaviour of the S slope that is revealed by monitoring is what might be expected if there is a main population following a regression line and a cluster of outliers at a position of high leverage. The right-hand panel of Fig. 3 shows five fitted lines for S estimation. Those for high bdp down to 0.21 follow the lower line of data. The fit calculated with bdp $= 0.20$ lies close to the upper line, for which there are more observations than for the lower one. As the bdp further decreases the lines move slightly towards lying between the two main lines, being attracted upwards by the presence of a few large disconnected outliers, some with appreciable leverage. The plot for MM lies throughout close to the lower line.

The conclusion from this analysis is that monitoring using S estimation alerts us to the presence of a structure in the data that would not be so trenchantly revealed by looking at the output from a single fit. Monitoring MM estimation, on the other

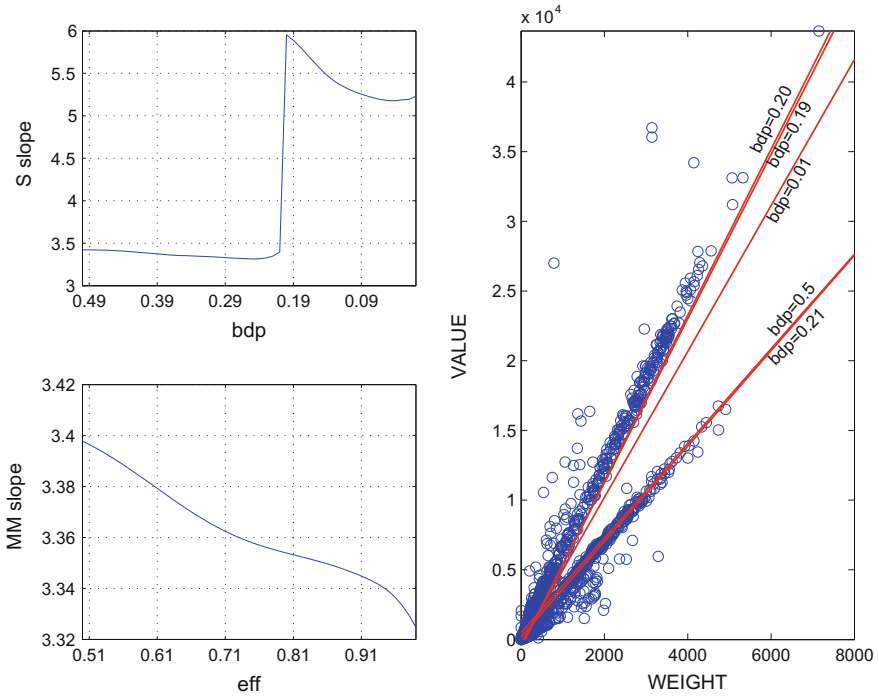


Fig. 3 Vegetable products data. *Left-hand panels*, the estimated slope parameters for *S* estimation (*upper panel*) and *MM* estimation (*lower panel*). *Right-hand panel*, the data and fitted lines using *S* estimation with five different values of *bdp*

hand, does not indicate that there is an important departure from the single model assumed to hold for the majority of the data. Perusal of Fig. 2 might, on the contrary, suggest that a transformation of the data is needed to achieve a symmetrical error distribution.

The results here from the comparison of S and MM estimation are in line with those of the extended study of this kind of monitoring by [33] who conclude that highly tuned methods like MM and τ estimation often reveal less about the structure of the data than does S estimation. Of the four ρ functions they compare, they show that Tukey's bisquare and the closely related optimal function provide the most informative monitoring. The hyperbolic ρ function, for some sets of data, is subject to numerical problems. Here we have used the optimal function.

We return to these data in the next section. Before we do so, we note that it might be expected that fitted lines for value against weight should go through zero. We did repeat our analysis setting the regression intercept to zero, but found that the conclusions were unaffected. Although, in some trading activities, there is a non-zero intercept, being the cost of setting up an order, such an effect is more common in domestic mail orders than in the kind of data we are analysing here.

4 Robustness Against What?

Standard robust methods were developed for fitting a single model. In this section we first describe a robust method for determining whether the data are a mixture from more than one model, although there is the restriction that the models are all of the same class. In the subsequent section we briefly discuss the more general, and far broader, problem of robustness when the class of model, or models, also needs to be identified.

4.1 Several Models: Clustering

The analysis of the trade data in Sect. 3 with monitoring shows that the robust S fit and least squares differ. However there is no clear indication of what is causing the difference. Of course, with a single explanatory variable, a simple scatterplot indicates the structure. But, in general, there may be several explanatory variables or so much data that perusal of individual scatterplots for all types of transaction is impossible. We use the FS to provide a robust analysis of data when there are several sources for the data. We need a robust method as we need to avoid the deleterious effect of the outliers, the presence of which is evident in the figure.

The forward search achieves robustness by fitting the model to subsets of the data of increasing size, where the subsets are sequentially chosen to contain observations as close as possible to the fitted model. The introduction of outliers into the subset is diagnostically revealed by plots of residuals against subset size as well as formally by statistically tuned tests using the minimum deletion residual among observations not in the subset. The method for a single population starts from a robustly chosen subset of m_0 observations. However, if the data are a mixture of observations generated from more than one model, the robustly chosen initial subset $S^*(m_0)$ may lead to a search in which observations from several models enter the subset haphazardly in such a way that the various models are not revealed. Searches from more than one starting point are necessary to reveal the more complicated structure of a mixture.

For finding clusters in multivariate data, [3] suggest running several hundred searches from randomly chosen initial subsets m_0. At the beginning of the search with regression models, a random start produces some very large residuals. But, because the search can drop units from the subset as well as adding them, some searches are attracted to specific regression lines. As the searches progress, the various random start trajectories converge, with subsets containing the same units. Once trajectories have converged, they cannot diverge again. As we see in Fig. 4, which is typical of those for many data structures, the search is rapidly reduced to relatively few trajectories, some of which show marked peaks. It is these that provide information on the number and membership of the clusters.

The two peaks in Fig. 4 indicate the two linear structures that are apparent in Fig. 3. The final peak in the plot results from the outliers, which are also evident in Fig. 3. The next step in the analysis of the data is to 'interrogate' the peaks, taking many of the units in the subset just before the peaks as large initial subsets for forward

searches to confirm cluster membership. The availability of automatic procedures for deciding cluster membership is an advance over many robust clustering procedures which require prior information on the number of clusters and on the proportion of the data to be trimmed, and so suffer from one of the main disadvantages of robust methods listed in Sect. 2.

A final word is in order about the interpretation of the forward plot of deletion residuals in Fig. 4. In all there are 1,558 observations. However, the two peaks come at $m = 1,174$ and 1,310, which total much more than all the observations. There are, however, an appreciable number of observations at low values of x; due to variability in the data, these could belong to either line. Straightforward clustering would be unable to decide to which line such observations should be allocated.

4.2 Which Model for the 'Good' Data and How Many Outliers?

The development of high-breakdown techniques, like S and MM estimation, has been the mainstream of theoretical work on robust statistics for at least 25 years. These methods are expected to work well in a contamination framework where the data generating distribution, say $G(y)$, is such that

$$G(y) = (1 - \gamma)G_0(y) + \gamma G_1(y). \tag{1}$$

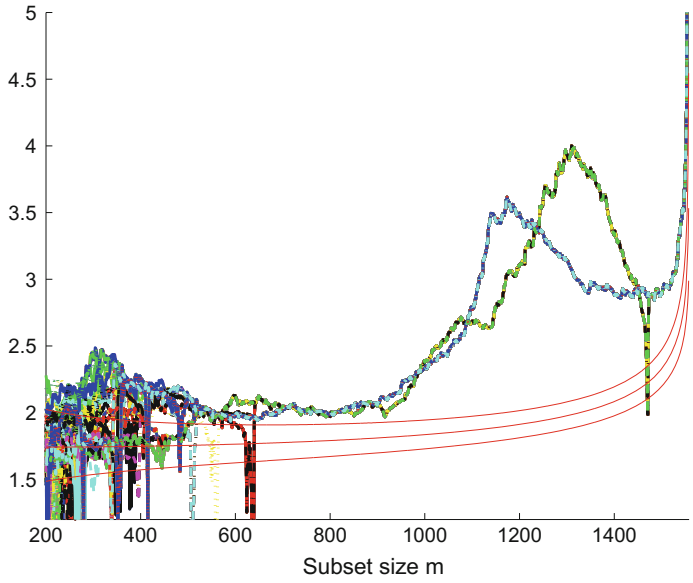


Fig. 4 Vegetable products data: forward plots of minimum deletion residuals from 200 random starts with pointwise 1 and 99 % limits. There appear to be two distinct groups (regression lines)

In model (1), $G_0(y)$ and $G_1(y)$ denote the distribution functions of the 'good' and of the contaminated part of the data, respectively, and $\gamma < 0.5$ is the unknown contamination rate.

We speculate that another reason for the limited appeal of robust methods in practical applications is the need to specify $G_0(y)$. Furthermore, very little is known about both the theoretical and empirical behaviour of the techniques when $G_0(y)$ is not normal. To motivate our claim, we observe that all high-breakdown estimators require computation of a normalizing constant which ensures consistency when $\gamma = 0$. In the case of hard trimming, this constant is a scaling factor for the estimate of dispersion and, in the case of soft trimming, a threshold above which observations are given zero weight. As far as we know, explicit and computable formulae for the normalizing constant exist only if $G_0(y)$ is the normal distribution and, indeed, relevant real-world applications have been confined to this model.

Reference [10] propose a method for testing the hypothesis that $G_0(y)$ in (1) is normal. The good power properties of their test seem to suggest that the empirical behaviour of high-breakdown techniques may be considerably different under non-normal models, especially when $G_0(y)$ is skewed. Furthermore, they show the potentially deleterious consequences of a naive approach to robustness which is often implemented in practice, when standard methods are applied to the observations that remain after outlier removal.

Even when $G_0(y)$ is the normal distribution, many high-breakdown procedures show poor finite sample properties for estimation of the contamination rate γ. The tendency to produce a plethora of spurious outliers has been shown in many studies, starting from [12] and including [9]. We argue that this tendency has also been a serious constraint on the dissemination of robust methods among practitioners. As a consequence, we strongly advocate the use of robust techniques that are able to provide effective control on the number of false discoveries, while keeping good detection properties. References [6,7] propose modified high-breakdown procedures that can achieve this goal, while [30,33] and this paper point towards a flexible monitoring approach.

5 Conclusion

We argue that there is compelling need for a reconciliation between robustness and applied statistics. In this paper we have investigated some of the reasons that we see as major disincentives to the routine use of standard robust methods. We have also provided empirical evidence, in a regression setting and in a real-world problem concerning international trade, of the advantages of a new approach to data analysis based on monitoring.

We conclude by noting that our monitoring approach deserves further theoretical investigation. A pioneering contribution in this direction, although in a somewhat simplified setting, is the study of the asymptotic properties of the radius process of [16]. Results for the forward search are provided by [11,26], while the properties of the trajectories of the residuals computed from other high-breakdown estimators,

like those given in Figs. 1 and 2, are still unexplored. Nevertheless, we trust that our work will provide a positive contribution towards the desired reconciliation.

Acknowledgments We thank the Scientific Program Committee of the 47th Scientific Meeting of the Italian Statistical Society for inviting us to present this work. We are also grateful to Dr. Domenico Perrotta of the European Commission Joint Research Centre at Ispra for providing the data on trade in vegetable products. Our work on this paper was partly supported by the project MIUR PRIN "*MISURA – Multivariate models for risk assessment*".

References

1. Andrews, D.F., Bickel, P.J., Hampel, F.R., Tukey, W.J., Huber, P.J.: Robust Estimates of Location: Survey and Advances. Princeton University Press, Princeton (1972)
2. Atkinson, A.C., Riani, M.: Robust Diagnostic Regression Analysis. Springer, New York (2000)
3. Atkinson, A.C., Riani, M., Cerioli, A.: Monitoring random start forward searches for multivariate data. In: Brito, P. (ed.) COMPSTAT, pp. 447–458. Physica-Verlag, Heidelberg (2008)
4. Atkinson, A.C., Riani, M., Cerioli, A.: The forward search: theory and data analysis (with discussion). J. Korean Stat. Soc. **39**, 117–134 (2010)
5. Box, G.E.P.: Non-normality and tests on variances. Biometrika **40**, 318–335 (1953)
6. Cerioli, A.: Multivariate outlier detection with high-breakdown estimators. J. Am. Stat. Assoc. **105**, 147–156 (2010)
7. Cerioli, A., Farcomeni, A.: Error rates for multivariate outlier detection. Comput. Stat. Data Anal. **55**, 544–553 (2011)
8. Cerioli, A., Perrotta, D.: Robust clustering around regression lines with high density regions. Adv. Data Anal. Classif. **8**, 5–26 (2014)
9. Cerioli, A., Riani, M., Atkinson, A.C.: Controlling the size of multivariate outlier tests with the MCD estimator of scatter. Stat. Comput. **19**, 341–353 (2009)
10. Cerioli, A., Farcomeni, A., Riani, M.: Robust distances for outlier-free goodness-of-fit testing. Comput. Stat. Data Anal. **65**, 29–45 (2013)
11. Cerioli, A., Farcomeni, A., Riani, M.: Strong consistency and robusteness of the forward search estimator of multivariate location and scatter. J. Multivar. Anal. **126**, 167–183 (2014)
12. Cook, R.D., Hawkins, D.M.: Comment on Rousseeuw and van Zomeren. J. Am. Stat. Assoc. **85**, 640–644 (1990)
13. Cox, D.R., Donnelly, C.A.: Principles of Applied Statistics. Cambridge University Press, Cambridge (2011)
14. Farcomeni, A., Greco, L.: Robust Methods for Data Reduction. Chapman and Hall/CRC, Boca Raton (2015)
15. Farcomeni, A., Ventura, L.: An overview of robust methods in medical research. Stat. Methods Med. Res. **21**, 111–133 (2012)
16. García-Escudero, L.A., Gordaliza, A.: Generalized radius processes for elliptically contoured distributions. J. Am. Stat. Assoc. **100**, 1036–1045 (2005)
17. Hampel, F.R.: Beyond location parameters: robust concepts and methods. Bull. Int. Stat. Inst. **46**, 375–382 (1975)
18. Hampel, F., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.: Robust Statistics. Wiley, New York (1986)
19. Hawkins, D.M., Olive, D.J.: Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm (with discussion). J. Am. Stat. Assoc. **97**, 136–159 (2002)

20. Heritier, S., Cantoni, E., Copt, S., Victoria-Feser, M.P.: Robust Methods in Biostatistics. Wiley, Chichester (2009)
21. Hoaglin, D.C., Mosteller, F., Tukey, J.W.: Understanding Robust and Exploratory Data Analysis. Wiley, New York (1983)
22. Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Stat. **35**, 73–101 (1964)
23. Huber, P.J.: Robust Statistics. Wiley, New York (1981)
24. Huber, P.J.: Data Analysis: What Can be Learned from the Past 50 Years. Wiley, New York (2011)
25. Huber, P.J., Ronchetti, E.M.: Robust Statistics, 2nd edn. Wiley, New York (2009)
26. Johansen, S., Nielsen, B.: Analysis of the forward search using some new results for martingales and empirical processes. Bernoulli **21** (2015, in press)
27. Markatou, M., Basu, A., Lindsay, B.G.: Weighted likelihood estimating equations with a bootstrap root search. J. Am. Stat. Assoc. **93**, 740–750 (1998)
28. Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust Statistics: Theory and Methods. Wiley, Chichester (2006)
29. Pearson, E.S.: Statistics in biological research. Nature **123**, 866–867 (1929)
30. Riani, M., Atkinson, A.C., Cerioli, A.: Finding an unknown number of multivariate outliers. J. R. Stat. Soc. Ser. B **71**, 447–466 (2009)
31. Riani, M., Perrotta, D., Torti, F.: FSDA: a MATLAB toolbox for robust analysis and interactive data exploration. Chemom. Intell. Lab. Syst. **116**, 17–32 (2012)
32. Riani, M., Atkinson, A.C., Perrotta, D.: A parametric framework for the comparison of methods of very robust regression. Stat. Sci. **29**, 128–143 (2014)
33. Riani, M., Cerioli, A., Atkinson, A.C., Perrotta, D.: Monitoring robust regression. Electron. J. Stat. **8**, 646–677 (2014)
34. Riani, M., Cerioli, A., Torti, F.: On consistency factors and efficiency of robust S-estimators. TEST **23**, 356–387 (2014)
35. Rousseeuw, P.J.: Least median of squares regression. J. Am. Stat. Assoc. **79**, 871–880 (1984)
36. Rousseeuw, P.J., Leroy, A.M.: Robust Regression and Outlier Detection. Wiley, New York (1987)
37. Rousseeuw, P.J., van Zomeren, B.C.: Unmasking multivariate outliers and leverage points. J. Am. Stat. Assoc. **85**, 633–639 (1990)
38. Stigler, S.M.: The changing history of robustness. Am. Stat. **64**, 277–281 (2010)
39. Tukey, J.W.: A survey of sampling from contaminated distributions. In: Olkin, I., et al. (eds.) Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, pp. 448–485. Stanford University Press, Palo Alto (1960)

Logistic Quantile Regression to Model Cognitive Impairment in Sardinian Cancer Patients

Silvia Columbu and Matteo Bottai

Abstract

When analyzing outcome variables that take on values within a finite bounded interval, standard analyses are often inappropriate. The conditional distribution of bounded outcomes given covariates is often asymmetric and bimodal (e.g., J- or U-shaped) and may substantially vary across covariate patterns. Analyzing this type of outcomes calls for specific methods that can constrain inference within the feasible range. The conditional mean is generally not an effective summary measure of a bounded outcome, and conditional quantiles are preferable. In this chapter we present an application of logistic quantile regression to model the relationship between Mini Mental State Examination (MMSE), a cognitive impairment score bounded between 0 and 30, with age and the results of a biochemical analysis (Oil Red O) for the determination of cytoplasmic neutral lipids in peripheral blood mononuclear cells in a sample of 124 cancer patients living in Sardinia, Italy. In addition we discuss an internal cross-validation method to optimally select the boundary correction in the logit transform.

S. Columbu (✉)
Dipartimento di Matematica e Informatica, Università di Cagliari, Cagliari, Italy
e-mail: silvia.columbu@unica.it

M. Bottai
Unit of Biostatistics, Institute of Environmental Health, Karolinska Institutet, Stockholm, Sweden
e-mail: matteo.bottai@ki.se

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_7

1 Introduction

Bounded outcomes are measurements that take on values on a known finite interval, which can be closed, open or half closed. Examples of bounded outcomes can be found in many research areas. Frequency distributions of this type of variables may assume a variety of shapes including unimodal, U-shape, and J-shape. To analyze bounded outcomes traditional statistical methods, such as least squares regression, mixed effects models, and even classic nonparametric methods, such as the Wilcoxon's test, may prove inadequate. Methods that constrain inference to lie within the feasible range of values should instead be considered. Reference [4] explored the use of a regression quantile model based on a logistic transformation of quantiles for values of the outcome at the boundaries of the range.

Quantile regression models conditional quantiles of the response variable. The basic idea dates back to the 18th century when Boscovich [3] introduced the criteria of minimization of the sum of absolute residuals to fit a median regression. More recent computational developments have encouraged the use and spread of this method. In 1959 Wagner [11] formulated the problem as a linear programming problem, and an efficient algorithm was introduced in 1973 by Barrodale and Roberts [2]. This regression method is becoming increasingly popular [6].

Compared with least squares regression quantile regression has numerous advantages: it makes no distributional assumptions about the regression error term, its inference is invariant to monotone transformations of the outcome variable, it is robust to outliers and it allows inference on the entire shape of the conditional distribution and not just the mean.

We used logistic quantile regression to analyze the relationship between a bounded outcome score and a set of covariates with data from the cancer therapy service of the University of Cagliari, Cagliari, Italy. We also investigated the use of a cross-validation algorithm to optimally define the boundary correction in the logit transform.

2 Logistic Quantile Regression

In this section we follow the description given by [4].

Consider a sample of n continuous observations $\{y_1, \ldots, y_n\}$ bounded from below and from above by two known constants y_{min} and y_{max}, and a set of s covariates $x = \{x_1, \ldots, x_s\}^T$. The p-th quantile of the conditional distribution of y_i given x_i is defined as $Q_y(p) = x_i^T \beta_p$. For example, if $p = 0.5$, $Q_y(0.5)$ indicates the conditional median.

We assume that for any pth quantile, with $p \in (0, 1)$, there exists a fixed set of parameters, $\beta_p = \{\beta_{p,0}, \beta_{p,1}, \ldots, \beta_{p,s}\}$, and a known nondecreasing function $h : (y_{min}, y_{max}) \to \mathbb{R}$ such that

$$h\{Q_y(p)\} = \beta_{p,0} + \beta_{p,1} x_1 + \cdots + \beta_{p,s} x_s.$$

The h function is usually called "link" function.

Because a continuous outcome bounded within the unit interval resembles a probability, or a propensity, among a variety of suitable choices for the link function h, Bottai et al. [4] opted for the logit transformation modified to constrain predictions in the feasible range (y_{min}, y_{max}). The selected function is defined as

$$h(y_i) = logit^*(y_i) = \log\left(\frac{y_i - y_{min}}{y_{max} - y_i}\right),$$

with inverse

$$Q_y(p) = \frac{\exp(\beta_{p,0} + \beta_{p,1}x_1 + \cdots + \beta_{p,s}x_s)y_{max} + y_{min}}{\exp(\beta_{p,0} + \beta_{p,1}x_1 + \cdots + \beta_{p,s}x_s) + 1}.$$

The logit transform permits interpreting the regression coefficient $\beta_{p,j}$, $j = 1, \ldots, s$, as a quantile-specific odds ratio. Logistic regression has been widely used in applications for analyzing the mean of categorical outcome variables as an alternative to the method of discriminant linear analysis. Similarly, logistic quantile regression can be seen as an alternative to linear quantile regression in the analysis of continuous bounded outcomes.

The parameter β_p can be estimated using quantile regression by regressing the transformed outcome $h(y_i)$ on x

$$Q_{h(y_i)}(p) = Q_{logit^*(y_i)}(p) = x_i^T\beta_p$$

The parameters estimates derive from the quantile minimization problem

$$\hat{\beta}_p = \min_{\beta\in\mathbb{R}^q}\sum_{i=1}^{n}\rho_p(h(y_i) - x_i^T\beta_p) = \min_{\beta\in\mathbb{R}^q}\sum_{i=1}^{n}\rho_p(logit^*(y_i) - x_i^T\beta_p),$$

where $\rho_p(u) = u(p - \mathbb{I}(u \leq 0))$ is a piecewise loss function and $\mathbb{I}$ is the indicator function.

The small- and large-sample properties of the estimator for β_p are the same as those of the quantile regression estimator of the non-transformed dependent variable y. Under assumption of *i.i.d.* errors the asymptotic distribution of the quantile estimator, as shown by Koenker and Bassett in 1978 [8], is normal with covariance matrix $\omega^2(p)(x^Tx)^{-1}$ where $\omega^2(p)$ denotes the quantity $p(1-p)/f^2(F^{-1}(p))$ and $f^2(F^{-1}(p))$ is the density of the error distribution evaluated at the pth quantile. That is, under the same conditions, the limiting behavior of the quantile estimator is similar to the behavior of the ordinary least squares estimator. Here the variance σ^2 of the underlying error distribution is replaced by the quantity ω.

It has been shown [5] that the boostrap resampling technique has some advantage over asymptotic approximations. In the application in the next Section, we therefore opted for the use of the bootstrap. Inference on estimates was based on the assumption that the sampling distribution is approximately normal and simple t-tests were calculated to evaluate the significance of parameters.

Once estimates for the regression coefficients β_p are obtained, inference on $Q_y(p)$ can then be made through the inverse transform. This is possible because of the property of invariance of quantiles to monotone transformations, $Q_{h(y)}(p) = h\{Q_y(p)\}$, which is not shared by the mean.

3 Modeling Mini Mental State Examination

Between September 2009 and April 2012, a total of 124 patients (66 females, 53 % and 58 males, 47 %) with solid tumors were admitted to the day hospital of anticancer therapy service of University of Cagliari. All patients received at least one previous chemotherapy regimen and were evaluated during chemotherapy cycles. Data on age and gender were obtained from questionnaires. Clinical information was obtained from medical charts. Blood sampling was performed during chemotherapy cycles. The age range was 29–94 years. The data collection for this study was approved by the Ethics Committee of the Cagliari University School of Medicine, and all subjects provided written informed consent before participating in this study.

The Mini Mental State Examination (MMSE) measured the participants' global cognitive status. MMSE assess orientation with respect to place and time, short-term memory, episodic long-term memory, ability to perform subtraction and construct a sentence, and oral language ability. MMSE is a questionnaire-based score bounded between 0 and 30. A score of 30 points indicates no cognitive impairment, and a score of 0 maximum cognitive impairment. Subjects with a MMSE score <24 are typically considered cognitive impaired.

We applied logistic quantile regression to make inference about quantiles of MMSE.

The covariates considered in the study were sex, age, presence of metastasis and a binary variable based on the result of a biochemical test performed to determine the concentration of citoplasmic neutral lipids in peripheral blood mononuclear cells. Oil Red O (ORO) [9] is a lipid-soluble dye which stains neutral lipids, including esterified cholesterol but not free cholesterol. It appears as bright red spots in the cytoplasm. The two levels of the variable ORO used in our analysis represent the red intensity scored on a semi-quantitative scale: 1 indicates an intense diffuse staining and higher concentration of neutral lipids and 0 a lower intensity of coloration in cells.

Our research interest was to study the behavior of patients with cognitive deficit, corresponding to lower values of MMSE, and investigate if cognitive impairment for cancer patients corresponded to higher concentration of neutral lipids in the brain [1]. We therefore decided to make inference on lower percentiles of the distribution of MMSE via logistic quantile regression.

We defined

$$logit^*_\varepsilon(MMSE) = \log\left(\frac{MMSE + \varepsilon}{30 - MMSE + \varepsilon}\right), \tag{1}$$

where ε was a small quantity added to ensure that the logit transform was defined for all values of MMSE.

We built three logistic quantile regression models corresponding to the percentiles $p \in \{0.1, 0.25, 0.50\}$. The fitted models were the following

$$Q_{logit^*_\varepsilon(MMSE)}(p) = \beta_{p,0} + \beta_{p,1} age + \beta_{p,2} ORO + \beta_{p,3} sex + \beta_{p,4} metastasis.$$

Because of the equivariance of quantiles to monotone transformations the constant ε in the logit function can be set as any value, and it should be selected to ensure

that the assumption of linearity in the model is met. We selected the constant ε based on a measure of goodness of fit. Given a set of possible ε values we chose the one that minimized the loss function that defines the quantile regression problem at any fixed p:

$$GOF = \min_{\varepsilon} \sum_{i=1}^{n} \{[logit_{\varepsilon}^{*}(MMSE_i) - x_i^T \beta_p][\omega_i - p]\}, \quad (2)$$

where $\omega_i = I(logit_{\varepsilon}^{*}(MMSE_i) \leq x_i^T \beta_p)$, $i = 1, \ldots, n$.

4 Results and Discussion

The analyses were performed with the statistical software R. We estimated logistic quantile regression with the *rq* function of the *quantreg* library [7] after logit-transforming the outcome. For a Stata command see Orsini and Bottai [10].

Patients baseline characteristics, reported in Table 1, were compared across the two ORO groups by Fisher's exact test for the categorical variables. For the continuous variables differences in the distributions were tested by Wilcoxon's rank-sum test.

The distribution of sex (P-value = 0.80) and that of metastasis groups (P-value = 0.61) did not significantly differ between the ORO levels, while that of MMSE and age did.

Figure 1 shows the boxplots of MMSE in the two ORO categories. The observed distribution of MMSE differed between the two groups, and patients with intense diffusion stain ($ORO = 1$) showed lower values of MMSE.

These preliminary descriptive analyses suggested an association between MMSE and ORO categories.

We applied logistic quantile regression to estimate the percentiles $p \in \{0.1, 0.25, 0.5\}$ of the conditional distribution of MMSE.

As discussed in Sect. 3 we considered a numerical criteria for the choice of the ε constant to be considered in the argument of the logit transform. The dependent variable MMSE is a score outcome. We assume that MMSE is the rounded value of a latent continuous variable $MMSE^*$. Its relationship with the observed values satisfies $MMSE - 0.5 \leq MMSE^* \leq MMSE + 0.5$. Predicted values of the proposed

Table 1 Descriptive characteristics of the study's participants

Characteristics	ORO 0 (N = 105)	ORO 1 (N = 19)	P-value
Female sex (no. %)	55 (52)	11 (58)	0.80
Metastasis (no. %)	64 (61)	10 (53)	0.61
Age (mean sd)	62.87 ± 11.4	68.05 ± 8.65	<0.001
MMSE (mean sd)	27.83 ± 2.26	25.37 ± 3.24	<0.001

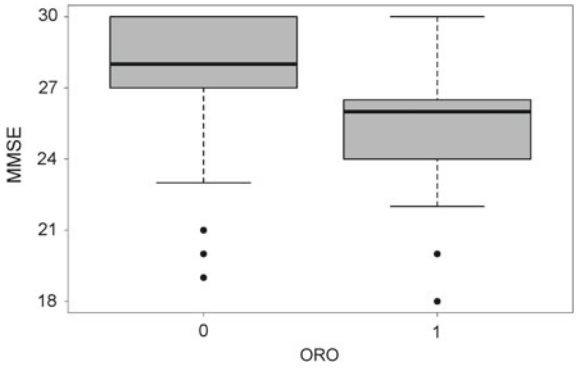


Fig. 1 Boxplot of Mini Mental State Examination (MMSE) by ORO's categories. $ORO = 0$ corresponds to a lower intensity of coloration in peripheral blood mononuclear cells; $ORO = 1$ corresponds to an intense diffuse staining. Patients in the $ORO = 1$ group show lower values of MMSE

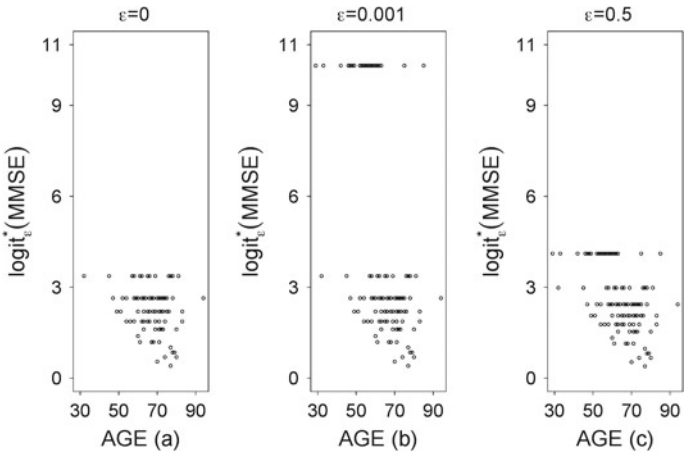


Fig. 2 Distribution of $logit^*_\varepsilon(MMSE)$ against age with the value of the constant ε set to 0 (panel **a**), to 0.001 (**b**), and to 0.5 (**c**)

model are in a continuous scale in the range $(MMSE_{min} - 0.5, MMSE_{max} + 0.5)$. We selected the constant ε based on a grid search over the interval from 0 to 0.5.

The goodness of fit criteria showed that for the three percentiles considered the best ε in the logit transform was 0.5. This conclusion could have also been taken after observing that, as shown in Fig. 2, for higher values of ε the distribution of $logit^*_\varepsilon(MMSE)$ against the continuous covariate age tended to be closer to that in which no constants, e.g. $\varepsilon = 0$, were added in the logit transform.

The explanatory variables in Table 1 were initially all included as covariates. Sex and metastasis were then removed because not statistically significant. Their

Table 2 Estimates of coefficients of the logistic quantile regression model for the 10th, the 25th percentile and the median of the logit transform of MMSE. Standard errors, confidence intervals and P-values were estimated with 1000 bootstraps samples

		Coefficients	Std error	t value	P-value	CI
$p = 0.10$	*Intercept*	4.33	0.61	7.05	0.00	(3.12, 5.53)
	$ORO = 1$ versus $ORO = 0$	−0.47	0.28	−1.71	0.09	(−1.02, 0.07)
	Age	−0.04	0.01	−4.18	<0.001	(−0.06, −0.02)
$p = 0.25$	*Intercept*	4.30	0.77	5.59	0.00	(2.79, 5.81)
	$ORO = 1$ versus $ORO = 0$	−0.66	0.23	−2.86	0.005	(−1.12, −0.21)
	Age	−0.03	0.01	−2.88	0.005	(−0.06, −0.01)
$p = 0.50$	*Intercept*	5.69	0.92	6.18	0.00	(3.89, 7.50)
	$ORO = 1$ versus $ORO = 0$	−0.75	0.20	−3.81	<0.001	(−1.14, −0.37)
	Age	−0.05	0.01	−3.65	<0.001	(−0.07, −0.02)

inclusion did not improve the goodness of fit for any of the percentiles considered and the estimates of the coefficients for age and ORO remained nearly unchanged.

The final model was

$$Q_{logit^*_{0.5}(MMSE)}(p) = \beta_{p,0} + \beta_{p,1} age + \beta_{p,2} ORO$$

which included ORO and age as predictors, and $\varepsilon = 0.5$ in the logit transform. Standard errors, confidence intervals and P-values were estimated with 1000 bootstraps samples [5]. The estimated coefficients for the three percentiles considered are shown in Table 2.

In the final models all the estimates of the regression coefficients were statistically significant for the $25th$ percentile and the median, while the estimate of ORO was not significant for the 10th percentile.

We were not interested in the average MMSE value, but rather in modeling the lower tail of the distribution. Because the dataset was quite small the information on the 10th percentile was insufficient. MMSE score was associated with ORO and with age for the 25th percentile and the median of the distribution. The interpretation of the regression coefficients was analogous to the interpretation of the coefficients of a logistic regression for binary outcomes. The adjusted logit for the 25th percentile of the MMSE score was estimated to be 0.66 lower in the group of individuals with $ORO = 1$ and decreased also with age with a difference of 0.03 for each year. The exponential of the coefficient estimate ($exp(-0.66) = 0.52$) represents the 25th

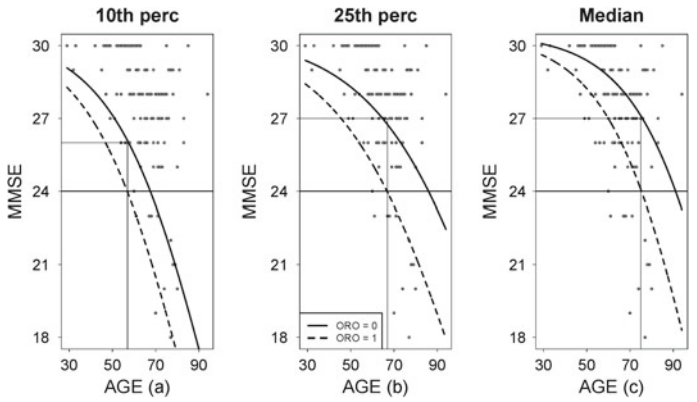


Fig. 3 MMSE distribution against age and predicted transformed values of logistic quantile regression for the 10th (panel **a**), the 25th (**b**), and the 50th percentile (**c**). The *solid line* represents the predicted quantile in the ORO = 0 group and the dashed line the predicted quantile in the ORO = 1 group

percentile odds ratio (OR) of MMSE score in patients with $ORO = 1$ versus $ORO = 0$. Something analogous can be said for the median where the adjusted logit of MMSE was 0.75 lower when $ORO = 1$ and decreased with age with a difference of 0.05 per year.

Patients with a MMSE < 24 were considered cognitive impaired.

A summary of the inference from the three models is showed in Fig. 3. MMSE score decreased along with age. Among patients aged >67 years, 25 % of those with $ORO = 1$ had MMSE values below the cut-point of 24 while 75 % of patients with $ORO = 0$ were above a MMSE score of 27 (Fig. 3b). Among patients that were >75 years old 50 % of individuals with $ORO = 1$ had a MMSE score lower than the threshold value, and 50 % of individuals with $ORO = 0$ had MMSE higher or equal to 27 (Fig. 3c). The figure relative to the 10th percentile did not add any information to the interpretation of the results (Fig. 3a).

5 Conclusions and Remarks

Our findings suggest that lower quantiles of MMSE were associated with high intensity of ORO staining, independently on the pathological cancer status of patients. Specifically, we observed that a high concentration of neutral lipids in peripheral mononuclear blood cells was associated with cognitive impairment and that older patients tended to have altered MMSE.

The use of logistic quantile regression allowed drawing a detailed picture of medical behavior for patients with altered cognitive functions while respecting the

bounded nature of the response variable. Thanks to the equivariance property of quantiles, modeling the dependence on the covariates was relatively easy.

The cross-validation criteria, defined as in (2), proved to be a valid and useful additional tool to reduce the uncertainty and arbitrarity related to the introduction of a correction constant when defining the logit transform (1) in the data.

The quantiles of discrete bounded outcomes, such as MMSE, are also discrete and should be modeled as a continuous function of a set of covariates [4]. Discrete bounded outcomes, however, can often be seen as the discretize version of a latent continuous variable. This generally facilitates the interpretation of the predictive values.

References

1. Anchisi, L., Dessí, S., Pani, A., et al.: Neutral lipid determination in peripheral blood mononuclear cells: a useful tool for diagnostic and therapeutic interventions in dementia. J. Mol. Biomark. Diagn. **3**(6), 3–6 (2012)
2. Barrodale, I., Roberts, F.D.K.: An improved algorithm for discrete L_1 linear approximation. S.I.A.M. J. Numer. Anal. **10**, 839–848 (1973)
3. Boscovich, R.J.: De Litteraria Expeditione per Pontificiam Ditionem De Litteraria Expeditione per Pontificiam Ditionem. Bononiensi Scientiarum et Artum Instituto Atque Academia Commentarii **4**, 353–396 (1757)
4. Bottai, M., Cai, B., McKeown, R.E.: Logistic quantile regression for bounded outcomes. Stat. Med. **29**, 309–317 (2010)
5. Buchinsky, M.: Estimating the asymptotic covariance matrix for quantile regression models: a Monte Carlo study. J. Econometrics **68**, 303–308 (1995)
6. Koenker, R.: Quantreg: Quantile Regression, R package version 4.76 (2011)
7. Koenker, R.: Quantile Regression. Econometric Society Monograph Series. Cambridge University Press, Cambridge (2005)
8. Koenker, R., Bassett, G.: Regression quantiles. Econometrica **46**(1), 33–50 (1978)
9. Mandas, A., Congiu, M., Abete, C., Dessí, S., Manconi, P., Musio, M., Columbu, S., Racugno, W.: Cognitive decline and depressive symptoms in late-life are associated with statin use: evidence from a population-based study of Sardinian old people living in their own home. Neurol. Res. **36**(3), 247–254 (2014)
10. Orsini, N., Bottai, M.: Logistic quantile regression in Stata. Stat. J. **11**, 327–344 (2011)
11. Wagner, W.H.: Linear programming techniques for regression analysis. J. Am. Stat. Assoc. **54**, 206–212 (1959)

Bounding the Probability of Causation in Mediation Analysis

A. Philip Dawid, Rossella Murtas and Monica Musio

Abstract

Given empirical evidence for the dependence of an outcome variable on an exposure variable, we can typically only provide bounds for the "probability of causation" in the case of an individual who has developed the outcome after being exposed. We show how these bounds can be adapted or improved if further information becomes available. In addition to reviewing existing work on this topic, we provide a new analysis for the case where a mediating variable can be observed. In particular, we show how the probability of causation can be bounded when there is no direct effect and no confounding.

1 Introduction

Many statistical analysis aim at a causal explanation of the data. In particular, in epidemiology many studies are conducted to try to understand if and when an exposure will cause a particular disease. Also in a Court of Law, when we want to assess

A.P. Dawid (✉)
Department of Pure Mathematics and Mathematical,
University of Cambridge, Cambridge, UK
e-mail: apd@statslab.cam.ac.uk

R. Murtas · M. Musio
Department of Mathematics and Computer Science,
University of Cagliari, Cagliari, Italy
e-mail: apd@statslab.cam.ac.uk

M. Musio
e-mail: mmusio@unica.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_8

legal responsibility we usually refer to causality. But when discussing this topic it is important to specify the exact query we want to talk about. For example it may be claimed in court that it was Ann's taking the drug that was the cause of her death. This type of question relates to the cause of an observed effect ("CoE") and is fundamental to the allocation of responsibility. On the other hand, much of classical statistical design and analysis, for example randomized agricultural or medical experiments, has been created to address questions about the effects of applied causes ("EoC"). When we address an EoC query, we are typically asking a hypothetical question: "What would happen to Ann if she were to take the drug?". At the very same time we can address alternative hypothetical questions: "What would happen to Ann if she were not to take the drug?".

Assessing the effects of causes can be achieved in straightforward fashion using a framework based on probabilistic prediction and statistical decision theory [2]. To formalize the problem, let X be a binary decision variable denoting whether or not Ann takes the drug, and Y the response, coded as 1 if she dies and 0 if not. We denote by P_1 [resp. P_0] the probability distribution of Y ensuing when X is set to the value 1 [resp. 0]. The two distributions P_1 and P_0 are all that is needed to address EoC-type queries: I can compare these two different distributions for Y, decide which one I prefer, and take the associated decision.

The situation is different for a CoE query, where the drug has already been taken and the outcome observed. A natural way to address a CoE question would be to try to imagine what would have happened to Ann had she not taken the drug. In other words, given the fact that Ann actually took the drug and died, how likely is it that she would not have died if she had not taken the drug? We can not address a CoE query using only the two distribution P_1 and P_0. In fact, we can no longer base our approach purely on the probability distribution of Y and X conditioned on known facts, since we know the values of both variables ($Y = 1$, $X = 1$), and after conditioning on that knowledge there is no probabilistic uncertainty left to work with. Nevertheless we want an answer. This query can be approached by introducing (for any individual) an associated pair of "potential responses" $\boldsymbol{Y} := (Y(0), Y(1))$, where $Y(x)$ denotes the value of the response Y that will be realized when the exposure X is set to x (which we write as $X \leftarrow x$). Both potential responses are regarded as existing, simultaneously, prior to the choice of X, the actual response Y then being determined as $Y = Y(X)$. However, for any each individual just one of the potential responses will be observable. For example, only $Y(1)$ will be observable if in fact $X \leftarrow 1$; $Y(0)$ will then be *counterfactual*, because it relates to a situation, $X \leftarrow 0$, which is contrary to the known fact $X \leftarrow 1$.

To address the court's query we use the formulation of *Probability of Causation*, *PC* as given by [5] who names it *Probability of Necessity*. In terms of the triple $(X_A, Y_A(0), Y_A(1))$, we define the Probability of Causation in Ann's case as

$$PC_A = P_A(Y_A(0) = 0 \,|\, X_A = 1, Y_A(1) = 1), \tag{1}$$

where $P_A(\cdot)$ denotes the probability distribution over attributes of Ann. Knowing that Ann did take the drug ($X_A = 1$) and the actual response was recovery ($Y_A(1) = 1$), this is the probability that the potential response $Y_A(0)$, that would have been observed

Table 1 Deaths in individuals exposed and unexposed to the same drug taken by Ann

	Die	Live	Total
Exposed	30	70	100
Unexposed	12	88	100

had Ann not taken the drug, would have been different ($Y_A(0) = 0$). But how are we to get a purchase on this quantity?

Suppose that a good experimental study, in which subjects were randomly assigned to be either exposed ($X = 1$) or unexposed ($X = 0$), tested the same drug taken by Ann, and produced the data reported in Table 1.

Since our analysis here is not concerned with purely statistical variation due to small sample sizes, we take proportions computed from this table as accurate estimates of the corresponding population probabilities (but see [3] for issues related to the use of small-sample data for making causal inferences). Thus we take

$$Pr(Y = 1 \mid X \leftarrow 1) = 0.30$$
$$Pr(Y = 1 \mid X \leftarrow 0) = 0.12,$$

where we use Pr to denote population probabilities.

We see that, in the experimental population, individuals exposed to the drug ($X \leftarrow 1$) were more likely to die than those unexposed ($X \leftarrow 0$), by 18 percentage points. So can the court infer that it was Ann's taking the drug that caused her death? More generally: Is it correct to use such experimental results, concerning a population, to say something about a single individual? This "Group-to-individual" (G2i) issue is discussed by [3] in relation to the question "When can Science be relied upon to answer factual disputes in litigation?". It is pointed out that in general we cannot obtain a point estimate for PC_A, but we can provide useful information, in the form of bounds between which this quantity must lie.

In this paper, we show how these bounds can be adapted or improved when further information is available. In Sect. 2 we consider the basic situation where we only have information on exposure and outcome. In Sect. 3 we bound the probability of causation when we have additional information on a pretreatment covariate. Section 4 considers the situation in which unobserved variables confound the exposure–outcome relationship. Finally in Sect. 5 we introduce new bounds for PC when a mediating variable can be observed. Section 6 presents some concluding comments.

2 Starting Point: Simple Analysis

In this section, we discuss the simple situation in which we have information, as in Table 1, from a randomized experimental study. We need to assume that the fact of Ann's exposure, X_A, is independent of her potential responses $\boldsymbol{Y}_A$:

$$X_A \perp\!\!\!\perp \boldsymbol{Y}_A. \tag{2}$$

Property (2) parallels the "no-confounding" property $X_i \perp\!\!\!\perp \boldsymbol{Y}_i$ which holds for individuals i in the experimental study on account of randomization. We further suppose that Ann is exchangeable with the individuals in the experiment, i.e., she could be considered as a subject in the experimental population.

On account of (2) and exchangeability, (1) reduces to $PC_A = Pr(Y(0) = 0 \mid Y(1) = 1)$, but we cannot fully identify this from the data. In fact, we can never observe the joint event $(Y(0) = 0; Y(1) = 1)$, since at least one of $Y(0)$ and $Y(1)$ must be counterfactual. In particular, we can never learn anything about the dependence between $Y(0)$ and $Y(1)$. However, even without making any assumptions about this dependence, we can derive the following inequalities [4]:

$$1 - \frac{1}{RR} \leq PC_A \leq \frac{Pr(Y = 0 \mid X \leftarrow 0)}{Pr(Y = 1 \mid X \leftarrow 1)} \tag{3}$$

where

$$RR = \frac{Pr(Y = 1 \mid X \leftarrow 1)}{Pr(Y = 1 \mid X \leftarrow 0)}$$

is the *experimental risk ratio* between exposed and unexposed. These bounds can be estimated from the experimental data using the population death rates computed in Sect. 1.

In many cases of interest (such as Table 1), we have

$$Pr(Y = 1 \mid X \leftarrow 0) < Pr(Y = 1 \mid X \leftarrow 1) < Pr(Y = 0 \mid X \leftarrow 0).$$

Then the lower bound in (3) will be nontrivial, while the upper bound will exceed 1, and hence be vacuous.

We see from (3) that, whenever $RR > 2$, the Probability of Causation PC_A will exceed 50 %. In a civil court this is often taken as the criterion to assess legal responsibility "on the balance of probabilities" (although the converse is false: it would not be correct to infer $PC_A < 0.5$ from the finding $RR < 2$). Since, in Table 1, the exposed are 2.5 times as likely to die as the unexposed ($RR = 30/12 = 2.5$), we have enough confidence to infer causality in Ann's case: we have $0.60 \leq PC_A \leq 1$.

3 Additional Covariate Information

In this section, we show how we can refine the bounds of (3) if further information about a pretreatment covariate S is available. For example, S might be a gene, possession of which enhances the dangerous effect of exposure to the drug. We now take the assumptions of Sect. 2 to hold after conditioning on S (indeed in cases where the original assumptions fail, it may well be possible to reinstate them by conditioning on a suitable covariate S). In particular, $X_A \perp\!\!\!\perp \boldsymbol{Y}_A \mid S_A$, and $X_i \perp\!\!\!\perp \boldsymbol{Y}_i \mid S_i$: adjusting for S is enough to control for confounding, both for Ann and in the study.

3.1 Fully Observable

Consider first the situation where we can observe S both in the experimental data and in Ann. In this case, (1) should be replaced by the more specific definition

$$PC_A = P_A(Y_A(0) = 0 \mid X_A = 1, Y_A(1) = 1, S_A = s_A)$$

where s_A is Ann's value for S. We can apply the analysis of Sect. 2, after conditioning on S, to obtain the estimable lower bound

$$1 - \frac{1}{RR\,(s_A)} \leq PC_A,$$

where

$$RR\,(s) = \frac{Pr(Y = 1 \mid X \leftarrow 1, S = s)}{Pr(Y = 1 \mid X \leftarrow 0, S = s)}.$$

3.2 Observable in Data Only

But even when we can only observe S in the population, and not in Ann, we can sometimes refine the bounds in (3). Thus suppose S is binary, and from the data we infer the following probabilities (which in particular imply the same values as given in Table 1):

$$\begin{aligned} P_A(S = 1) &= 0.50 \\ P_A(Y = 1 \mid X \leftarrow 1, S = 1) &= 0.60 \\ P_A(Y = 1 \mid X \leftarrow 0, S = 1) &= 0 \qquad (4) \\ P_A(Y = 1 \mid X \leftarrow 1, S = 0) &= 0 \qquad (5) \\ P_A(Y = 1 \mid X \leftarrow 0, S = 0) &= 0.24. \end{aligned}$$

Since we know $X_A = 1$ and $Y_A = 1$, from (5) we realize we cannot have $S_A = 0$, so we must have $S_A = 1$. Then from (4) we see that, when we set X to 0, we can not obtain $Y = 1$, so we must have $Y_A(0) = 0$. That is, in this special case we can infer causation in Ann's case—even though we have not directly observed her value for S.

More generally [1] we can refine the bounds in (3) as follows:

$$\frac{\Delta}{Pr(Y = 1 \mid X \leftarrow 1)} \leq PC \leq 1 - \frac{\Gamma}{Pr(Y = 1 \mid X \leftarrow 1)}$$

where

$$\Delta = \sum_s Pr(S = s) \times \max\{0, Pr(Y = 1 \mid X \leftarrow 1, S = s) - Pr(Y = 1 \mid X \leftarrow 0, S = s)\}$$

and

$$\Gamma = \sum_s Pr(S = s) \times \max\{0, Pr(Y = 1 \mid X \leftarrow 1, S = s) - Pr(Y = 0 \mid X \leftarrow 0, S = s)\}$$

These bounds are never wider than those obtained from (3), which ignores S.

Table 2 Observational data

	Die	Live	Total
Exposed	18	82	100
Unexposed	24	76	100

4 Unobserved Confounding

So far we have assumed no confounding, $X \perp\!\!\!\perp \boldsymbol{Y}$ (perhaps conditionally on a suitable covariate S), both for Ann and for the study data. Now we drop this assumption for Ann. Then the experimental data cannot be used, by themselves, to learn about $PC_A = P(Y_A(0) = 0 \mid X_A = 1, Y_A(1) = 1)$.

We might however be able to gather additional *observational* data, where there was no possibility of experimental control over subjects' exposure, X, which might thus be related to unobserved personal aspects affecting the response Y. However—importantly—we now assume that the dependence between X and $\boldsymbol{Y}$ for subjects in the sampled population is just the same as it is for Ann. Let Q denote the joint observational distribution of (X, Y), which is estimable from such data. Reference [7] obtain the following bounds for PC_A, given both experimental and observational data:

$$\max\left\{0, \frac{Q\,(Y=1) - Pr(Y=1 \mid X \leftarrow 0)}{Q\,(X=1, Y=1)}\right\} \leq PC_A \leq \min\left\{1, \frac{Pr(Y=0 \mid X \leftarrow 0) - Q\,(X=0, Y=0)}{Q\,(X=1, Y=1)}\right\}. \tag{6}$$

For example, suppose that, in addition to the data of Table 1, we have observational data as in Table 2.

Thus

$$\begin{aligned} Q\,(Y=1) &= 0.21 \\ Q\,(X=1, Y=1) &= 0.09 \\ Q\,(X=0, Y=0) &= 0.38. \end{aligned}$$

Also, from Table 1 we have $Pr(Y = 1 \mid X \leftarrow 0) = 0.12$ (so $Pr(Y = 0 \mid X \leftarrow 0) = 1 - 0.12 = 0.88$). From (6) we thus find $1 \leq PC_A \leq 1$. We deduce that Ann would definitely have survived had she not taken the drug.

5 Mediation Analysis

In this section, we bound the Probability of Causation for a case where a third variable, M, is involved in the causal pathway between the exposure X and the outcome Y. Such a variable is called a *mediator*. In general, the total causal effect of X on Y can

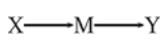


Fig. 1 Directed Acyclic Graph representing a mediator M, responding to exposure X and affecting response Y. There is no "direct effect," unmediated by M, of X on Y

be split into two different effects, one mediated by M (the *indirect effect*) and one not so mediated (the *direct effect*). Here we shall only consider the case of no direct effect, as intuitively described by Fig. 1. An application in which such an assumption is plausible is in the treatment of ovarian cancer [6], where X represents management either by a medical oncologist or by a gynaecological oncologist, M is the intensity of chemotherapy prescribed, and Y is death within 5 years.

We shall be interested in the case that M is observed in the experimental data but is not observed for Ann, and see how this additional experimental evidence can be used to refine the bounds on PC_A.

To formalize our assumption of "no direct effect", we introduce $M(x)$, the potential value of M for $X \leftarrow x$, and $Y^*(m)$, the potential value of Y for $M \leftarrow m$, where the irrelevance of the value x of X to Y^* encapsulates our assumption that X has no effect on Y over and above that transmitted through its influence on the mediator M. The potential value of Y for $X \leftarrow x$ (in cases where there is no intervention on M, which we here assume) is then $Y(x) := Y^*\{M(x)\}$.

In the sequel, we restrict to the case that all variables are binary, and define $\boldsymbol{M} := (M(0), M(1))$, $\boldsymbol{Y}^* := (Y^*(0), Y^*(1))$, and $\boldsymbol{Y} := (Y(0), Y(1))$. In particular, we have observable variables $(X, M, Y) = (X, M(X), Y(X))$. We denote the bivariate distributions of the potential response pairs by

$$\begin{aligned} m_{ab} &:= Pr(M(0) = a, M(1) = b) \\ y^*_{rs} &:= Pr(Y^*(0) = r, Y^*(1) = s) \\ y_{rs} &:= Pr(Y(0) = r, Y(1) = s). \end{aligned}$$

Then

$$\begin{aligned} m_{a+} &= Pr(M = a \mid X \leftarrow 0) \\ m_{+b} &= Pr(M = b \mid X \leftarrow 1) \\ y^*_{r+} &= Pr(Y = r \mid M \leftarrow 0) \\ y^*_{+s} &= Pr(Y = s \mid M \leftarrow 1) \\ y_{r+} &= Pr(Y = r \mid X \leftarrow 0) \\ y_{+s} &= Pr(Y = s \mid X \leftarrow 1), \end{aligned}$$

where m_{a+} denotes $\sum_{b=0}^{1} m_{ab}$, *etc.*

In addition to the assumptions of Sect. 2 we further suppose that none of the causal mechanisms depicted in Fig. 1 are confounded—expressed mathematically by assuming mutual independence between X, $\boldsymbol{M}$ and $\boldsymbol{Y}^*$ (both for experimental individuals, and for Ann). Then, m_{a+}, m_{+b}, y^*_{r+}, y^*_{+s}, y_{r+}, y_{+s} are all estimable from experimental data in which X is randomized, and M and Y are observed.

It is also easy to show the Markov property:

$$Y \perp\!\!\!\perp X \mid M.$$

This observable property can serve as a test of the validity of our conditions.

The assumed mutual independence implies

$$\begin{aligned} y_{rs} &= Pr(Y^*(M(0)) = r, Y^*(M(1)) = s) \\ &= \sum_{a,b=0}^{1} Pr(Y^*(a) = r, Y^*(b) = s) Pr(M(0) = a, M(1) = b). \end{aligned}$$

This yields

$$\begin{aligned} y_{00} &= m_{00} y^*_{0+} + (m_{01} + m_{10}) y^*_{00} + m_{11} y^*_{+0} \\ y_{01} &= m_{01} y^*_{01} + m_{10} y^*_{10} \\ y_{10} &= m_{01} y^*_{10} + m_{10} y^*_{01} \\ y_{11} &= m_{00} y^*_{1+} + (m_{01} + m_{10}) y^*_{11} + m_{11} y^*_{+1}, \end{aligned}$$

and

$$y_{r+} = m_{0+} y^*_{r+} + m_{1+} y^*_{+r} \tag{7}$$

$$y_{+s} = m_{+0} y^*_{s+} + m_{+1} y^*_{+s}. \tag{8}$$

Suppose now that we observe $X_A = 1$ and $Y_A = 1$, but do not observe M_A. We have

$$PC_A = \frac{y_{01}}{y_{+1}} = \frac{m_{01} y^*_{01} + m_{10} y^*_{10}}{y_{+1}}. \tag{9}$$

The denominator of (9) is $Pr(Y = 1 \mid X \leftarrow 1)$, which is estimable from the data.

As for the numerator, this can be expressed as

$$2\mu\eta + A\mu + B\eta + AB = 2(\mu + B/2)(\eta + A/2) + AB/2 \tag{10}$$

with $\mu = m_{01}$, $\eta = y^*_{01}$, $A = y^*_{+0} - y^*_{0+}$, and $B = m_{+0} - m_{0+}$. Note that A and B are identified from the data, whereas for μ and η we can only obtain inequalities:

$$\begin{aligned} \max\{0, -B\} \le \mu \le \min\{m_{0+}, m_{+1}\} \\ \max\{0, -A\} \le \eta \le \min\{y^*_{0+}, y^*_{+1}\}, \end{aligned}$$

so that

$$\begin{aligned} |B/2| \le \mu + B/2 \le \min\{\tfrac{1}{2}(m_{0+} + m_{+0}), \tfrac{1}{2}(m_{1+} + m_{+1})\} \\ |A/2| \le \eta + A/2 \le \min\{\tfrac{1}{2}(y^*_{0+} + y^*_{+0}), \tfrac{1}{2}(y^*_{1+} + y^*_{+1})\}. \end{aligned} \tag{11}$$

The lower [resp., upper] limit for (10) will be when $\mu + B/2$ and $\eta + A/2$ are both at their lower [resp., upper] limits. In particular, the lower limit for (10) is $\max\{0, AB\}$. Using (7) and (8), we compute $AB = y_{+1} - y_{1+}$, which leads to the lower bound

$$PC_A \ge 1 - \frac{Pr(Y = 1 \mid X \leftarrow 0)}{Pr(Y = 1 \mid X \leftarrow 1)} = 1 - \frac{1}{RR},$$

exactly as for the case that M was not observed. Thus the possibility to observe a mediating variable in the experimental data has not improved our ability to lower bound PC_A.

Table 3 Upper bound for numerator of (9)

	$m_{1+} + m_{+1} \geq 1$	$m_{1+} + m_{+1} < 1$
$y^*_{1+} + y^*_{+1} \geq 1$	$m_{0+}y^*_{0+} + m_{+0}y^*_{+0}$	$m_{1+}y^*_{+0} + m_{+1}y^*_{0+}$
$y^*_{1+} + y^*_{+1} < 1$	$m_{0+}y^*_{+1} + m_{+0}y^*_{1+}$	$m_{1+}y^*_{1+} + m_{+1}y^*_{+1}$

We do however obtain an improved upper bound. Taking into account the various possible choices for the upper bounds in (11), the upper bound for the numerator of (9), in terms of experimentally estimable quantities, is given in Table 3.

It can be shown that this upper bound is never greater than that in (3), which ignores the mediator M, and is strictly smaller unless $y^*_{1+} + y^*_{+1} \geq 1$ and $m_{1+} + m_{+1} = 1$.

5.1 Example

Suppose we obtain the following values from the data:

$$\begin{aligned} Pr(M = 1 \mid X \leftarrow 1) &= 0.25 \\ Pr(M = 1 \mid X \leftarrow 0) &= 0.025 \\ Pr(Y = 1 \mid M \leftarrow 1) &= 0.9 \\ Pr(Y = 1 \mid M \leftarrow 0) &= 0.1. \end{aligned}$$

Again, these imply the values given in Table 1. Then we find $0.60 \leq PC_A \leq 0.76$; whereas without taking account of the mediator we would have no nontrivial upper bound.

6 Discussion

In this paper, we have considered estimation of the Probability of Causation in a number of contexts, including a novel analysis for the case of a mediating variable in the absence of a direct effect. As we saw in Sect. 5, considering such a third variable in the pathway between exposure and outcome will lead to an improved upper bound, although conclusions about the lower bound remain the same.

The next step will be to generalize our analysis to more general cases of mediation, allowing for the possibility of a direct effect, and for unobserved confounding.

References

1. Dawid, A.P.: The role of scientific and statistical evidence in assessing causality. In: Goldberg, R. (ed.) Perspectives on Causation, pp. 133–147. Hart Publishing, Oxford (2011)
2. Dawid, A.P.: Statistical causality from a decision-theoretic perspective. Ann. Rev. Stat. Appl. **2**, 273–303 (2015)
3. Dawid, A.P., Fienberg, S., Faigman, D.: Fitting science into legal contexts: assessing effects of causes or causes of effects? Sociol. Methods Res. **43**, 359–390 (2014)
4. Dawid, A.P., Musio, M., Fienberg, S.E.: From statistical evidence to evidence of causality. Bayesian Analysis. Advance Publication, 26 August 2015. http://projecteuclid.org/euclid.ba/1440594950 (2014)
5. Pearl, J.: Probabilities of causation: three counterfactual interpretations and identification. Synthese **121**, 93–149 (1999)
6. Silber, J.H., Rosenbaum, P.R., Polsky, D., Ross, R.N., Even-Shoshan, O., Schwartz, J.S., Armstrong, K.A., Randall, T.C.: Does ovarian cancer treatment and survival differ by the specialty providing chemotherapy? J. Clin. Oncol. **10**, 1169–1175 (2007)
7. Tian, J., Pearl, J.: Probabilities of causation: bounds and identification. Ann. Math. Artif. Intell. **28**, 287–313 (2000)

Analysis of Collaboration Structures Through Time: The Case of Technological Districts

Maria Rosaria D'Esposito, Domenico De Stefano and Giancarlo Ragozini

Abstract

In the present work we propose to analyze, through Multiple Factorial Analysis (MFA), the relational structure embedded in collaboration networks observed across time occasions. We show, through a case study, how the solutions provided by the MFA can be interpreted in a suitable way in the relational setting which arises in complex and heterogeneous networks. Valuable information about the strength and typology of the collaboration structure and its evolution can be obtained. As case study, we analyze a Technological District located in South Italy.

1 Introduction

Collaboration among private companies, institutions, and public research organizations is a topic of growing importance in the agendas of both research and development (R&D) policymakers and governmental administrators. Most policies are based, either in the design and decision phase or in the implementation phase, on a

M.R. D'Esposito (✉)
Dipartimento di Scienze Economiche e Statistiche, Università di Salerno, Fisciano, Italy
e-mail: mdesposi@unisa.it

D. De Stefano
Dipartimento di Scienze Politiche e Sociali, Università di Trieste, Trieste, Italy
e-mail: ddestefano@units.it

G. Ragozini
Dipartimento di Scienze Politiche, Università di Napoli Federico II, Naples, Italy
e-mail: giragoz@unina.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_9

network of agents who concur in the policies [9] to create a critical mass of firms, research labs, and universities.

In the last decade, in Italy the efforts have been directed in promoting firms innovation capability through their systemic aggregations to foster R&D activities on key technologies. To this aim, the Italian government, through the Minister of University and Research, MIUR, has created the so-called Technological Districts (TDs) in some carefully chosen geographic locations in the national territory.[1] TDs may be defined as geographical concentration of interconnected companies and associated institutions including end-producers, universities, research laboratories, and service providers, all focused on a specialized area of economic activity.

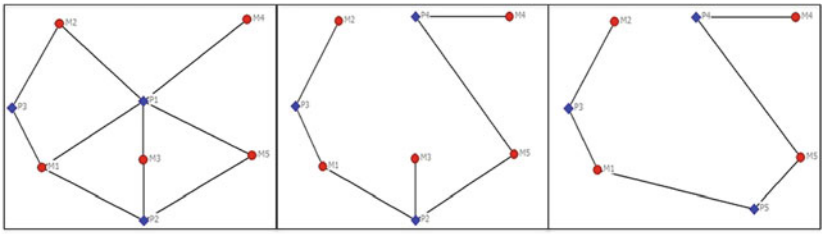
Often, when the interest is on monitoring the collaboration among the TDs members, it is worth to consider the former as a collaboration network. In such a case, it has to be considered that the collaboration networks under study are complex structures in which each collaboration tie can be observed across time occasions that span from the birth of the analyzed TD until its current state. There are several possibilities to measure collaboration in terms of relational ties; among them we focus on the joint participation to research projects. This latter gives rise to two-mode networks (actors-by-events). When considering TDs collaboration structures we observe a time-varying two-mode networks (actors-by-events under different types of relationships across time occasions). Relational data observed in such conditions can be organized into multidimensional arrays, and statistical methods from the theory of multiway data analysis [3,8] may be exploited to reveal the underlying data structure. Among these latter, in the current work propose to explore the relational structure of collaboration emerging in a TD setting through the use of Multiple Factor Analysis (MFA). We show how Social Network Analysis (SNA) techniques jointly with MFA, can be fruitfully used to obtain valuable information about the strength and typology of the collaboration structure among TD members. We focus the attention on the collaboration network of a TD located in south Italy.

2 MFA for Network Data: Main Concepts

Two-mode networks are particular networks that consist of N actors (the first mode) and J events to which they participate (the second mode) and are represented by a bipartite graph or by a so-called affiliation matrix.

A time-varying two-mode network is a two-mode network observed along K occasions (such as time points) in which the N actors are fixed, whereas the events depend on the occasion. That is, at the k-th occasion we observe J_k events. In Fig. 1 we depict a two-mode network referring to the same set of actors that participate to different (or even the same) events in each of the $K = 3$ occasions.

[1] http://hubmiur.pubblica.istruzione.it/web/ricerca/ricerca-internazionale/technological-district.



	Occasion 1			Occasion 2			Occasion 3		
	P1	P2	P3	P2	P3	P4	P3	P4	P5
M1	1	1	1	1	1	0	1	0	1
M2	1	0	1	0	1	0	1	0	0
M3	1	1	0	1	0	1	0	1	0
M4	1	0	0	0	0	0	0	0	0
M5	1	1	0	1	0	1	0	1	1

Fig. 1 Example of a grand table $\mathbf{F}_k$ for affiliation networks of five actors (from M1 to M5, *red circles*) observed over three occasions (K = 3). Actors are fixed, whereas the event occurrences (from P1 to P5, *blue diamonds*) depend on occasion k

For each k we have a binary affiliation matrix $\mathbf{F}_k = (f_{ijk})$, $i = 1, \ldots, N$, $j = 1, \ldots, J_k, k = 1, \ldots, K$, with $f_{ijk} = 1$ if the i-th actor attends the j-th event at the occasion k, and 0 otherwise. A grand table F can be built up by stacking each subtables $\{\mathbf{F}_k\}_{k=1,\ldots,K}$ side by side (Fig. 1).

In the present paper, we look at the relational structure embedded not only in each $\mathbf{F}_k$, but also in F. To this end, we propose to use MFA, which provides a unifying and general framework to deal with multiple-way matrix, like F. MFA is an extension of factorial techniques [1,5,6] tailored to handle multiple data tables. This allows to jointly analyze quantitative and qualitative variables, providing displays in which representations of the set of individuals associated to each group of variables are superimposed. By applying MFA to time-varying affiliation networks, we can perform four different analyses [11]: (i) analysis of each $\mathbf{F}_k, k = 1, \ldots, K$ through a suitable factorial method (partial analysis); (ii) analysis of F (global analysis); (iii) analysis of structural changes among occasions; (iv) analysis of actor/event variation over the occasion by projecting the weighted affiliation matrices $\mathbf{F}_k$ on the global factorial plane. To perform MFA, a factorial method has to be chosen for the analysis of both $\mathbf{F}_k$ and F. We choose the use of Multiple Correspondence Analysis (MCA) because of its nice properties in the analysis of network data [4,11].

Relational patterns at each occasion can be visually analyzed by: (i) *representing events in the actor space*: each event in the actor space is represented by two opposite vectors, corresponding to the two poles, lying on the same direction and passing through the origin. The cosine of the angle between two event segments is the "correlation" between event attendance patterns; (ii) *representing actors in the event space*: each actor corresponds to a point and proximity between two actors means they have similar participation patterns. Actors corresponding to points close to the axes' origin have a common participation habit, while actors with corresponding points far from the center have unusual participation patterns; (iii) *jointly representing actors*

and events: In order to represent actors and events in a joint two-dimensional map we can use the asymmetric biplot [7]. The direction vector defined by each event is the biplot axis. By projecting the actor points onto each biplot axis, we can appreciate approximately their event participation profiles. This allows us to characterize actors' closeness or farness in terms of event participation.

In the MFA approach, actors with similar participation patterns in all the occasions will be located close together on the factorial planes. Events can be represented, for all the occasions, on the same factorial plane by looking at the correlations both with respect to the events on the same occasion and with respect to the events referring to different occasions.

3 The Collaboration Structure of the TD Under Analysis

3.1 The Data

Our case study refers to a TD established in 2006 and located in Campania Region. It represents a Knowledge Integrator that designs and develops specific network mechanisms to promote links between scientific research organizations and private companies mainly in the field of engineering of polymeric materials and composites. This case is interesting as it refers to an Italian successful story of technological development in an economic poor context. The focus on the collaboration structure is related to the explicit TD mission to design and developed specific network mechanism to foster collaboration at local, national, and international level. Then, here we look at scientific collaboration networks given by the joint participation in R&D projects.

The data we are working on is a proprietary data set collected in the time period from 2007 to 2013, and refer to research projects lasting from one to three years which involve only TD members. Following the setting in Sect. 2, a member of the TD is an actor; an R&D project at time k is an event, and the participation of a member to a project at time k gives rise to a collaboration tie, with $i = 1, \ldots, 52$, $j = 1, \ldots, J_k, k = 1, \ldots, 7$, and $J_1 = 12$, $J_2 = 13$, $J_3 = 14$, $J_4 = 12$, $J_5 = 11$, $J_6 = 14$, $J_7 = 12$ for a total of $J = 88$ events. The 52 TD members consist of 11 private companies, 19 Research Centers, 16 University Departments, and 6 other institutions.

In previous works [2,4] the collaboration structure arising among the TD member have been analyzed considering the time span as whole without looking at evolution over time. Figure 2 displays the bipartite graph for the participation of TD members to research projects over the whole time span. Apart from few projects (e.g., ITALNANONET and TECOP) which are characterized by the exclusive participation of groups of members, the patterns of participation to all the other project is not so clear. Furthermore, the graph does not tell the whole story as it does not incorporate the

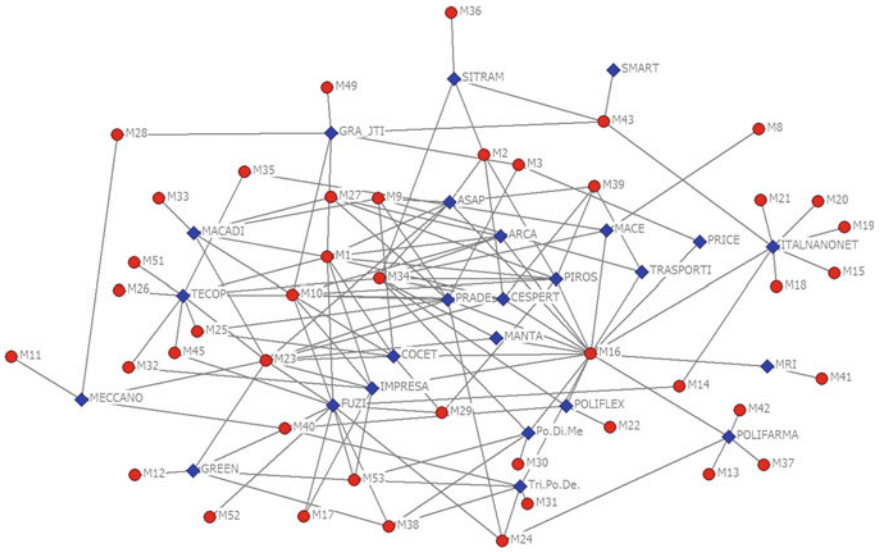


Fig. 2 Bipartite graph for the participation of TD Members (*red circles*) in the granted R&D research projects (*blue diamonds*)

dynamic perspective and is not able to highlight changes in collaboration patterns. In this paper, we focus on this latter aspect and we aim at investigating if the collaboration networks are active and if the participation to scientific projects is changing and evolving over time. Using MFA based on MCA we can detect differences over time, and highlight actor participation and event attendance trajectories over the time levels and the structural changes with respect to time occasions.

3.2 Analysis of the Affiliation Structure over Time

We can use the properties of MCA applied to affiliation networks [4] to interpret the results in both actor and event space. For the sake of presentation we consider only the global analysis. In order to have an idea of the overall quality of 2D approximation of the MFA solution, we look at the unadjusted and adjusted values of the proportion of inertia explained by the first two factorial axes [11]. With respect to unadjusted proportion, the adjusted one increases moving from 40.5 to 76.9 %. In the actor space, the angles between event segments are of interest as they express the similarities of the attendance patterns (the smaller the angle, the greater the similarity in attendance pattern between two projects). For instance, from Fig. 3a, looking at the angle between segments (and using event attributes), it is possible to identify groups of projects referring to: Aerospace sector, Transport sector, and Chemical

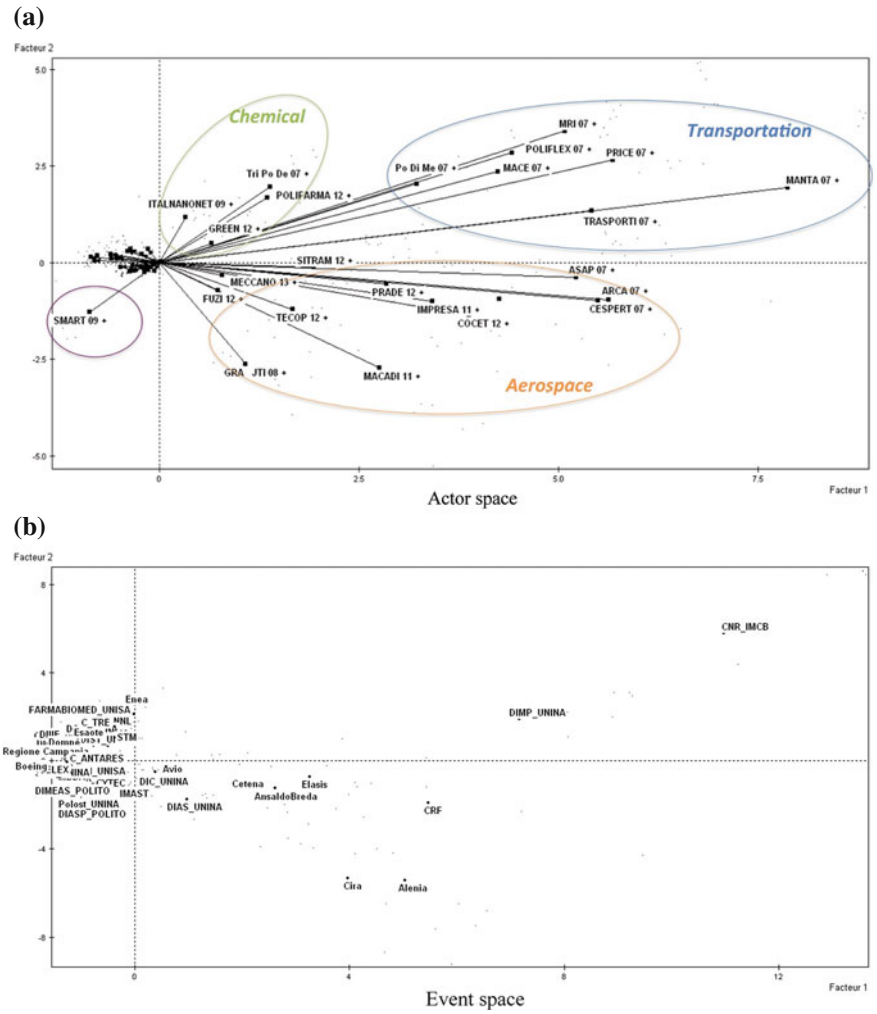


Fig. 3 Global analysis: **a** Events' representation in the actor space for all the occasions, events' attributes are used to interpret clusterized events; **b** actors' representation in the event space. Each point represents one actor and its coordinates are the weighted average of the coordinates over the three years span. Actor and project labels are hidden

sector. This means that over the years these groups of projects are mainly attended by the same TD members. Only one project results very different from the others, namely the SMART project.

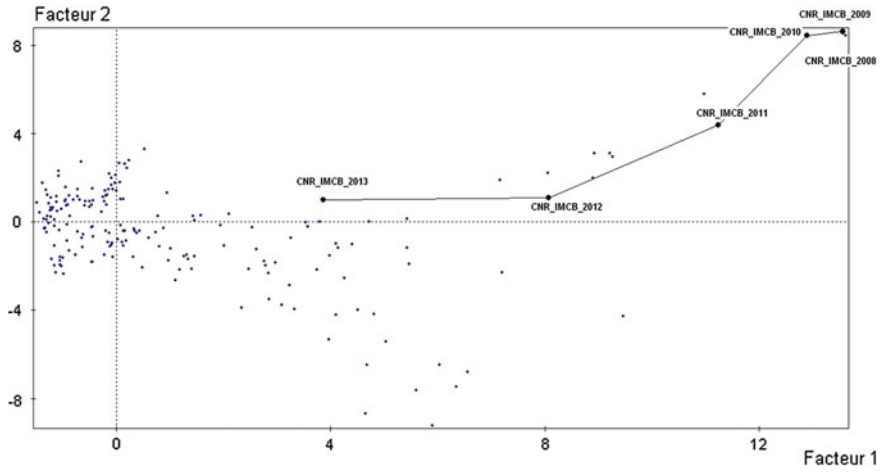


Fig. 4 Global analysis: Actors' representation in the event space. Each point represents one actor at a given time points. Example of actor CNR-IMCB trajectory in the time occasions is highlighted

In the event space (Fig. 3b), actors are placed close to each other to the extent they have similar relational patterns. In this space, TD members are grouped with respect to their rate of participation (first axis) and the type of projects (second axis). For example, Alenia and CIRA have high joint participation in aerospace projects (bottom right), while DIMP-UNINA and CNR-IMCB present high participation rate in transportation projects (top right).

We can also interpret the combination of the partial analyses (for each of the K time occasions) by means of the so-called trajectories. An example is depicted in Fig. 4. Here we draw the line connecting points that represent in each partial analysis the actor CNR-IMCB$_k$, for $k = 1, \ldots, 7$. In such trajectory drawn for exemplification, the actor moves from participation to nonparticipation over time. Indeed, the projects in which the actor was involved from the beginning ended, and the actor was not able to enter in new projects, due also to change in project topics.

In analogy with the actor trajectories, the actor (event) attribute trajectories could be represented. MFA allows to use attributes in place of the individual actors to interpret trajectories and relational position of whole groups of actors. We look at the changes occurred for the types of organizations involved in the project (e.g., private companies (C), research centers (RC), Universities (U), other institutions (O), Fig. 5a). A change in the participation pattern can be appreciated over time from private-public to private-private. All these changes lead to global structural changes that can be represented using the whole table representation, in which the $\mathbf{F}_{[k]}$ are taken as unit of analysis (Fig. 5b). In such a case we can observe that the first five years are very similar in terms of $actor - to - project$ participation, while in the last two years a structural break appears. This is due to the participation in large European projects in new research fields that involves a variety of new actors.

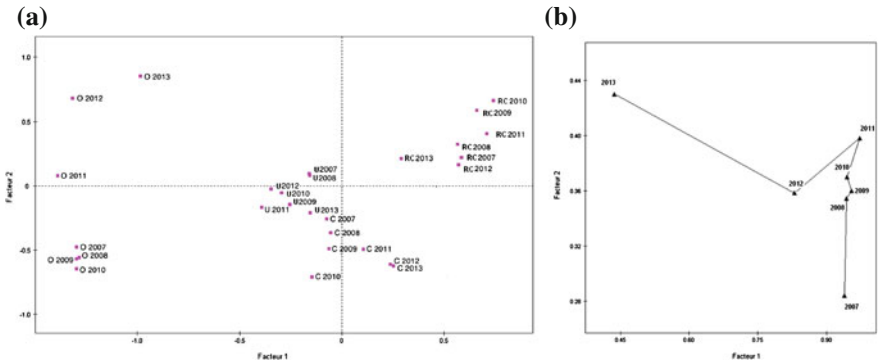


Fig. 5 **a** Representation of actor typology over time: private companies (C), research centers (RC), Universities (U), other institutions (O). **b** Representation of the time occasions relational distance: each year is represented by an individual point

Summing up, MFA represents a flexible tool to visualize and analyze the complex relational structure embedded in a three-way two-mode networks. Different type of information can be obtained using cluster-type techniques proper of network analysis, such as blockmodeling. For the data at hand, the results from the use of blockmodeling are in [10].

Acknowledgments The authors acknowledge the financial support of the REPOS Project (Networks, Public Policies, and Development) FSE POR Campania 2007–2013.

References

1. Abdi, H., Williams, L.J., Valentin, D.: Multiple factor analysis: principal component analysis for multitable and multiblock data sets. WIREs Comput. Stat. **5**, 149–179 (2013)
2. Capuano, C., De Stefano, D., Del Monte, A., D' Esposito, M.R., Vitale, M.P.: The analysis of network additionality in the context of territorial innovation policy: the case of Italian Technological Districts. In: Giudici, P., Ingrassia, S., Vichi, M. (eds.) Statistical Models for Data Analysis, pp. 81–88. Springer, Heidelberg (2013)
3. Coppi, R.: An introduction to multiway data and their analysis. Comput. Stat. Data Anal. **18**, 3–13 (1994)
4. D' Esposito, M.R., De Stefano, D., Ragozini, G.: On the use of multiple correspondence analysis to visually explore affiliation networks. Soc. Netw. **38**, 28–40 (2014)
5. Escofier, B., Pagès, J.: Analyses Factorielles Simples et Multiples: Objectifs, Mèthodes, Interprètation. Dunod, Paris (1988)
6. Escofier, B., Pagès, J.: Multiple factor analysis (AFMULT package). Comput. Stat. Data Anal. **18**, 121–140 (1994)
7. Greenacre, M.: Biplots in Practice. Fundaciòn BBVA, Madrid (2010)
8. Kroonenberg, P.M.: Applied Multiway Data Analysis. Wiley, New York (2008)

9. Lazzeroni, M.: High-tech activities, system innovativeness and geographical concentration insights into technological districts in Italy. Eur. Urban Reg. Stud. **17**, 45–63 (2010)
10. Prota, L., Vitale, M.P.: A pre-specified blockmodeling to analyze structural dynamics in innovation networks. In: Vicari, D., Okada, A., Ragozini, G., Weihs, C. (eds.) Analysis and Modeling of Complex Data in Behavioral and Social Sciences. Springer, Switzerland (2014)
11. Ragozini, G., De Stefano, D., D' Esposito, M.R.: Multiple factor analysis for multirelational/time-varying two-mode networks. Netw. Sci. **3**, 18–36 (2015)

Bayesian Spatiotemporal Modeling of Urban Air Pollution Dynamics

Simone Del Sarto, M. Giovanna Ranalli, K. Shuvo Bakar, David Cappelletti, Beatrice Moroni, Stefano Crocchianti, Silvia Castellini, Francesca Spataro, Giulio Esposito, Antonella Ianniello and Rosamaria Salvatori

Abstract

This work deals with the spatiotemporal analysis of urban air pollution dynamics in the town of Perugia (Central Italy) using high-frequency and size resolved data on particular matter (PM). Such data are collected by an Optical Particle Counter (OPC) located on a cabin of the Minimetro, a public transport system that moves on a monorail on a line transect of the town. Hierarchical Bayesian models are used that allow to model a quite large dataset and include an autoregressive term in time, in addition to spatially correlated random effects. Models are fitted for three response variables (fine and coarse particle counts, nitric oxide concentration) and using covariate information such as temperature and humidity. Results show a large temporal autocorrelation, relatively larger for particle counts; moreover,

S.D. Sarto (✉)
Department of Economics, University of Perugia, Perugia, Italy
e-mail: delsarto@stat.unipg.it

M.G. Ranalli
Department of Political Sciences, University of Perugia, Perugia, Italy
e-mail: giovanna.ranalli@stat.unipg.it

K.S. Bakar
Deparment of Statistics, Yale University, New Haven, CT, USA
e-mail: shuvo.bakar@yale.edu

D. Cappelletti · B. Moroni · S. Crocchianti · S. Castellini
Department of Chemistry, Biology and Biotechnologies, University of Perugia, Perugia, Italy
e-mail: david.cappelletti@unipg.it

F. Spataro · G. Esposito · A. Ianniello · R. Salvatori
Institute of Atmospheric Pollution Research (IIA), CNR, Roma, Italy
e-mail: spataro@iia.cnr.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_10

all variables show a significant spatial correlation, with larger ranges for fine PM rather than for coarse PM and nitric oxide concentration.

1 Introduction

Urban pollution has an important impact on human health and environment. Investigating the behavior of pollutants and understanding air quality of particular geographical areas has been one of the central issues in environmental public policy and decision-making. In this paper, we analyze data from the PMetro project (http://www.pmetro.it), which studies urban pollution dynamics in the town of Perugia (Italy) since September, 2012. Unlike classical monitoring of pollutants concentration using fixed stations, fast measure of gases and size resolved particulate matter (PM) is coupled with information on the evolution of urban microclimate. In particular, data is collected using an Optical Particle Counter (OPC) located on a cabin of the Minimetro, a public conveyance that moves on a monorail on a line transect of the town.

Air pollution often shows a spatiotemporal structure that is interesting to study and understand: for this purpose, hierarchical Bayesian modeling provides useful tools to investigate spatial and/or temporal patterns also in large datasets [3,8–11]. In this paper, we aim at applying such models to this fairly novel dataset to describe the spatiotemporal structure of the data and understand whether there is a different behavior for fine and coarse particle counts, and for nitric oxide concentration. The paper is organized as follows. Section 2 illustrates in more detail the data set at hand and Sect. 3 provides the description of the spatiotemporal model employed. Then, in Sect. 4 results from the application of the model fitting are shown, while some concluding remarks are given in Sect. 5.

2 Data

An OPC integrated on a cabin of Minimetro is used to get a snapshot of the urban pollution dynamics along a sector of the town at high spatial and temporal resolution. Figure 1 provides all the details on the metro path. It is about 3 km long with seven stations: a single travel takes about twenty minutes, so that each cabin runs along the same path about forty times a day. The path is outdoors for the most part and passes through parks, high-traffic roads, residential areas, and two tunnels. The OPC takes a sample of air every 6 seconds while the cabin moves on the monorail and counts the number of particles with a diameter between 0.28 and 10 μm (μm, 10^{-6} m), dividing the total count into 22 size channels. The location and the speed of the cabin is continuously recorded by the central control software of the Minimetro transport system and transmitted to the OPC data logger. Therefore, sampling points

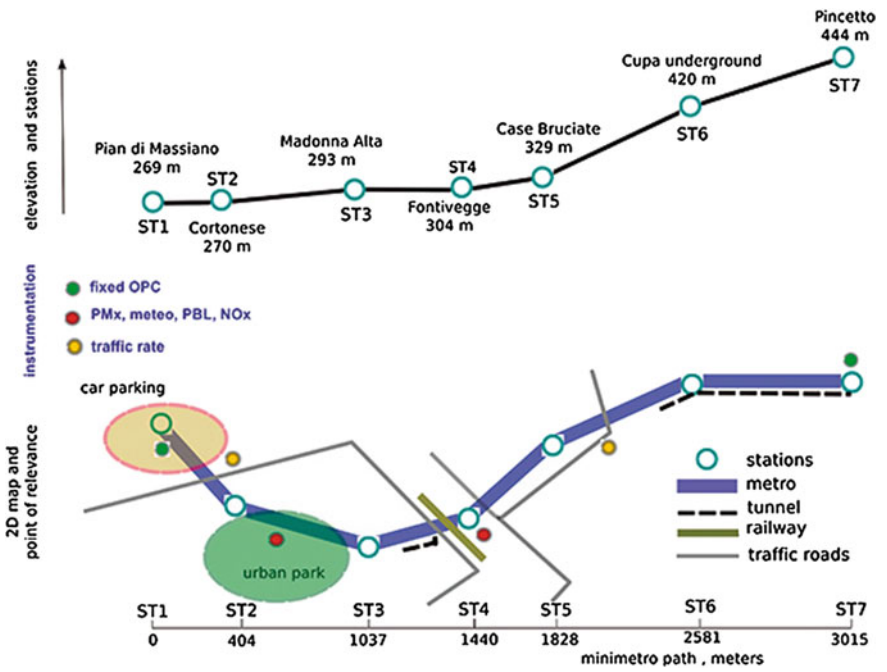


Fig. 1 Schematic map of the Minimetro path and of the sources of data. In the upper panel station names and elevation (meters a.s.l.) are indicated. In the lower panel a 2D sketch of the area suggests the main intersections with traffic roads and indicates also some points of interest (car parking, urban park, tunnels). The metro path is shown at the bottom together with distances along the path (in meters). Position and typology of the instruments employed in the project are also shown

are variable along the path and depend on the speed of the cabin, which is not constant along the path or during the day. Since March 2014, the instrument also collects nitric oxide (NO) concentration ($\mu g/m^3$), temperature and relative humidity at the same space-time resolution. As far as particulate matter is concerned, we use the following usual classification on the basis of the diameter: *fine* PM has diameter between 0.28 and 1.10 μm, while *coarse* PM has diameter larger than 1.10 μm.

We analyze data collected on March, April, and May, 2014, for overall 60 days. Due to the operation time of the Minimetro, we only have observations from 6 am to 7 pm. In addition, we average measurements for each hour, so that we have 14 hourly observations for each day. Furthermore, we divide the entire path of the Minimetro into 45 equally spaced bins, for which we determine the coordinates of the centroid in latitude and longitude. Therefore, the final dataset is made up of 37,800 observations (60 days $\times$ 14 h $\times$ 45 space-bins). It is possible to have some missing data due to maintenance and/or malfunctioning of the OPC, but we can consider missingness to be completely at random.

We perform the analysis separately for three response variables (NO concentration, fine and coarse PM) using the following covariates: temperature (°C), relative

humidity (RH) (%), altitude of the centroid of the space-bin (meters a.s.l.), day of the week (factor, reference day Sunday), dummy variable for tunnel (takes value 1 if the centroid of the space-bin is located inside a tunnel), dummy variable for park (takes value 1 if the centroid of the space-bin is located in a park), dummy variable for station (takes value 1 if the centroid of the space-bin is one of the metro stations).

3 Spatiotemporal Modeling

The main purpose of this paper is to better understand the mechanisms underlying urban air pollution dynamics and its spatiotemporal features. Since data are collected by a mobile station continuously moving during the day from the first suburbs to old-town center, a first objective is to detect whether response variables are spatially correlated and, if so, analyze the structure of such correlation. Similarly, since data are analyzed on an hourly base, it can be interesting to see if and how much observed variables are temporally autocorrelated. Finally, studying whether the three response variables have different spatial and/or temporal correlation would also allow a better understanding of the underlying urban air pollution dynamics. To this end, the spatiotemporal structure of the data has been explored.

As far as the spatial component is concerned, we compute the empirical variogram for each response variable, following [12] and accounting for the spatiotemporal nature of the data. Note that, given their skew distribution, the three response variables have been transformed and analyzed on a logarithmic scale. Variogram plots (see Fig. 2) exhibit a clear spatial variation for all the three response variables. In addition, the shape of the variograms suggests that an exponential correlation function can be considered plausible. The structure of the spatial correlation is similar for NO and coarse PM, while fine PM seems to have a larger range. Points in the variogram clouds are color coded according to whether semivariances are computed with respect to locations of which at least one is placed inside a tunnel. It can be noted that for fine PM, semivariance is considerably larger when one location is inside the tunnel, since its level is significantly different from that observed outside the tunnel.

Looking at the temporal component, on the other end, the autocorrelation function (ACF) and the partial autocorrelation function (PACF) have been computed for each response and for each space-bin. Inspection of the three sets of plots (not shown here for reasons of space) provide evidence of a strong lag-1 autoregressive structure for most sites, stronger for PM counts and slightly weaker for NO concentration.

Considering the results obtained in the exploratory analysis, we use the spatiotemporal autoregressive model proposed in [11]. Let l and t denote the two units of time, where $l = 1, \ldots, r$ denotes the longer unit, i.e., the day, and $t = 1, \ldots, T_l$ denotes the shorter unit, i.e., the hour. Let $Z_l(\boldsymbol{s}_i, t)$ be the observed point referenced data at space-bin $\boldsymbol{s}_i, i = 1, \ldots, n$ at time denoted by the two indexes l and t, and let $O_l(\boldsymbol{s}_i, t)$ be the true value corresponding to $Z_l(\boldsymbol{s}_i, t)$. Then let $\boldsymbol{Z}_{lt} = \left[Z_l(\boldsymbol{s}_1, t), \ldots, Z_l(\boldsymbol{s}_n, t)\right]^T$ be a $n \times 1$ vector of observed values at n space-bins and let $\boldsymbol{O}_{lt}$, ε_{lt} and η_{lt} be defined

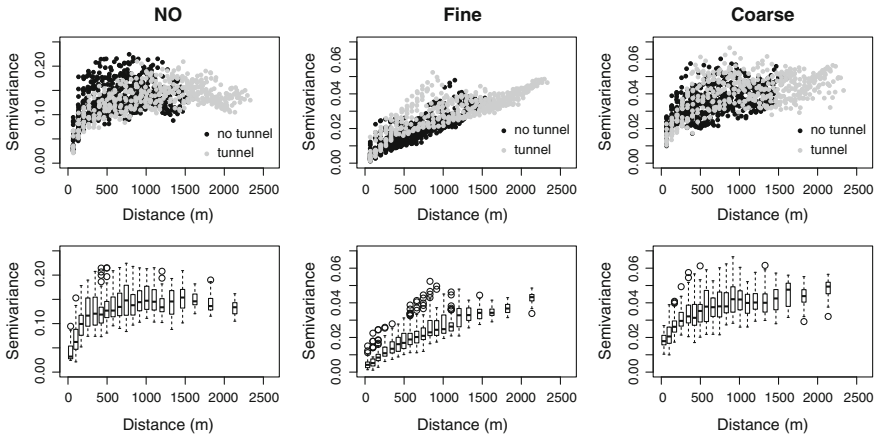


Fig. 2 Semivariogram cloud (*top*) and corresponding boxplot (*bottom*) for each variable of interest. Points in the cloud are color coded according to whether at least one location is placed inside a tunnel. *Boxplots* are produced from 5 %–quantiles of the distance

similarly in terms of $O_l(\boldsymbol{s}_i, t)$, $\varepsilon_l(\boldsymbol{s}_i, t)$ and $\eta_l(\boldsymbol{s}_i, t)$, respectively. The structure of this model is a simplified version with respect to the model used by [11] and is specified hierarchically as follows:

$$\boldsymbol{Z}_{lt} = \boldsymbol{O}_{lt} + \boldsymbol{\varepsilon}_{lt}, \tag{1}$$

$$\boldsymbol{O}_{lt} = \boldsymbol{X}_{lt}\boldsymbol{\beta} + \rho \boldsymbol{O}_{lt-1} + \boldsymbol{\eta}_{lt}, \tag{2}$$

$$for\, l = 1, \ldots, r \quad t = 1, \ldots, T_l,$$

where $\boldsymbol{\varepsilon}_{lt}$ is the so-called nugget effect (or the pure error term) and is assumed to have distribution $N(\mathbf{0}, \sigma_\varepsilon^2 \boldsymbol{I}_n)$, where σ_ε^2 is the unknown variance and $\boldsymbol{I}_n$ is an identity matrix of order n; β is a $p \times 1$ vector of regression coefficients and $\boldsymbol{X}_{lt}$ is a $n \times p$ matrix of covariates. Furthermore, we also assume that the spatiotemporal random effects, $\boldsymbol{\eta}_{lt}$, follow a normal distribution $N(\mathbf{0}, \Sigma_\eta)$ independently in time, with $\Sigma_\eta = \sigma_\eta^2 \boldsymbol{S}_\eta$, where σ_η^2 is the site invariant spatial variance and $\boldsymbol{S}_\eta$ is the spatial correlation matrix. This matrix can be determined using the general Matérn correlation function or, more simply, using the exponential function, that may depend on a parameter ϕ, controlling the rate of decay of the correlation as the distance between two locations increases [2]. We use the exponential form for the spatial correlation, so each element of $\boldsymbol{S}_\eta$ is given by $\exp(-\phi d_{ij})$, where d_{ij} is the distance between the centroids of the space-bins $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, $i, j = 1, \ldots, n$. Finally, ρ denotes the unknown temporal correlation parameter which lies in the interval $(-1, 1)$.

The parameters of the model are, therefore, β, ρ, σ_ε^2, σ_η^2 and ϕ. Their posterior distribution (except for ϕ) is determined using Gibbs sampling and each full conditional distribution is provided in the Appendix of [1]. We specify a flat Normal prior for β and ρ, with mean 0 and variance 10^{10}. The prior distribution for the two precision parameters ($1/\sigma_\varepsilon^2$ and $1/\sigma_\eta^2$) is chosen to be a Gamma distribution with shape $a = 2$ and rate $b = 1$ so as to guarantee a proper prior distribution for

each variance parameter [5,11]. The estimate for ϕ is obtained using an Empirical Bayes approach as in [10]: we search the best fitting model (1)–(2) over a grid of different values for ϕ. We determine the interval of possible values of ϕ using the definition of *effective range*, i.e., the distance at which there is essentially no lingering spatial correlation [2]: for this purpose we use the relationship $\exp(-\phi d_{ij}) \approx 0.05, i, j = 1, \ldots, n$, using the minimum and the maximum among all distances between space-bins. The optimization criterion is based on the formal Predictive Model Choice Criterion (PMCC) [4,6] given by

$$\text{PMCC} = \sum_{i=1}^{n} \sum_{l=1}^{r} \sum_{t=1}^{T_l} \left\{ E\left[Z_l(\boldsymbol{s}_i, t)_{rep} - Z_l(\boldsymbol{s}_i, t)\right]^2 + Var\left[Z_l(\boldsymbol{s}_i, t)_{rep}\right] \right\}, \quad (3)$$

where $Z_l(\boldsymbol{s}_i, t)_{rep}$ is a replicate observation of $Z_l(\boldsymbol{s}_i, t)$ under the assumed model, sampled from the posterior predictive distribution. In (3), the first term is a goodness-of-fit term while the second can be seen as a penalty term, so that a model is considered less desirable if it has large predictive variance.

4 Results

To estimate model (1)–(2) above, we use the R package `spTimer` [1] version 1.0–2. This package requires that covariates do not have missing values; to overcome this issue, we perform imputation for temperature and relative humidity using again model (1)–(2), with the same space-time structure and only the intercept term. The model has been fitted separately for each response, running MCMC with 10,000 iterations and burn-in 1,000. Results are summarized in Table 1. All responses show a significant temporal autocorrelation (with respect to the previous hour), although NO shows a lower estimated value for ρ, than fine and coarse PM (0.260 versus 0.872 and 0.886, respectively). Analysis of the (partial) autocorrelation function of the residuals of the models does not provide evidence of a lag-2 dependence.

Fine PM has a lower Empirical Bayes estimates for ϕ (0.006, corresponding to about 475 m) than coarse PM (0.018, corresponding to about 165 m). Both fine and coarse PM are chemically stable in atmosphere. Therefore the different spatial correlation is due to the larger mobility of fine particles, in turn related to their lower mass. Coarse particles tend to settle more fastly and, therefore, are transported for smaller distances; this implies they are more localized near the sources, among which there is, typically, resuspension form vehicular motion. For NO, estimated ϕ takes value 0.021 (143 m), very similar to the one obtained for coarse PM as already noted inspecting variograms in Fig. 2. This is related with the higher chemical reactivity of NO which may be oxidized to NO_2 and enter into the, rather complex, ozone production cycle. For these reasons also NO is localized near the sources, which include not only traffic but also domestic heating.

Parameter estimates for the covariates are often significant, although their values are not particularly large. Those related to the day of the week deserve a closer

Table 1 Posterior estimates for Nitric Oxide, fine and coarse PM: mean and 95 % Credible Interval. Estimates for ϕ are obtained using Empirical Bayes. Sunday is the reference day of the week

Parameter	Nitric oxide		Fine (0.28–1.10 μm)		Coarse (≥1.10 μm)	
	Mean	(95 % CI)	Mean	(95 % CI)	Mean	(95 % CI)
Intercept	1.421	(1.350; 1.490)	1.551	(1.495; 1.610)	0.682	(0.641; 0.722)
Mon	0.170	(0.142; 0.198)	−0.006	(−0.021; 0.010)	−0.042	(−0.055; −0.029)
Tue	0.192	(0.166; 0.220)	−0.073	(−0.089; −0.057)	−0.030	(−0.043; −0.017)
Wed	0.136	(0.110; 0.163)	−0.092	(−0.107; −0.076)	−0.070	(−0.083; −0.057)
Thu	0.112	(0.084; 0.139)	−0.038	(−0.054; −0.022)	−0.094	(−0.108; −0.080)
Fri	0.294	(0.267; 0.322)	0.066	(0.050; 0.082)	0.034	(0.021; 0.047)
Sat	0.059	(0.033; 0.087)	−0.032	(−0.048; −0.017)	−0.060	(−0.073; −0.047)
Tunnel	−0.083	(−0.103; −0.064)	−0.006	(−0.012; 0.001)	−0.001	(−0.010; 0.008)
Park	−0.043	(−0.067; −0.019)	−0.004	(−0.012; 0.004)	−0.002	(−0.014; 0.009)
Station	−0.027	(−0.042; −0.013)	0.007	(0.004; 0.010)	0.013	(0.006; 0.019)
Temp (°C)	−0.030	(−0.032; −0.028)	−0.010	(−0.011; −0.009)	0.005	(0.004; 0.006)
RH (%, ×10)	−0.0047	(−0.0079; −0.0015)	0.0048	(0.003; 0.006)	−0.0014	(−0.0029; 0.0002)
Altit. (m, ×100)	0.040	(0.023; 0.057)	−0.008	(−0.016; −0.0006)	−0.005	(−0.014; 0.003)
ρ	0.260	(0.252; 0.268)	0.872	(0.868; 0.877)	0.886	(0.881; 0.890)
σ_ε^2	0.104	(0.099; 0.109)	0.0096	(0.009; 0.010)	0.010	(0.0096; 0.0104)
σ_η^2	0.308	(0.303; 0.314)	0.037	(0.0368; 0.0381)	0.066	(0.065; 0.067)
ϕ	0.021	–;–	0.006	–;–	0.018	–;–

look: in the model for NO, they all take positive significant values with a peak on Fridays and relatively smaller concentrations on Saturdays and Sundays. This is in line with the fact that NO can be considered as a proxy of vehicular traffic and combustion in general. Fine and coarse PM display a different pattern: posterior means are all negative (except for Friday) and some are not significant. These results can be explained by a peculiar condition due to atmospheric stability problems, observed in some of the days considered here.

5 Conclusions

In this paper, we analyze a fairly new dataset on PM counts and NO concentration in the town of Perugia (central Italy) using Bayesian spatiotemporal models. Such data are collected using a mobile station and are indexed by time (hours and days) and space (position of the station). In order to take the spatiotemporal structure into account, we employ an autoregressive spatiotemporal model, in which an AR(1) term is included at true process level, while the spatial structure is incorporated using spatially correlated random effects. Such model has been fit separately for three variables of interest: fine and coarse particle counts, and NO concentration. Results reveal that all responses have a consistent temporal autocorrelation component, with fine and coarse PM more correlated in time than NO, while fine PM shows a larger spatial correlation than coarse PM and NO.

Further research will focus on the use of multivariate spatiotemporal models, using a bivariate response (e.g., fine PM and NO) or a multivariate response, consisting in all the 22 size channels counts collected by the OPC. Finally, Integrated Nested Laplace approximation (INLA) proposed by [7] is another technique to get inference in Bayesian models. It is developed as a computationally efficient alternative to MCMC and is designed for latent Gaussian models, a general and wide class of flexible models, among which we can include the spatiotemporal models considered here. INLA produces a numerical approximation to the posterior distribution of interest. Therefore, a comparison between MCMC and INLA in our context could be an interesting and tempting challenge, particularly when dealing with a large amount of data.

References

1. Bakar, K.S., Sahu, S.K.: spTimer: spatiotemporal bayesian modelling using R. J. Stat. Softw. **63**(15), 1–32 (2015)
2. Banerjee, S., Carlin, B.P., Gelfand, A.E.: Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall/CRC, Boca Raton (2004)

3. Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H.: Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc. B **70**, 825–848 (2008)
4. Gelfand, A.E., Ghosh, S.K.: Model choice: a minimum posterior predictive loss approach. Biometrika **85**, 1–11 (1998)
5. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis. Chapman & Hall/CRC, Boca Raton (2004)
6. Laud, P., Ibrahim, J.: Predictive model selection. J. R. Stat. Soc. B **57**, 247–262 (1995)
7. Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. J. R. Stat. Soc. B **71**, 319–392 (2009)
8. Sahu, S.K., Bakar, K.S.: A comparison of Bayesian models for daily ozone concentration levels. Stat. Method. **9**, 144–157 (2012a)
9. Sahu, S.K., Bakar, K.S.: Hierarchical Bayesian autoregressive models for large space time data with applications to ozone concentration modelling. Appl. Stoch. Models Bus. Ind. **28**, 395–415 (2012b)
10. Sahu, S.K., Gelfand, A.E., Holland, D.M.: Spatiotemporal modeling of fine particulate matter. J. Agric. Biol. Environ. Stat. **11**, 61–86 (2006)
11. Sahu, S.K., Gelfand, A.E., Holland, D.M.: High-resolution space-time ozone modeling for assessing trends. J. Am. Stat. Assoc. **102**, 1221–1234 (2007)
12. Sahu, S.K., Mardia, K.V.: A Bayesian kriged kalman model for short-term forecasting of air pollution levels. J. R. Stat. Soc. C **54**, 223–244 (2005)

Clustering Functional Data on Convex Function Spaces

Tonio Di Battista, Angela De Sanctis and Francesca Fortuna

Abstract

The curves in a functional data set often present a variety of distinctive patterns corresponding to different shapes that can be identified by clustering the functions. However, clustering functional data is a difficult task because the function space is, generally, of infinite dimension. Thus, the distance among functions may have infinity solutions and can be approximated in different ways leading to different clustering results. The paper deals with this problem and focuses on cases in which the functional form of the observations is known in advance. In this setting, the approximation of the function underlying the data is not required and the functional distance may be computed directly in the explicit form of the functions. Moreover, we restrict the space of the functions to a closed and convex subset in an Hilbert space to achieve desirable properties. In the proposed framework, an L^2 metric is applied combined clustering algorithms for finite dimensional data. The method is applied to a real data set concerning lichen biodiversity in the province of Genoa, North Western Italy.

T. Di Battista (✉) · F. Fortuna
DISFPEQ,
University G. d' Annunzio, Chieti-Pescara, Chieti, Italy
e-mail: dibattis@unich.it

F. Fortuna
e-mail: francesca.fortuna@unich.it

A. De Sanctis
Department of Management and Business Administration, University G. d' Annunzio, Chieti-Pescara, Chieti, Italy
e-mail: a.desanctis@unich.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_11

1 Introduction

Functional data analysis (FDA) [18,24] addresses problems in which the observations are described by functions rather than finite dimensional vectors.

The curves in a functional data set often present a variety of distinctive patterns corresponding to different shapes and variation that can be identified by clustering the functions [1,27].

However, clustering functional data is generally a difficult task because of the infinite dimensional space that data belong to. For this reason, many approaches are based on dimension reduction before clustering. The most simple method is called raw-data clustering. It consists in applying classical clustering methods directly on the discretization of the functions [4]. This procedure, obviously, presents many limitations because it ignores the functional nature of the observations. Indeed, the most commonly used approaches reduce the infinite dimension problem to a finite one by approximating data with elements from some finite dimensional space, such as coefficients of functional data expansion [1] or a given number of principal component scores [22]. Then, classical heuristic clustering algorithms can be performed. It is common to refer to this procedure with the term two-stage approach. Alternatively, nonparametric clustering methods consist in defining specific distances or dissimilarities among the curves and then in applying clustering algorithms for finite dimensional data [18]. Finally, model-based clustering methods can be performed assuming a probabilistic distribution on some finite dimensional coefficients describing the data [26].

The disadvantage of the above methods is that clustering results can differ depending on how the curves are fitted to the data. For example, nonparametric methods can be assimilated to raw-data clustering or to a two-stage method, depending on whether the distance is approximated using directly the discrete observations of curves or using an approximation of the curves into a finite basis.

This paper focuses on a nonparametric clustering method and deals with a particular aspect of FDA. Indeed, it refers to cases in which the functional form underlying the observations is known in advance [6,8,12–16] since it is expressed by a parametric model and the functions constitute a convex subset of an L^p space. In these cases, the approximation by means of basis functions is not required and the functional distance can be computed directly on the explicit form of the functions. Thus, the problem of how the curves are fitted to the data does not arise.

The paper is organized as follows. Section 2 introduces functional data that belong to a specific parametric family of functions and deals with nonparametric clustering methods for functional data. In particular, the functional L^2 distance is presented focusing on cases in which the function space is parametric and convex in L^2. In the same section the functional k-means algorithm is presented by specifying its critical issues in the functional framework. Finally, Sect. 3 shows an empirical application dealing with epiphytic lichen biodiversity of the province of Genoa, Liguria (Italy).

2 Functional Distances on Convex Function Spaces

The classical FDA approach assumes the existence of certain unknown smooth functions $f(\cdot)$ which generate and underlie the data. However, since in real cases functional data are often observed as a sequence of point data, the first FDA task is to fit the true form of the underlying function through some techniques such as basis functions expansion and regularization [24]. These techniques are largely used in the literature as demonstrated by numerous empirical applications based on them (see [23], for some examples). Nevertheless, there are situations in which their use is not suitable because it mystifies the intrinsic characteristics of the data; for example when the functions show points of discontinuity or singularities or lie in a discrete functional domain or when the underlying data process is known in advance. In these cases, functional data are not intrinsically smooth, thus, it is preferable to work directly on the reference functional space, when it is possible.

This paper focuses on a particular aspect of functional data analysis, i.e., when the functional datum is expressed by a specific function known in advance. Examples of this kind of data are present in many disciplines. In economics, one can refer to the Cobb–Douglas production functions, which are used to study the relationship between input factors and the level of production. Another example can be found in biology where the logistic growth function is used to describe growth processes. In ecology, the biodiversity is evaluated by means of diversity profiles, which express diversity as a function of the relative abundance vector [8,19].

In these cases, the functional space S is constituted by a set of functions belonging to the same parametric family

$$S = \{f(\boldsymbol{\theta}; x)\}, \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_s)'$ represents a set of parameters taking values in a parameter space $\boldsymbol{\Theta}^x$; x is the functional domain and S is a subset of some L^p space. In particular, we focus on cases in which S is a convex subset of L^p.

Starting from n parametric functional data, $f(\boldsymbol{\theta}_1, x)$, $f(\boldsymbol{\theta}_2, x), \ldots, f(\boldsymbol{\theta}_n, x)$, we aim to identify a set of homogeneous clusters in L^p by determining a partition of the space according to the minimal distance. Since the functional observations belong to an infinite dimensional space, the equivalence between norms and distances, typical of finite dimensional Euclidean spaces, fails. In this context, the choice of the preliminary norm becomes crucial [18], because the resulting metric might not be complete. For example, let us consider the uniform norm, the L^1 norm and the L^2 norm on the space of continuous functions in $(0, 1)$. The three norms are inequivalent and the space is complex for the first norm but not for the other two.

Several proximity measures for functional data can be used, depending on the characteristics of the data and on the target of interest in clustering. For example, in [18] the use of a family of semi-metric is suggested because it is more flexible with respect to the metric. However, this choice emphasizes features of the derivatives rather than of the curves themselves. A suitable measure of distance between two functions, f and g, is the L^p distance for which different p can be chosen (the use of L^1, L^2 and L^∞ is very common in the literature). In this paper, we focus on the

L^2 distance

$$d\Big(f, g\Big) = ||f - g||^2 = \left(\int_X |f(x) - g(x)|^2 d\mu(x)\right)^{\frac{1}{2}}. \tag{2}$$

We recall that $L^2(\mu)$ is, among all $L^p(\mu)$ spaces, the only Hilbert space, that is its norm is induced by the inner product

$$(f, g) = \int_X f(x)g(x)d\mu(x) \tag{3}$$

whenever $f, g \in L^2(\mu)$. Therefore, an orthogonality notion between two functions f and g is defined as $(f, g) = 0$.

The main problem caused by the infinite dimensionality of the spaces is that there may be infinitely many solutions for the distance in Eq. (2). Indeed, different norms may lead to different conclusions about convergence of a given sequence. Thus, different clustering results can be obtained. For this reason, we restrict the space of the functions to closed and convex subsets, S, in Hilbert spaces. An essential property of Hilbert space is that the distance of a point to a closed set is always attained. Indeed, if S is a closed convex set in a Hilbert space, H, and $h \in H$, then, there exists a unique point $s \in S$ that minimize the distance between h and a point in S [25]

$$d(h, S) = ||h - s|| = \min\{||h - s|| : s \in S\}. \tag{4}$$

Thus, a closed convex subset of a Hilbert space has a unique minimum norm. Generally, this issue holds in any uniformly convex Banach space.

Other functional spaces can be considered, such as appropriate Sobolev spaces, $W^{k,p}$, that is, vector spaces of functions equipped with a norm that is a combination of L^p-norms of the function itself as well as its derivatives up to a given order. Indeed, for $p = 2$, the Sobolev space is an Hilbert space endowed with the Hilbert inner product and the Hilbert norm; whereas, for $1 \leq p \leq \infty$, it is a Banach space and it is uniformly convex. Generally, Sobolev spaces are powerful in demonstrating existence of solutions to partial differential equations [5] but their use can be considered also in clustering problems. For example, in [2] Sobolev spaces are considered for clustering functional data using wavelet-based similarity measures.

2.1 Functional k-means

In this proposed setting an L^2 metric in function space is applied combined with a k-means algorithm for finite dimensional data. The k-means algorithm is an iterative procedure that is initialized by fixing the number k of clusters, $\{C_1, C_2, \ldots, C_k\}$ and by selecting in S a set of k arbitrary and distinct initial centroids, $\{\phi_1^{(0)}(x), \ldots, \phi_k^{(0)}(x)\}$, one for each cluster. At the m-th algorithm iteration, $m > 0$, each function is assigned to the cluster whose centroid, at the $(m-1)$-th iteration, is the nearest according to the chosen distance

$$\underset{q=1,2,\ldots,k}{\arg\min} \left(\int_X |f_i(x) - \phi_q^{m-1}(x)|^2 dx\right)^{\frac{1}{2}}. \tag{5}$$

Once all of the functions have been assigned to a cluster, the cluster means are updated as the mean of the functions belonging to it, as follows:

$$\phi_q^{m+1}(x) = \sum_{f_i \in c_q} \frac{f_i(x)}{n_q}, \tag{6}$$

where n_q is the number of functions in the q-th cluster, C_q. This procedure continues until no function changes cluster.

In our specific case S is a parametric set of functions. Thus, the functional distance in Eq. (5) can be computed directly on the explicit form of the functions. For this reason, clustering results do not depend on how the curves are smoothed to the data, contrary to the classical FDA approach. Indeed, it is well known that, functional k-means clustering results vary according to the method used for fitting the curves [27]. In the classical FDA framework, thus, the primary question of interest is how best to linearly transform the data prior to clustering.

Our approach allows us to obtain some desirable properties because it considers cases in which S is a closed and convex subset in a Hilbert space; then it contains a unique element of smallest norm [25]. Moreover, the convexity allows us to define the functional mean in the usual way obtaining an element of S. Thus, the centroid in Eq. (6) belongs to the same family as the functions. Indeed, this essential requirement is not always achieved with the classical functional approach [6,14]. In particular, only when S is a linear vectorial subspace in L^p, it is possible to express the functional statistics as a straightforward statistics of the functions obtaining a statistics of the same functional form of the observed data [8].

3 Application

The framework described previously has been applied to a real data set concerning epiphytic lichens biodiversity of the province of Genoa, Liguria (Italy). Lichen biodiversity provides useful information about the global conditions affecting the environment over a given area [20]. These organisms, indeed, are particularly sensitive to environmental stresses, especially with regard to pollution, eutrophication, and climate change [3].

Data on lichen abundance were collected following the standards suggested by [3]; the survey lasted from 2002 to 2003 and involves a total of 196 epiphytic lichen species and 47 plots [20].

For every i-th environmental site ($i = 1, \ldots, 47$) and for each j-th species ($j = 1, \ldots, 196$), we consider an abundance vector, $\boldsymbol{N}_i = (N_{i1}, \ldots, N_{is})'$, calculated as the sum of the lichen frequencies found on every plot and a relative abundance vector, $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{is})'$, with $p_j = N_j / \sum_{j=1}^{s} N_j$, such that $0 \leq p_j \leq 1$ and $\sum_{j=1}^{s} p_j = 1$ [10].

In order to evaluate lichen biodiversity we refer to parametric families of diversity indices [21], which are usually referred to as diversity profiles. They portray the

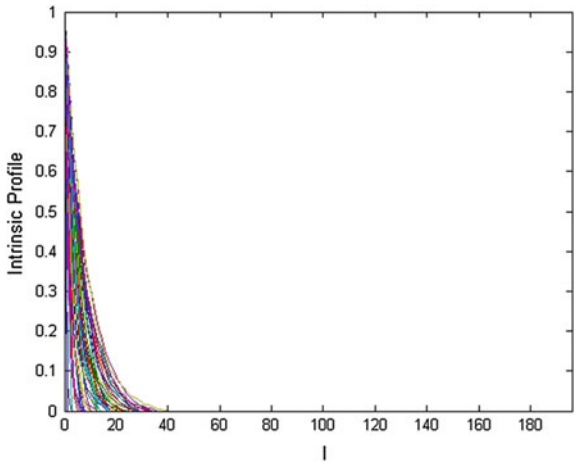


Fig. 1 Intrinsic profiles for the province of Genoa

simultaneous values of a large collection of diversity indices in a single diversity spectrum. In particular, the intrinsic diversity profile proposed by [21] has been applied. It is defined as the plotting of the (l, T_l) pairs, where

$$T_l = \sum_{j=l+1}^{s} p(j)^{\#} \qquad l = 0, 1, \ldots, s, \tag{7}$$

is the right-tail sum diversity index, $p(j)^{\#}$ is the relative abundance vector ranked in descending order and l is the species abundance rank. The intrinsic diversity profile plays a fundamental role in comparing different community using the analysis of a graph [7,9,11,17]. Indeed, if the profiles do not intersect, the higher curve corresponds to the community with greater diversity. On the contrary, communities with intersecting profiles are not comparable. However, in real cases, the profiles intersect one or more times leading to ranking problems among communities [15]. Figure 1 displays the intrinsic profiles for the province of Genoa. However, since there are many rare species, Fig. 2 shows the intrinsic profiles focusing on the first forty ordered species.

According to the analysis of the graph, it is no possible to distinguish a site with greater diversity because the profiles cross each other.

Biodiversity comparison is an important issue for planning environmental policies. For this reason, our aim is to characterize the sites of the province by determining different biodiversity patterns. However, due to the large number of curves in Fig. 2, it is difficult to pick out distinct and representative curve shapes. Thus, we proceed with the identification of homogeneous groups of data.

Despite the intrinsic diversity profile is a discrete function observed for each ranked species, in the literature it is showed by means of a curve. For this reason, Di Battista et al. [8,19] suggested to analyze it in a functional framework. Since the parametric functional form of the intrinsic profile is known in advance, we propose to

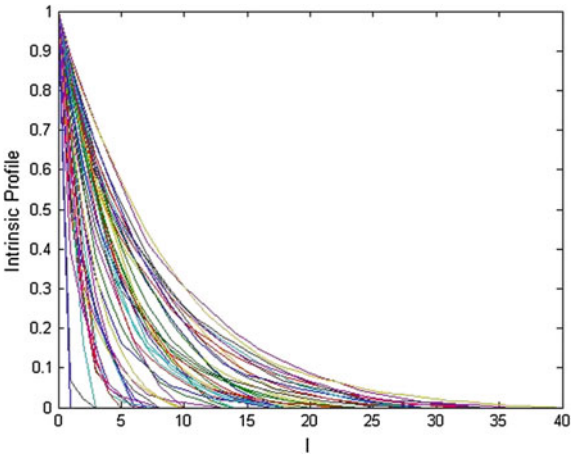


Fig. 2 Intrinsic profiles for the first 40 ordered species of the province of Genoa

work directly on the reference functional space preserving its typical characteristics [14].

In order to identify specific common patterns among the sites, a nonparametric clustering method has been applied. In particular, a functional k-means algorithm has been implemented computing the functional distance directly on the explicit known form of the functions. The k-means procedure has been initialized by choosing in S two arbitrary centroids $\phi_1^{(0)}(x), \phi_2^{(0)}(x)$. Then, the functions are assigned to the clusters according to the minimal distance between them and the centroids. Since the set of the sequences of the intrinsic profiles in Eq. (7) lies in a discrete domain, the functional space S is a subset of l^p and the usual Euclidean metric can be used. Once all functions have been assigned to a cluster, the cluster mean has been computed in the usual way as the average of the functions, obtaining an element of S due to the convexity property

$$\overline{T_l} = \frac{1}{n} \sum_{i=1}^{n} T_{li}. \tag{8}$$

Figure 3 shows the clustering results focusing on the first forty ordered species. The first cluster (red lines) is composed of 27 sites with high biodiversity, that is sites with great species richness and low dominance. The second cluster (blue lines) presents the opposite situation and it is composed of 20 sites. Figure 4 displays the $k = 2$ cluster mean curves obtained from the k-means algorithm. The procedure shows good results, in fact, the profiles belonging to the two clusters do not intersect except for three sites. Indeed, these latter present some singularities. For example, a site belonging to the second cluster intersects with the profiles of the first cluster because it has greater evenness with respect to the group it belongs to, but, at the same time, shows high species dominance. In the same way, two profiles of the first cluster intersect with those of the second one because they have a smaller number of species compared to the first group.

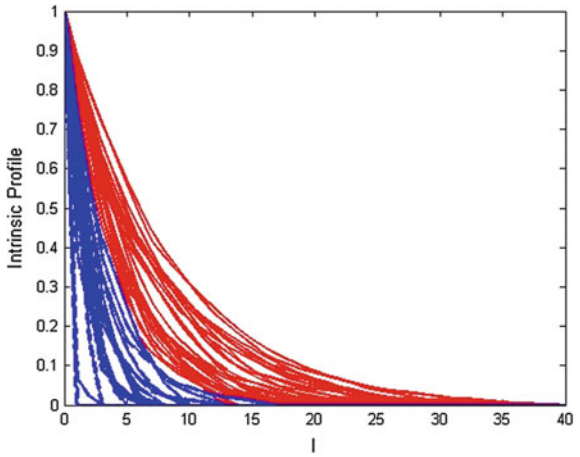


Fig. 3 Two clusters for T_{jl} for the first 40 ordered species

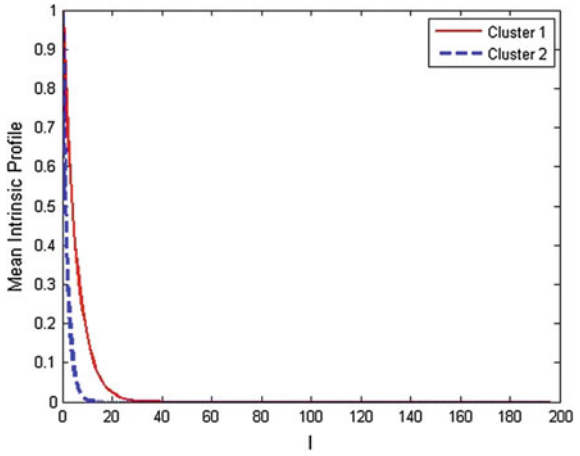


Fig. 4 Mean Intrinsic profile for the two clusters for the first 40 ordered species

In order to evaluate the performance of the clustering results obtained, following we compare our approach with classical FDA nonparametric clustering methods. For example, we can fit the intrinsic profiles using a B-spline basis and the k-means algorithm can be applied on the coefficients of the basis functions approximation. In this case, clustering results are equal to those obtained with our method except for two sites. Also in this case, problems of intersection for the same three sites arise. However, using an approximation of the curves into a finite basis, some problems may arise. For example, when the raw data points are converted into the continuous function there is no best choice of the basis function and the functional k-means algorithm can vary considerably depending on how the functional data is transformed prior to clustering [27]. Thus, the choice of the functional basis becomes crucial for clustering results. Obviously, a great deal also depends on how efficient the basis functions are in reproducing the behavior of the original functions. We point out that

the classical FDA approach works in a continuous domain while the intrinsic profile lies in a discrete functional domain which is represented by the relative abundance vector ranked in descending order. Accordingly, in this case, the functional datum is not intrinsically smooth and the use of basis function approximation could hide some characteristics of the phenomenon.

Our method allows to overcome some critical aspects of classical FDA approach, taking advantage of the known form of the functions and of the convexity of the reference functional space. In particular, the approximation of the function underlie the data is not required since it is expressed by a specific parametric model. For this reason, it is possible to classify the intrinsic profiles computing the functional distance directly on the explicit form of the observations, leading to clustering results that does not depend on how the curves are fitted to the data. In addition, the convexity property allows us to define the functional mean in the usual way obtaining an element belonging to the same family as the functions.

From an ecological point of view, the proposed method allows us to identify different patterns of biodiversity when the intrinsic profile does not highlight an explicit ranking of biodiversity among sites.

References

1. Abraham, C., Cornillon, P.A., Matzner-Løber, E., Molinari, N.: Unsupervised curve clustering using B-splines. Scandinavian J. Stat. Theory Appl. **30**(3), 581–595 (2003)
2. Antoniadis, A., Brossat, X.: Clustering functional data using wavelets. Int. J. Wavelets Multiresolution Inf. Process. **11**(1), 1350003-1–1350003-30 (2013)
3. Asta, J., Erhardt, W., Ferretti, M., Fornasier, F., Kirschbaum, U., Nimis, P.L., Purvis, O.W., Pirintos, S., Scheidegger, C., Van Haluwyn, C., Wirth, V.: Mapping lichen diversity as an indicator of environmental quality. In: Nimis, P.L., Scheidegger, C., Wolseley, P.A. (eds.) Monitoring with Lichens-Monitoring Lichens. Nato Science Program-IV, pp. 273–279. Kluwer Academic Publisher, The Netherlands (2002)
4. Bouveryon, C., Brunet, C.: Model-based clustering of high-dimensional data: A review. Technical report, Laboratoire SAMM, Universit Paris 1 Pantheon-Sorbonne (2003)
5. Brezis, H.: Functional Analysis, Sobolev Spaces and Partial Differential Equations. Springer, New York (2011)
6. De Sanctis, A., Di Battista, T.: Functional analysis for parametric families of functional data. Int. J. Bifurc. Chaos **22**(9), 1250226-1–1250226-6 (2012)
7. Di Battista, T.: Diversity index estimation by adaptive sampling. Environmetrics **13**(2), 209–214 (2002)
8. Di Battista, T., Fortuna, F.: Assessing biodiversity profile through FDA. Statistica **1**, 69–85 (2013)
9. Di Battista, T., Gattone, S.A.: Non parametric tests and confidence regions for intrinsic diversity profiles of ecological populations. Environmetrics **14**(8), 733–741 (2003)
10. Di Battista, T., Gattone, S.A.: Multivariate bootstrap confidence regions for abundance vector using data depth. Environ. Ecol. Stat. **11**, 355–365 (2004a)
11. Di Battista, T., Gattone, S.A.: Simultaneous inference on diversity of biological communities. Stat. Methods Appl. **13**, 129–136 (2004b)

12. Di Battista, T., Fortuna, F., Maturo, F.: Parametric functional analysis of variance for fish biodiversity. In: International Conference on Marine and Freshwater Environments. iMFE (2014)
13. Di Battista, T., Fortuna, F., Maturo, F.: Recent advances in functional data stream classification. In: Proceedings of the 60th World Statistics Congress of the International Statistical Institute, ISI2015, The Hague, The Neverlands (2015)
14. Di Battista, T., De Sanctis, A., Fortuna, F.: Functional statistics on function spaces. Statistical Methodology. Under review (2016)
15. Di Battista, T., Fortuna, F., Maturo, F.: Environmental monitoring through functional biodiversity tools. Ecol. Indic. **60**, 237–247 (2016a)
16. Di Battista, T., Fortuna, F., Maturo, F.: Parametric functional analysis of variance for fish biodiversity assessment. J. Environ. Inf. (2016b, to appear)
17. Fattorini, L., Marcheselli, M.: Inference on intrinsic diversity profiles of biological populations. Environmetrics **10**(5), 589–599 (1999)
18. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis. Springer, New York (2006)
19. Gattone, S.A., Di Battista, T.: A functional approach to diversity profiles. J. R. Stat. Soc. **58**, 267–284 (2009)
20. Giordani, P., Brunialti, G., Alleteo, D.: Effects of atmospheric pollution on lichen biodiversity (LB) in a Mediterranean region (Liguria, NW Italy). Environ. Pollut. **118**, 53–64 (2002)
21. Patil, G.P., Taillie, C.: An overview of diversity. In: Grassle, J.F., Patil, G.P., Smith, W., Taillie, C. (eds.) Ecological Diversity in Theory and Practice, pp. 23–48. International Co-operative Publishing House, Fairland (1979)
22. Peng, J., Muller, H.-G.: Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. Ann. Appl. Stat. **2**(3), 1056–1077 (2008)
23. Ramsay, J.O., Ramsay, J.B.: Functional data analysis of the dynamics of the monthly index of non durable goods production. J. Econ. **107**, 327–344 (2001)
24. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd edn. Springer, New York (2005)
25. Rudin, W.: Real and complex analysis. McGraw-Hill (1986)
26. Sam, A., Chamroukhi, F., Govaert, G., Aknin, P.: Model based clustering and segmentation of times series with changes in regime. Adv. Data Anal. Classif. **5**(4), 301–322 (2011)
27. Tarpey, T.: Linear transformations and the k-means clustering algorithm: applications to clustering curves. Am. Stat. **61**(1), 34–40 (2007)

The Impact of Demographic Change on Sustainability of Emergency Departments

Enrico di Bella, Paolo Cremonesi, Lucia Leporatti and Marcello Montefiori

Abstract

The progressive ageing of the population and the increasing migration flows are affecting the population structure in most of the western countries. Because of this demographic change, the demand for public services is expected to rise, creating potential problems to the economic sustainability of major public services. Our paper is focused on Accident and Emergency Departments (AEDs) services and it aims at estimating how the AED demand and costs will change adapting to the demographic trend in a specific Italian administrative region (Liguria) in the next decades (2012–2065). This is done as follows: first, we split the patients assisted over a whole year by one of the most relevant Italian AEDs into several categories per severity level (i.e., triage colour) and demographic characteristics (age span, gender, and nationality); using actual accounting data we estimate the average assistance cost per typology of patient; after we derive an estimate of the probability for each category of patient to ask for emergency assistance; finally we use official ISTAT 2012 – 2065 residential population forecasts to provide an estimate of the expected number of accesses per patient category and the overall expected AEDs' cost of the whole Liguria region. Our results suggest

E. di Bella (✉) · L. Leporatti · M. Montefiori
Departments of Economics, University of Genoa, Genoa, Italy
e-mail: edibella@economia.unige.it

L. Leporatti
e-mail: lucia.leporatti@unige.it

M. Montefiori
e-mail: montefiori@unige.it

P. Cremonesi
Emergency Department E.O. Ospedali Galliera, Genoa, Italy
e-mail: paolo.cremonesi@galliera.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_12

that, although immigration seems to be a more relevant aspect for future AEDs' sustainability than ageing, the inappropriate use of emergency departments by nonurgent patients is the biggest threat which policymakers will really have to deal with.

1 Introduction and Background

Accidents and Emergency Departments (AEDs) represent one of the most demanding hospital departments in terms of human and devices absorption and, as a consequence, they have been widely studied by the literature [9,10,26,29]. AEDs aim is to supply health care assistance to people in emergency, providing a fast and effective response towards urgent medical problems. AED use is significantly changing over time. Two main elements are affecting this change [32,37]: demographic change and the attitude of patients to consider AEDs as a source of primary services rather than the structure where emergency services are provided (we refer to the latter behaviour as inappropriate use).

Demographic factors have been long considered key drivers of future health care demand: traditionally, age, gender, and nationality are considered the three most important demographic determinants of health care and emergency care usage.[1] Most western countries are experiencing a dramatic demographic change, mainly caused by the progressive ageing of population, induced by the increase in life expectancy, and by the massive migration flows that originate from non-western countries. As a consequence elderly and foreign people are rising impressively and Italy does not make exception: according to the estimation provided by the Italian Institute for Statistics (ISTAT), the proportions of elderly (i.e., people aged more than 65) and foreign residents are expected to rise in Italy respectively from 20 to 33 % and from 9 to 23 % in the period 2013–2065. As the use of the AED is not uniformly distributed over the different demographic groups, demographic changes will affect the AED demand: older individuals tend to consume more emergency services than younger due to their worse general health status and, similarly, foreign individuals, especially Temporary Foreign Residents and non-Registered European, tend to consume more emergency services because they are socially and economically excluded from specialized expensive examinations or simply because they have a scarce knowledge of the health care system of the host country [25]. In addition, medical needs tend to differ among demographic groups: elderly people have needs generally connected to chronic health conditions while foreign patients tend to have more specific medical

[1] Previous researches show that also non demographic factors [24] can have a role in the determination of future AED sustainability. In particular, medical progress and technology may impact the type and cost of the emergency services offered [14,33] but these topics exceed the aims of this paper and will not be taken into account in the following.

needs, often connected to particular life events (e.g., pregnancy and work-related injuries).

Literature on AED use, suggests that elderly people represent the largest share of AED patients for absolute number of visits, length of staying, and costs [4,16, 36]. Generally, elderly tend also to access AED with more severe health conditions (yellow or red triage code[2]) due to their frailty and, as often occurs, the presence of multiple morbidities [31]. According to the literature [27,34,41], gender differences derive from three categories of factors: biological risk (e.g., life expectancy, medical condition), risk acquired from different lifestyles (e.g., diet, smoking habits, working condition), attitude towards health status (e.g., perception of health status). The majority of previous researches found out that women tend to use more AED services than men [7,17,43, e.g.]. Mustard et al. [30] found out that the 22 % of expenditure for female patients is connected to conditions specifically dependent to their gender (i.e., pregnancy and childbirth) while for men this percentage drops to 3 %. In addition, as women tend to live longer than men and to consume more emergency services, older women will become "frequent users" during the next decades [15,23,30,34].

For what concerns the use of emergency services among foreign people, apart from differences in biological risk derived from the country of origin [39,40], the different use of health care services between immigrants and natives is generally attributed to gaps in habits, lifestyles, and cultural differences [18] gaps in socio-economical status [12,13] and in the level of insurance coverage and barriers to access health care services [5,12]. As AEDs are usually free of charge and their services are available 24 h, immigrants tend to make large use of emergency services as they are prevented to get health care from general practitioner, private specialized examinations, or other care facilities [19]. This means that, when immigrants run into barriers to health care services or when they have a scarce knowledge on the health care system of the host country [8,35], they tend to use emergency services as substitutes of other health care services (e.g., general practitioner services, specialized visits), increasing the number of inappropriate accesses to AED.

Inappropriate accesses represent the second element affecting present and future AED usage; indeed, several researches [2,37, e.g.] show that more and more patients access AEDs for primary healthcare services rather than using general practitioner services. Previous researches found out that the number of accesses in white and green triage code (less urgency) is significantly higher than the number of accesses in yellow and red triage (high urgency) [3]: this means that inappropriate use of AED leads to organizational problems for AEDs (waiting times, overcrowding) and it may lead to a drop in the quality of services offered [6,28]. This study is intended

[2]Triage coding is the most common European classification criterion of patients at AED check-in. It is based on four colours, each of them associated to a different severity/urgency patient condition. The most critical patients are classified with a red code and they have to be attended immediately because their life is in peril. Yellow is attributed to urgent patients for which some waiting is possible. Green codes require medical care but it is not urgent. Finally, white codes are nonurgent patients. White code patients and a share of green codes are generally amenable to inappropriate use of AEDs.

to provide policymakers with new informative tools to forecast future AEDs demand and expenditure based on demographic pattern. We believe this may help in correctly programming AED activity, costs and resource allocation. With this purpose, we generalize data on cost and accesses by severity levels and demographic characteristics using data from one of the biggest Italian hospitals' AED (E.O. Ospedali Galliera) in Genoa (594,904 residents in 2014) in order to obtain the expected expenditure and number of accesses to AEDs for the entire administrative region around Genoa (Liguria district, 1,591,939 residents in 2014) for the period 2012–2065 based on the residents forecasts provided by the Italian Institute of Statistics (ISTAT).[3]

2 Data and Methods

The E.O. Galliera AED registry has been used to get information on demo-graphic characteristics (i.e., age, gender, nationality, zip code of residence) of patients accessing the AED during 2012 and on the medical condition (triage code, diagnosis) and treatments (exams and visits) connected to each access.[4] To estimate the cost associated to each access we matched Galliera registry with the official AED fees derived from the Italian Ministry of Health (Law nr. 23/2013). In Italy, these tariffs represent the standard cost established by the Ministry of Health for each AED visit or exam and they are used as a benchmark for the reimbursements given from the Government to the single hospital for each treatment. In order to estimate the cost associated to each access we consider the three main categories of direct costs imputable to each event: AED general visits, non laboratory exams (e.g., X-rays) and laboratory exams (e.g., blood analysis). If we cut abandonments, the total number of accesses to Galliera AED in 2012 was 44,969.[5] Older individuals (aged more than 65) represent the 31 % of the total number of accesses while the 21 % of the total number of accesses is due to foreign people[6] (Table 1). The vast majority (73 %) of individuals access AED in white or green triage code, while only a minor percentage (3 %) of the patients access for an acute and severe medical condition (red triage code).

Generally, older individuals tend to be classified with red or yellow codes more often than younger due to their general worse health status. On the contrary, the inappropriate use of AED is more frequent among younger individuals and, above all, among foreign people who record the largest percentage of white triage code (13 %). Several methods have been proposed to estimate the evolution of demand and cost of health care services based on demographic change [11,20,22,38,42]. Przywara

[3]This analysis is not run directly at regional level as data for the whole Liguria Region are not available to the authors.

[4]As E.O. Galliera is not a paediatric hospital and the number of children accessing is small our analysis will be focused only on adults patients (aged more than 14).

[5]The analysis is run by episode and not by patient, thus it does not take into account issues connected to re-access of the same individual more than once during the year.

[6]Our analysis only include legal resident immigrants without considering illegal immigration.

Table 1 Number of accesses by triage code and demographic group (Year 2012, Galliera AED)

		White triage	Green triage	Yellow triage	Red triage	Total	Total %
Age class	15–24	474	3,840	618	31	4,963	10
	25–44	2,034	12,430	1,943	110	16,517	33
	45–64	1,005	8,713	2,589	248	12,555	25
	65–84	576	5,773	4,192	666	11,207	22
	>85	92	1,810	2,344	481	4,727	9
Gender	Male	2,434	17,341	5,949	698	26,422	47
	Female	1,747	15,225	5,737	838	23,547	53
Nationality	Italian	2,650	25,011	10,227	1,404	39,292	79
	Foreign	1,531	7,555	1,459	132	10,677	21
Total		4,181	32,566	11,686	1,536	49,969	100
(%)		8	65	23	3	100	

et al. [32] identify three possible methods of projecting health care expenditure in the future depending on available data: time-series methods which extrapolate into the future the past observed trends; macro-simulation models which work by a disaggregation of the aggregated spending data into a number of groups homogeneous for demographic features and micro-simulation models which start from datasets on single unit (e.g., individuals, households) gathered by doctors and hospitals rather than from aggregated data.

Generally, when suitable datasets are available, micro simulation model predictions are more reliable than macro models; however, in order to have reliable results, high quality, consistent datasets are needed [21]. In this paper, we start from micro data, considering the impact on AED demand of three demographic variables (class of age, gender and nationality) and of one clinical variable (triage code); subsequently we extend results obtained from one single AED to predict future AED demand for all the administrative region around Genoa (Liguria). The estimations are computed using a pure demographic scenario (i.e., we only take into account the impact of change in size and structure of the population) and therefore demand and consumption of AED services are assumed to be constant over time, within age, gender and nationality class and to be independent of medical and technological progress [1]. We will partially relax this assumption in the last part of the paper, where a few scenarios based on possible changes in AED's use pattern among foreign patients will be investigated. The analysis consists of four steps:

1. we derive the average cost ($c_{g,a,n,tc}$) associated to each category of patients split by gender (g = male, female), 5 age classes ($a = 15-24, 25-44, 45-64,$

65 − 84, 85+), 2 nationality groups (n = Italian, Foreign) and 4 triage codes, (tc = white, green, yellow, red) for a total of 80 average group costs[7];

2. we compute the probabilities of accessing the AED during the year ($AR_{g,a,n,tc}$) for each of the 80 categories of patients as the ratio between the number of patients who actually attended the AED ($P_{g,a,n,tc}$) and the number of residents living in the catchment area of the Galliera AED ($R_{g,a,n}$)[8] (14 urban areas of the city sufficiently close to the selected AED for a total number of patients resident in the catchment area of 19,064);
3. a forecast of the estimated number of accesses for each year t of the period 2012–2065 is obtained by multiplying the access rates computed in step 2 ($AR_{g,a,n,tc}$) for each category by the expected number of residents in Liguria for each demographic group obtained by the main scenario[9] forecasts provided by ISTAT ($R^*_{g,a,n,t}$)

$$A^*_{g,a,n,tc,t} = AR_{g,a,n,tc} * R^*_{g,a,n,t} \tag{1}$$

4. we estimate future AED expenditure (and its composition by group) by multiplying the average per capita cost of accessing AED (step1) by the estimated number of accesses (step3):

$$E^*_{g,a,n,tc,t} = c_{g,a,n,tc} * A^*_{g,a,n,tc,t} \tag{2}$$

3 Results

Table 2 summarizes the expected number of accesses and the corresponding expenditure for the period 2012–2065 according to models (1) and (2). Liguria is a peculiar region under a demographic point of view: indeed, this Italian region re-cords a particularly high proportion of older individuals (28 % in 2013 compared to an Italian mean of 21 %) and the lowest fertility rate in Italy (with an average child per woman equal to 0.99). Looking at the demographic trend forecasted by ISTAT for Liguria we see that this pattern is going to persist over the next decades. The total number of residents is estimated to decrease by 8 % as a result two opposing forces: a contraction in the number Italian residents (−26 %) and an increase in foreign residents (+210 %); the number of older Italian and foreign residents (aged more than 85) are

[7]For 14 group categories of foreign people the number of observations in the group was too small (less than 20 units) to allow for inference: thus we approximated average cost for these categories equal to the one of corresponding Italian patients. This is a conservative assumption as foreign individuals tend to have higher average cost for each access.

[8]For same reasons cited in note 9, 17 access rates relative to foreign residents were conservatively set equal to the ones of corresponding Italian patients.

[9]ISTAT provides three different demographic scenarios, the low, the central and the high, based on different assumptions on the dynamics (projections of) in the number of residents over the period 2012–2065. Our analysis is mainly based on the central scenario, however, estimates using the low and high scenarios are also provided.

Table 2 Expected residents, accesses and expenditure between 2012 and 2065 in Liguria

	Age	15–24	25–44	45–64	65–84	+85	Total
Italian nationality	Residents 2012	114,010	337,403	436,557	364,719	67,308	1,319,997
	Residents 2065	93,208	213,407	269,318	266,547	134,495	976,975
	% variation	−18 %	−37 %	−38 %	−27 %	100 %	−26 %
	Accesses 2012	15,746	43,878	48,845	66,574	29,206	204,249
	Accesses 2065	12,870	27,788	30,112	48,885	59,084	178,740
	% variation	−18 %	−37 %	−38 %	−27 %	102 %	−12 %
	Expenditure 2012	1,026,660	2,803,937	4,022,307	7,607,787	4,013,475	19,474,167
	Expenditure 2065	838,781	1,773,844	2,479,044	5,590,998	8,064,371	18,747,038
	% variation	−18 %	−37 %	−38 %	−27 %	101 %	−4 %
Foreign nationality	Residents 2012	17,807	59,629	28,393	4,540	311	110,680
	Residents 2065	44,505	110,428	96,450	70,164	21,193	342,740
	% variation	150 %	85 %	240 %	1,445 %	6,714 %	210 %
	Accesses 2012	4,145	15,993	6,671	2,622	364	29,796
	Accesses 2065	10,252	29,375	23,010	40,644	24,842	128,123
	% variation	147 %	84 %	245 %	1,450 %	6716 %	330 %
	Expenditure 2012	235,326	956,571	488,982	280,144	52,922	2,013,945
	Expenditure 2065	583,180	1,760,969	1,675,157	4,346,722	3,607,212	11,973,240
	% variation	148 %	84 %	243 %	1,452 %	6,716 %	495 %
Total	Residents 2012	131,817	397,032	464,950	369,259	67,619	1,430,677
	Residents 2065	137,713	323,835	365,768	336,711	155,688	1,319,715
	% variation	4 %	−18 %	−21 %	−9 %	130 %	−8 %
	Accesses 2012	19,891	59,871	55,516	69,196	29,570	234,045
	Accesses 2065	23,122	57,163	53,122	89,529	83,926	306,863
	% variation	16 %	−5 %	−4 %	29 %	184 %	31 %
	Expenditure 2012	1,261,986	3,760,508	4,511,289	7,887,931	4,066,397	21,488,112
	Expenditure 2065	1,421,961	3,534,813	4,154,201	9,937,720	11,671,583	30,720,278
	% variation	13 %	−6 %	−8 %	26 %	187 %	43 %

estimated to boost leading to an ageing society. The impact of demographic change can be considered under three points of view: on demand, on emergency level, and of inappropriate use of AEDs.

3.1 Impact on Demand

As a consequence of the expected demographic pattern, overall, the total adult number of accesses is estimated to increase by 31 %. With the exception of accesses connected to older people (+102 %), the AED events connected to Italian residents are estimated to decrease as a consequence of the contraction of the population. On the contrary, foreign accesses (in all age categories) are estimated to increase over the period (+330 %). Older foreign residents will become a significant category of patients in the next 50 years: indeed, nowadays the number of foreign residents (and therefore patients) aged more than 85 is particularly low in Liguria but, according to ISTAT forecasts, their number is likely to increase notably as a consequence of social integration of foreign people.

3.2 Impact on Emergency Level

As triage code level is not uniformly distributed among different demographic groups, the increase in accesses will impact on the average level of emergency and thus on the services needed by patients (Table 3). The expected evolution in other triage codes show that yellow and red triage codes are estimated to increase more than white and green ones (+63 and +46 %, respectively): this phenomenon will lead to an increase in the average level of emergency and thus to an increase in services needed during urgent situations. The percentage of not severe medical condition (white and green triage code accesses) on the total number of accesses is going to decrease from roughly 70 % in 2012 to 63 % in 2065; consequently the percentage of urgent medical condition (yellow and red triage code accesses) will become more relevant moving from 30 % in 2012 to 37 % in 2065; this will increase the needs for specific services addressed to patients accessing AED in particularly vulnerable conditions (e.g., increase need for operating tables and for ambulance services).

Table 3 Expected triage code composition of accesses (2012–2065) in Liguria

Year	White triage		Green triage		Yellow triage		Red triage	
	Number	Total %	Number	Total %	Number	Total %	Number	Total %
2012	14,664	6	148,361	63	63,103	27	7,916	3
2065	17,362	6	175,035	57	102,899	34	11,567	4
% variation	18 %		18 %		63 %		46 %	

3.3 Impact on Inappropriate Use of AED

From Table 3, we can see that the total amount of white triage codes is expected to increase by 18 % over the 50-years period moving from 14,664 to 17,362 accesses. This means that the problem of inappropriate use of emergency services is going to become much more relevant during the next decades. Moreover, whereas in 2012, Italian patients represented the largest category of white triage codes (75 %), our model suggests that in 2065 this percentage will drop to 47 % and that foreign individuals will represent the largest share of inappropriate users of AEDs.

3.4 Change in Costs

Change in demand can have dramatic consequences on the departments' ability to provide adequate services to all help-seeking patients. On the other side, increase in expenditure connected to AEDs can lead to balance problems for the government; according to our estimations, total expenditure is likely to increase significantly during the next decades (+43 %). Expenditure is estimated to increase more than the number accesses (+31 %); this is caused by the vast increase in the number patients with higher average cost for each access (i.e., older and foreign). Similarly to accesses variation, older Italian and foreign residents will cause the largest increase in economic resources that should be devoted to AEDs. On the other side, the contraction in the number of younger Italian residents will bring to a reduction in costs associated to them. For what concerns triage code composition of expenditure, Table 4 shows that white triage code related expenditure is estimated to increase by 21 %, whereas red triage code expenditure will increase by 42 %. Expenditure in white triage codes is particularly relevant for policymakers as it represents the cost of inappropriate use of AEDs and thus, in theory, it is could be avoided or reduced through specific policies addressed to the fight against inappropriate accesses; one the other side, expenditure connected to urgent accesses is more difficult to be reduced without compromising the outcome for vulnerable patients. The last column of Table 4 shows how the ratio between expenditure devoted to urgent accesses (yellow and red triage) and not urgent accesses (white and green triage) is estimated to change in the next decades. This measure is increasing over time (from 0.95 in 2012 to 1.28 in 2065) meaning that the resources devoted to urgent patients are estimated to increase in the future with respect to resources devoted to less urgent patients.

3.5 Sensitivity Analysis

Due to the long run perspective of our estimations, caution is needed in the interpretation of the forecasting: several unexpected economical and political changes may affect the future demographic pattern (e.g. migration flows) over a 50-year period. For this reason, we compare the results of Table 2 with those obtained using the low and high demographic forecasting provided by ISTAT which represent, respectively,

Table 4 Expected expenditure by triage code (2012–2065) in Liguria

Year	White triage	Green triage	Yellow triage	Red triage	Total expenditure	$\frac{\text{Expend(Y+R)}}{\text{Expend(W+G)}}$
2012	523,837	10,505,834	8,903,266	1,555,176	21,488,113	0.95
2065	631,741	12,851,766	15,027,403	2,209,368	30,720,278	1.28
% variation	21 %	22 %	69 %	42 %	43 %	

W = White triage, G = Green triage, Y = Yellow triage, R = Red triage

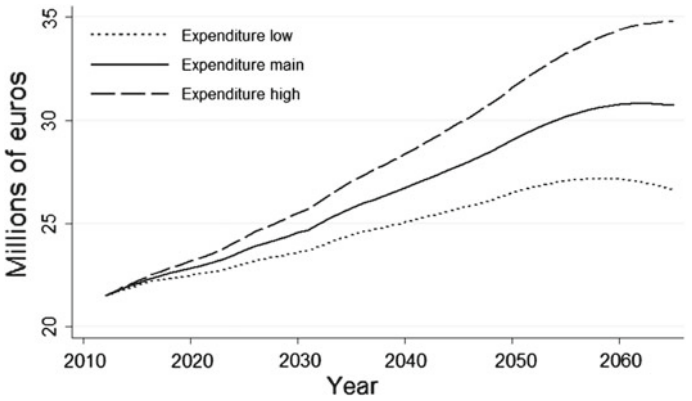


Fig. 1 Expected expenditure under the main, the low and the high demographic scenarios (millions of €)

a more and less conservative estimations of the expected number of residents (Fig. 1). The total expenditure, between 2012 and 2065 is estimated to increase by 24, 43 and 62 %, respectively under the low, the main and the high scenarios, with an expected increase in the number of accesses equal to +15 % (low scenario), +31 % (main scenario) and +47 % (high scenario). Taking this as a measure of the uncertainty connected to our estimations, we will now perform alternative scenarios based on possible changes in the use of AEDs across some demographic groups.

3.6 Scenarios Analysis

It has been showed that the expected demographic change (especially for what concerns the foreign factor) is going to affect impressively the future activity of AEDs, increasing the needs for emergency services. Even if ageing and migration flows are not under the direct control of the government, some actions could be taken into account to mitigate the expected sustainability problems.

We identify two possible actions that may significantly reduce the future demand and costs of AEDs: the convergence of foreign people to the Italian way of using AED services and the reduction in inappropriate use of AED.

The former may be pursued thanks to an increase into foreign awareness of the Italian health care system and through an increase in alternative services offered to the foreign individuals now excluded from general practitioner services (e.g., illegal immigrants). To analyze this issue we compare three scenarios in order to understand the savings in expenditure that may be obtained in 2065 controlling the foreign excessive use of emergency services through specific policies. The first scenario (i.e., Base Scenario) is our reference and it simply reports the result already discussed in Table 2. The second scenario (Alternative Scenario A) considers instead the situation in which costs and access rates of older foreign (aged more than 65) converge to those of Italian residents as a consequence of a greater level of integration and knowledge. The third scenario (Alternative Scenario B) is an extension of Scenario B in which costs and access rates of all foreign (for each age class) converge to those of Italian residents also thanks to specific policies ad-dressed towards foreign (e.g., institution of specific aid services). All this scenarios have been performed using the low, main and high ISTAT demographic forecasting.

Table 5 shows the results: we consider the estimated expenditure in 2065 under the base scenario equal to 100 %, and we compare this value with those obtainable across the described scenarios. For instance, if the evolution of the population will follow a low demographic scenario, expenditure in 2065 is estimated to be equal to the 87 % of that estimated under the main scenario. Let us consider the results obtained under the main scenario (second column of Table 5). We see that, if older foreign residents way of using AEDs will converge to the Italian way of using in terms of both accesses and cost for each access, the expenditure in 2065 will become the 83 % of that recorded under the base scenario. The total expenditure will drop to the 77 % of that recorded under a base scenario if we consider Alternative Scenario B.

This means that, proper policies aimed at reducing the differences in using services across nationality will lead to significant economic savings: the estimated expenditure for 2065 under this last scenario is equal to 23.7 millions of euros. If we consider that

Table 5 Expected expenditure by triage code (2012–2065) in Liguria

		ISTAT demographic scenario		
		Low	Main	High
Base scenario	Expenditure in 2065 (ml of €)	26.63	30.72	34.80
	% base scenario	87 %	100 %	113 %
Alternative scenario A	Expenditure in 2065 (ml of €)	22.10	25.50	28.96
	% base scenario	72 %	83 %	94 %
Alternative scenario B	Expenditure in 2065 (ml of €)	20.44	23.69	27.01
	% base scenario	67 %	77 %	88 %

the current expenditure in 2012 is estimated to be 21.5 millions of euros (Table 2), it is evident that, under Scenario B the expected increase in expenditure will account to 10 % rather than to 43 % over the period 2012–2065. The other relevant problem that policymakers will face in the next decades concerns the increase in inappropriate use of AED: inappropriate users generally are associated with low economical costs as they do not require sophisticated treatments: however, they are responsible for crowding and consuming of time resources of the AED staff which is an economical cost not included in our cost estimation. The provision of alternative services to be offered to inappropriate users of AED (e.g., first aid services) may partially mitigate the problems: according to our estimations, a reduction of 40 % in white triage codes accesses will lead to a apparently small reduction in expenditure (−1 % under the main scenario) but it may be responsible for a significant reduction in future crowding, organizational problems and expenditure for personnel. Another action that may be pursued by policymakers concerns the search for a better management of chronic conditions such as asthma, heart failure, Chronic Obstructive Pulmonary Disease (COPD). These pathologies are often connected to high rates of re-access and they could be managed through specifically addressed services outside the AED. In E.O. Galliera, these chronic pathologies represented in 2012 roughly the 4 % of total accesses (asthma = 0.3 %; heart failure = 1.7 %; COPD = 1.8 %) and they recorded an average cost for each access generally higher than the mean (asthma = 75 €; heart failure = 152 €; COPD = 140 €). If we assume that the prevalence of these chronic conditions will stay constant over time, the ability to deal with these problems outside AEDs will bring to a reduction of accesses by more than 12.000 accesses in 2065, with a total saving in terms of expenditure higher than 1.6 millions of euros.

4 Discussion and Conclusions

The results of our analysis suggest that the increase in older and foreign patients will bring to a significant increase in the number of accesses and that expenditure is going to rise more than demand over the period: the average per access cost will move from 92 to 100 euros. This fact is a consequence of the increase in accesses connected to expensive patients (i.e., red triage codes, older, foreign). The expenses of AEDs relating the treatment of all the patients are estimated (at 2012 values) to be 30.7 millions of euros in 2065 against the 21.5 of 2012 (Table 2) and the share of costs dedicated to the assistance of elderly people is expected to increase from the 56 % (12.0 millions of euros) to the 70 % (21.6 million of euros). The expenses to assist foreign patients will increase from 2 up to 12 millions of euros and a share of these future costs will be required to assist an increasing number of older foreign patients. It is therefore possible to sustain that ageing and immigration will both have a relevant and similar effect on AEDs expenses with a slightly dominant role of immigration. From a policymaking point of view we also point out that the expected in-crease in the number of white triage code (+18 %) may suggest the need for alter-native services

addressed to inappropriate users of AEDs. Future composition of inappropriate users of emergency services is an important element that should be taken into account by policymakers to address specific policies aiming at redirecting these patients toward more adequate services (e.g., general practitioner, specialized visits) through informative campaign addressed to foreign or at the institution of special first-aid services outside AEDs able to meet the needs of inappropriate users of AEDs. The analysis of different scenarios shows that policies specifically addressed towards the integration of foreign individuals may partially mitigate the future expected problems. The most relevant effect that the demographic change will cause is a further increase in the expenses for the assistance of nonurgent patients that is not the focal scope of AEDs. Moreover, the blind enlargement of AEDs to assist also nonurgent patients may have the only drawback of encouraging this undesirable behaviour of inappropriate use, making it more and more convenient for nonurgent patients. This policy may enlarge even more the offer for a demand that is potentially unlimited as it is alternative to the other services provided by the national health care systems. The first step for a future sustainability of AEDs is to reduce inappropriate accesses.

Acknowledgments The authors would like to thank E.O. Ospedali Galliera for their precious cooperation in providing the data.

References

1. Abel-Smith, B., Titmuss, R. M. and others: The Cost of the National Health Service in England and Wales. The Cost of the National Health Service in England and Wales. Cambridge: University Press, London: Cambridge University Press, Bentley House, 200, Euston Road, NWI (1956)
2. Afilalo, J., Marinovich, A., Afilalo, M., Colacone, A., Leger, R., Unger, B., Giguere, C.: Nonurgent emergency department patient characteristics and barriers to primary care. Acad. Emerg. Med. **11**(12), 1302–1310 (2004)
3. Ameri, M., Cremonesi, P., Montefiori, M.: The Effects of Inappropriate Emercency Department Use. Studi Economici. FrancoAngeli Editore, Milano (2011)
4. Aminzadeh, F., Dalziel, W.B.: Older adults in the emergency department: a systematic review of patterns of use, adverse outcomes, and effectiveness of interventions. Ann. Emerg. Med. **39**(3), 238–247 (2002)
5. Angel, R.J., Angel, J.L., Markides, K.S.: Stability and change in health insurance among older Mexican Americans: longitudinal evidence from the Hispanic established populations for epidemiologic study of the elderly. Am. J. Public Health **92**(8), 1264–1271 (2002)
6. Bamezai, A., Melnick, G., Nawathe, A.: The cost of an emergency department visit and its relationship to emergency department volume. Ann. Emerg. Med. **45**(5), 483–490 (2005)
7. Balnk, F.S., Li, H., Henneman, P.L., Smithline, H.A., Santoro, J.S., Provost, D., Maynard, A.M.: A descriptive study of heavy emergency department users at an academic emergency department revels heavy ED users have better access to care than average users. J. Emerg. Nurs. **31**, 139–144 (2005)

8. Cots, F., Castells, X., García, O., Riu, M., Felipe, A., Vall, O.: Impact of immigration on the cost of emergency visits in Barcelona (Spain). BMC Health Serv. Res. **7**(1), 9 (2007)
9. Cremonesi, P., Di Bella, E., Montefiori, M.: Cost analysis of emergency department. J. Prev. Med. Hyg. **51**(4) (2015)
10. Cremonesi, P., Di Bella, E., Montefiori, M., Persico, L.: The robustness and effectiveness of the triage system at times of overcrowding and the extra costs due to inappropriate use of emergency departments. Appl. Health Econ. Health Policy 1–8 (2015)
11. Dang, T., Antolin, P., Oxley, H.: Fiscal implication of ageing: projections of age-related spending. Organisation for Economic Co-operation and Development (OECD) Working Paper. **305** (2001)
12. Derose, K.P., Escarce, J., Lurie, N.: Immigrants and health care: sources of vulnerability. Health Aff. **26**(5), 1258–1268 (2007)
13. Dinesen, C., Nielsen, S.S., Mortensen, L.H., Krasnik, A.: Inequality in self-rated health among immigrants, their descendants and ethnic Danes: examining the role of socioeconomic position. Int. J. Public Health **56**(5), 503–514 (2011)
14. Dormont, B., Oliveira, M.J., Pelgrin, F., Suhrcke, M.: Health expenditures, longevity and growth. Florian and Suhrcke, Marc, Health Expenditures, Longevity and Growth (2008)
15. Dunlop, D.D., Manheim, L.M., Song, J., Chang, R.W.: Gender and ethnic/racial disparities in health care utilization among older adults. J. Gerontol. Ser. B: Psychol. Sci. Soc. Sci. **57**(4), S221–S233 (2002)
16. Frazier, S.C. and others: Health outcomes and polypharmacy in elderly individuals: an integrated literature review. J. Gerontol. Nurs. **31**(9), 4 (2005)
17. Fuda, K.K., Immekus, R.: Frequent users of Massachusetts emergency departments: a statewide analysis. Ann. Emerg. Med. **48**(1), 9–16 (2006)
18. Gadd, M., Johansson, S., Sundquist, J., Wändell, P.: Are there differences in all-cause and coronary heart disease mortality between immigrants in Sweden and in their country of birth? A follow-up study of total populations. BMC Public Health **6**(1), 102 (2006)
19. Gravelle, H., Morris, S., Sutton, M.: 18 Economic studies of equity in the consumption of health care. The Elgar Companion to Health Economics, vol. 193. Edward Elgar, Cheltenham (2006)
20. Gray, A.: Population ageing and health care expenditure. Ageing Horiz. **2**, 15–20 (2005)
21. Holly, A., Gardiol, L., Eggli, Y., Yalcin, T., Ribeiro, T.: Health-based Risk Adjustment in Switzerland: an exploration using medical information from prior hospitalisation. First Version of the Final Report: November (2003)
22. House of Commons Expenditure Committee: Ninth Report of the Expenditure Committee: Chapter V: Spending on the Health and Personal Social Services. HMSO. **341** (1977)
23. Irizarry, A.: Utilization of health services among the Puerto Rican elderly: gender considerations. P. R. Health Sci. J. **7**(3), 215–224 (1988)
24. LaCalle, E.J., Rabin, E.J., Genes, N.G.: High-frequency users of emergency department care. J. Emerg. Med. **44**(6), 1167–1173 (2013)
25. Levaggi, R., Montefiori, M.: Considerazioni economiche nella gestione sanitaria del mi-grante. Medicina delle migrazioni: la salute del migrante e i fattori di rischio associati. Società Italiana di Igiene, Medicina Preventiva e Sanità Pubblica Sezione Lombardia (2012)
26. Levaggi, R., Montefiori, M.: Definition of a prospective payment system to reimburse emergency departments. BMC Health Serv. Res. **13**(1), 409 (2013)
27. Macintyre, S., Hunt, K., Sweeting, H.: Gender differences in health: are things really as simple as they seem? Soc. Sci. Med. **42**(4), 617–624 (1996)
28. McCarthy, M.L., Aronsky, D., Jones, I.D., Miner, J.R., Band, R.A., Baren, J.M., Desmond, J.S., Baumlin, K.M., Ding, R., Shesser, R.: The emergency department occupancy rate: a simple measure of emergency department crowding? Ann. Emerg. Med. **51**(1), 15–24 (2008)
29. Moskop, J.C., Sklar, D., Geiderman, J.M., Schears, R.M., Bookman, K.J.: Emergency department crowding, part concept, causes, and moral consequences. Ann. Emerg. Med. **53**(5), 605–611 (2009)

30. Mustard, C.A., Kaufert, P., Kozyrskyj, A., Mayer, T.: Sex differences in the use of health care services. New Engl. J. Med. **338**(23), 1678–1683 (1998)
31. Peters, M.: The older adult in the emergency department: aging and atypical illness presentation. J. Emerg. Nurs. **36**(1), 29–34 (2010)
32. Przywara, B. and others: Projecting future health care expenditure at European level: drivers, methodology and main results. Directorate General Economic and Monetary Affairs (DG ECFIN), European Commission (2010)
33. Polder, J.J., Bonneux, L., Meerding, W.J., Van Der Maas, P.J.: Age-specific increases in health care costs. Eur. J. Public Health **12**(1), 57–62 (2002)
34. Redondo-Sendino, A., Guallar-Castillón, P., Banegas, J., Rodríguez-Artalejo, F.: Gender differences in the utilization of health-care services among the older adult population of Spain. BMC Public Health **6**(1), 155 (2006)
35. Rubio, D.: Equity in the use of health care services by immigrants in the Spanish national health care system. Estudios de Economía Aplicada **50**, 26–33 (2008)
36. Samaras, N.C.T., Samaras, D., Gold, G.: Older patients in the emergency department: a review. Ann. Emerg. Med. **56**(3), 261–269 (2010)
37. Sempere-Selva, T., Peiró, S., Sendra-Pina, P., Martínez-Espín, C., López-Aguilera, I.: Inappropriate use of an accident and emergency department: magnitude, associated factors, and reasons? An approach with explicit criteria. Ann. Emerg. Med. **37**(6), 568–579 (2001)
38. Smith, J.P.: Healthy bodies and thick wallets: the dual relation between health and economic status. J. Econ. Perspect.: J. Am. Econ. Assoc. **13**(2), 144 (1999)
39. Sol-Auró, A.S., Crimmins, E.M.: Health of Immigrants in European countries. Documents de Treball (IREA). Institut de Recerca en Economia Aplicada. **9**, 1 (2008)
40. Soléi A.A., Guillén, M., Crimmins, E.M.: Health care utilization among immigrants and native-born populations in 11 European countries. Results from the Survey of Health, Ageing and Retirement in Europe. IREA–Working Papers, 2009, IR09/020. Universitat de Barcelona. Institut de Recerca en Economia Aplicada Regional i Pública (2009)
41. Verbrugge, L.M.: Gender and health: an update on hypotheses and evidence. J. Health Soc. behav. 156–182 (1985)
42. Vilpert, S., Ruedin, H.J., Trueb, L., Monod-Zorzi, S., Yersin, B., Büla, C.: Emergency department use by oldest-old patients from 2005 to 2010 in a Swiss university hospital. BMC Health Serv. Res. **13**(1), 344 (2013)
43. Zuckerman, S., Shen, Y.: Characteristics of occasional and frequent emergency department users: do insurance coverage and access to care matter? Med. Care **42**(2), 176–182 (2004)

Bell-Shaped Fuzzy Numbers Associated with the Normal Curve

Fabrizio Maturo and Francesca Fortuna

Abstract

Statisticians often focus on fuzzy numbers with triangular or trapezoidal membership functions because they are very easy to apply. Although they offer a good approximation of a fuzzy variable, several doubts arise about the appropriateness of these kind of shapes. As known, fuzzy sets are useful for interval data when the "degree of truth" of the values varies within this range. In particular, they are desirable for translating human language into numbers. In this paper, we propose an alternative membership function that appears more appropriate to deal with linguistic variables. We refer to this function as "bell-shaped fuzzy number associated with the normal curve". In particular, we highlight the specific properties of the proposed fuzzy number and illustrate the utility of linking this function with the normal distribution.

F. Maturo (✉)
Department of Management and Business Administration, University G. D'Annunzio, Chieti and Pescara, Italy
e-mail: f.maturo@unich.it

F. Fortuna
Department of Philosophical, Pedagogical and Quantitative Economic Sciences, University G. D'Annunzio, Chieti and Pescara, Italy
e-mail: francesca.fortuna@unich.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_13

1 Introduction

Science is based entirely on Aristotelian logic, thus it gives an image of an hard-edged world in which things change insensibly. The assumption of bivalence is the basis of the scientific mentality, but the reality is quite different. Indeed, statisticians and mathematicians often use rigid conventions to deal with real phenomena [12]. This problem has been solved for the first time by Prof. Lotfi A. Zadeh by introducing the concept of degree of membership [24].

Fuzzy logic challenges and changes the concept of binary logic, according to which a predicate can have only two alternative states such as true or false, black, or white. The latter is the basis of computers operations but anyone can evaluate how inaccurate and inconsistent with the reality it may be. The fuzzy approach solves these issues by eliminating sharp edges, blurring the boundaries and overcoming the paradoxes of choosing whether an element belongs to a set rather than another.

From a statistical point of view, the indeterminacy may derive from four main aspects: imprecision related to measurement of phenomena, vagueness of language, ignorance about the values of a phenomenon and, finally, the link between the observed data and the universe of possible data.

Some variables, for their own nature, are better described by a pair of ordered values, like daily temperatures (in terms of minimum and maximum) or financial data (in terms of opening and closing daily prices); thus, by summarizing this measurement with a single value, there is a loss of information. In these situations, data are better described by interval values rather than single values. The margin of error in the value of measurement, that refers to the lack of knowledge about the value of a parameter, is knows as "imprecision" [22]; interval arithmetic analyzes this type of imprecision. If the intervals has no sharp boundaries, there is a different kind of imprecision called "vagueness"; fuzzy set theory is the right tool for the analysis of vague concepts [26]. In particular, fuzzy logic is suitable to deal with variables affected by vagueness of human language. In case of lack of knowledge about the occurrence of some event whose result is not known in advance, we have random variables; in this case, we talk about "randomness" [9].

In the past decades, it has been observed a substantial misunderstanding between probabilistic and fuzzy approaches; in fact, often the membership values have been confused with probabilities and the membership functions with probability distribution [10]. Nowadays, in the literature, this conceptual distinction is widely accepted and statistical data analysis can be conducted using interval arithmetic, fuzzy set theory, or the probabilistic approach according to the nature of the inspected variables.

Fuzzy logical rules found numerous applications in classification [6], regression [4,5,16], approximation and control problems [7,8,11,17,19,20]. In most of the real applications, both stochastic and deterministic uncertainties exist simultaneously. However, the traditional fuzzy theory and probabilistic models are only good at processing one aspect of uncertainties. Thus, several researchers integrated the probability theory with the fuzzy theory [14,18,20].

A major advantage of using this approach is that it tends to overcome some typical limitations of the classical one, such as the introduction of restrictive assumptions

about the nature and distribution of the data [25]. On the other hand, a main drawback is the fuzzification process, that consists in translating a nuanced variable into real numbers; of course, this phase is crucial because it can affect the whole analysis. A membership function (MF) is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. The relations between input uncertainties and fuzzy rules are systematically explored and several new types of membership functions discovered [3]. Fuzzy numbers are mathematical tools introduced in the context of fuzzy logic to codify vague data. Symmetric and not symmetric triangular fuzzy numbers or trapezoidal membership function are largely used for their simplicity, but it is widely shared that they are not always suitable and present some drawbacks; for example, sometimes, they represent a forced approximation of real phenomena that could be better described by different functions [3]. Another important problem is that triangular fuzzy numbers converge very quickly to low values of the "degree of truth" because they are not smooth. Also, Gaussian and bell membership functions are popular methods for specifying fuzzy sets; thus, in the literature, several kinds of bell-shaped fuzzy numbers have been proposed. For different reasons, they are often preferred to triangular ones even if they present some algebraic issues. One of the main reasons is that both of these curves have the advantage of being smooth and nonzero at all points [3].

There is a direct, although rarely explored, relation between uncertainty of input data and fuzziness expressed by membership functions. For this reason, we propose a specific bell-shaped membership function and suggest a link with the normal distribution. This provides a connection between the concepts of fuzziness and probability and helps researchers in the alpha-cut choice. The paper is organized as follows: in Sect. 2 we provide the definition and the properties of a fuzzy number and we present a brief review of the most common membership functions. In Sect. 3, we highlight some remarks about fuzziness and some drawbacks of triangular fuzzy numbers. In Sect. 4, we introduce the bell-shaped membership function associated to the normal distribution and we focus on its properties. In the same Section we propose a new index of uncertainty and its characteristics. In Sect. 5, we present some conclusions and discuss some possible limitations of our approach.

2 Fuzzy Numbers

A fuzzy number (FN) $\mu(x)$ is a special case of a fuzzy set; thus, it is a fuzzy set, defined on real numbers, with a normal and convex membership function, such that there exists at least one point where the membership function takes the value "one" [2,13,27]; it is a very useful tool for the analysis of imprecise numerical quantities, such as "approximately 3," "close to 3," "many," etc. [24]. A fuzzy number is a function having as domain the set of real numbers and with values in [0, 1]

$$\mu : \mathbb{R} \rightarrow [0, 1] \tag{1}$$

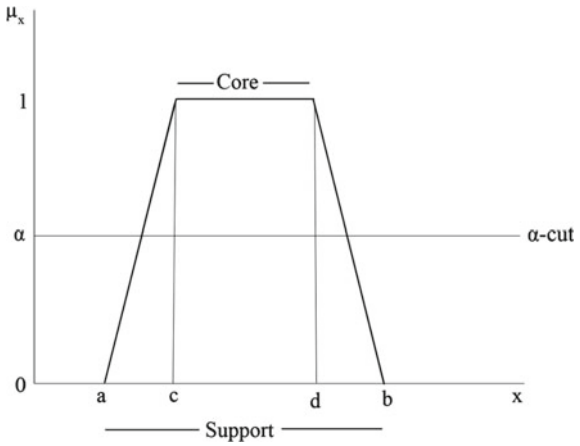


Fig. 1 The structure of a fuzzy number

with the following characteristics (1):

- Bounded support: there are two real numbers a and b, with $a \leq b$, called the endpoints of μ, such that:

$$\begin{cases} \mu(x) = 0 & \text{for} \quad x \notin [a, b] \\ \mu(x) > 0 & \text{for} \quad x \in (a, b); \end{cases} \tag{2}$$

- Normality: there are two real numbers c and d, with $a \leq c \leq d \leq b$ such that:

$$\mu(x) = 1 \qquad \text{if and only if} \qquad x \in [c, d]. \tag{3}$$

- Convexity: $\mu(x)$ is a function increasing in the interval [a, c] and decreasing in the interval [d, b];
- Compactness: for every $\alpha \in (0, 1)$, the set $\{x \in \mathbb{R} : \mu(x) = \alpha\}$ is a closed interval.

As shown in Fig. 1, the set of the real numbers x such that $\mu(x) > 0$ is said the support of the fuzzy number, and the interval [c, d] is said the core or central part of it. The intervals [a, c) and (d, b] are, respectively, the left part and the right part. The real numbers $\mu_L = c - a$, $\mu_C = d - c$, and $\mu_R = b - d$ are the left, middle, and right spreads, respectively. Their sum $\mu_T = b - a$ is the total spread of the fuzzy number. For every α such that $0 \leq \alpha \leq 1$ the set of the $x \in [a, b]$ such that $\mu(x) \geq \alpha$ is said $\alpha - cut$ of the fuzzy number [9].

The main membership functions representing fuzzy variables are triangular, trapezoidal, bell-shaped fuzzy numbers, fuzzy numbers with a flat. In the following section, we define the triangular fuzzy number (TFN) and the bell-shaped fuzzy number (BSFN), because the first is the most used and the second is a topic of this research.

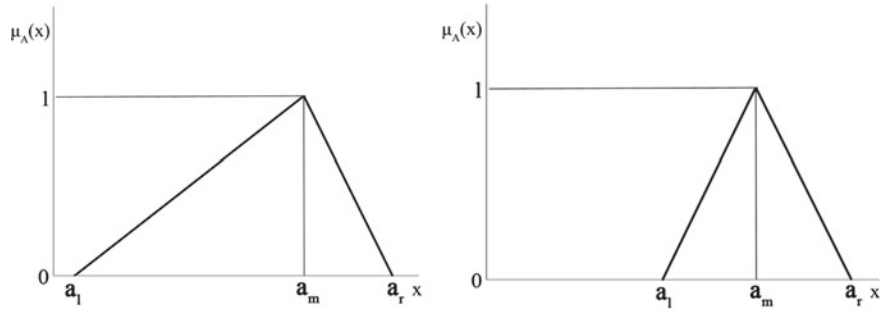


Fig. 2 Triangular fuzzy numbers and symmetric triangular fuzzy numbers

2.1 Triangular Fuzzy Numbers

A triangular fuzzy number (TFN) A is defined by the following membership function [21,23,24]:

$$\mu_A(x) = \begin{cases} \frac{x-a_l}{a_m-a_l} & \text{for} \quad a_l \leq x \leq a_m \\ \frac{x-a_r}{a_m-a_r} & \text{for} \quad a_m \leq x \leq a_r \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $[a_l, a_r]$ is the support and a_m is the core. As illustrated in Fig. 2 a_l and a_r are respectively the left and right endpoints while a_m is the point where the membership function is equal to one.

A TFN is often indicated using a simple notation like the following:

$$A = (a_l, a_m, a_r)$$

A particular kind of TFN is the central triangular fuzzy number (or symmetric triangular fuzzy number - STFN). This type of membership function is often used in applications such as managerial decision making, social science and fuzzy controllers; the membership functions of two TFNs is shown in Fig. 2.

2.2 Bell-Shaped Fuzzy Numbers

In this paper, we focus on bell-shaped fuzzy numbers (BSFNs). Different type of BSFNs have been created in the literature but the most used is the Gaussian bell-shaped [1]. It is known that the normal distribution is defined by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot(\frac{x-\mu}{\sigma})^2}$$

where μ is the mean and the σ standard deviation.

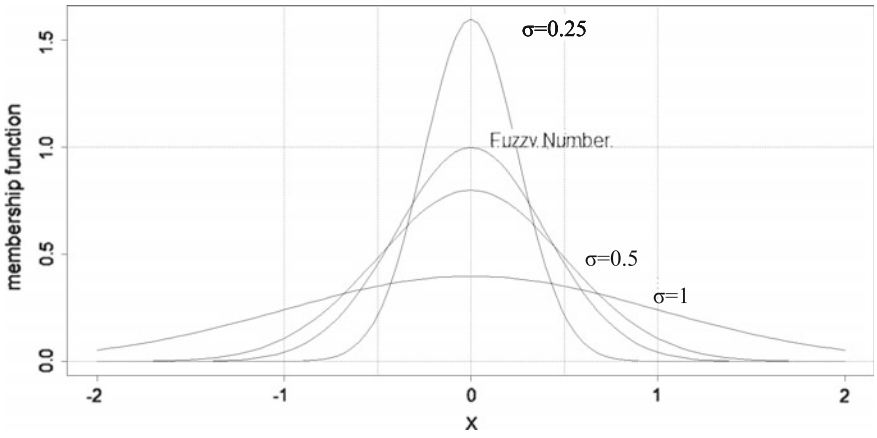


Fig. 3 The normal distribution and the normal fuzzy number

In Fig. 3, there are three examples of the variation of the shape of the normal function for $\sigma = 0.25, \sigma = 0.5, \sigma = 0.1$.

It is easy to show that for $\sigma = \frac{1}{\sqrt{2\pi}}$ the peak becomes $(\mu, 1)$; so, since a fuzzy number must have a membership function which assumes a maximum of 1, the membership function is defined by

$$\mu_A(x) = e^{-\pi(x-\mu)^2} \tag{5}$$

In this case, the parameter that determines the shape of the fuzzy number is only μ; in fact, it does not matter how it changes but the maximum is always 1.

This type of fuzzy number has the advantage of being very suitable for some real variables, but it is little used in practice because it is not easy to treat as the triangular fuzzy numbers. Furthermore, the supporting interval is $[-\infty, +\infty]$ so this kind of fuzzy number is unbounded.

3 The Choice of a Membership Function

The majority of scholars over the years have preferred to focus on triangular fuzzy numbers. Surely, this type of membership function is much easy to use and often gives a good approximation of real phenomena; nevertheless, it must be emphasized that a different shape for membership function would be preferable in some circumstances even if more difficult to treat.

There is a direct, although rarely explored, relation between uncertainty of input data and fuzziness expressed by membership functions. The assumptions about the type of input uncertainty distributions change the discontinuous mappings provided by crisp logic systems into more smooth mappings that are implemented in a natural

way by fuzzy rules using specific types of membership functions. Different assumptions about input uncertainty lead in a natural way to different types of membership functions [3].

Triangular fuzzy numbers present some drawbacks. The first drawback of TFNs is that they have a constant and linear rate of increase and decrease, before and after the center; this strong assumption is not suited for all kind of data, in particular for linguistic variables. Thus, in some circumstances, TFNs are forced approximations of real phenomena. A shared practice in the literature, is to transfer the uncertainty in the choice of the input variable rather than in the fuzzy rules, because it is much more intuitive. At this point the choice of an adaptable function is crucial. This issue highlights a second limit of TFNs that is the choice of the support for the membership function: once the researcher choices the support of a fuzzy number excludes a-priori the possibility, even minimal, that a fuzzy number can assume values that go beyond the support. In practical applications it is more convenient to transfer this choice in deciding the alpha cut, to control the fuzziness, ie the variability in the number fuzzy. A third aspect to be emphasized is a direct consequence of the practical applications. Indeed, in many real cases, analytical formulas for fuzzy membership functions have been derived using Monte Carlo methods; in fact, generalizing the results, a good guiding principle is to require that probabilities generated from Monte Carlo sampling should be the same as those obtained from the equivalent fuzzy system [3]. Thus, [3] demonstrated that, dealing with real phenomena, a good estimation of input uncertainty is often given by bell-shaped fuzzy numbers.

For these reasons, in this paper, we focus on bell-shaped membership functions. BSFNs are popular methods for specifying fuzzy sets and have the advantage of being smooth and nonzero at all points. Although more difficult to handle, these characteristics make them more realistic in some practical cases because BSFNs solve the three problems mentioned above for TFNs.

The practical usefulness of FNs is visible when we cut the number and consider only a part of the support. The cut is made with the choice of an alpha-cut, that sets the minimum degree of truth of the possible values. Of course, this choice greatly reduces the fuzziness. If we consider the whole support of a FN and we multiply it, or even raise to a power, we will have a greater spread. Large spreads mean useless information; indeed, taking all the support of a fuzzy number does not make much sense in practice. Another advantage of BSFNs is a direct consequence of the alpha-cut. If we cut a TFN, we get a decrease of fuzziness proportionally to the point where we cut because the decrease of the membership function is constant; on the contrary, the cut of a rounded shape keeps the values which are actually more "true" and the choice is more easy, because the trend of the function is not constant and gives an important indication. Therefore, the cut of a BSFNs leads to an higher saving of the "truest" part of the support and to a reduction of the fuzziness more meditated. In other words, if we cut a BSFNs, we delete many unnecessary values and keep a good level of credibility.

The great problem of BSFNs is that they are unbounded; this circumstance is in contradiction with the definition of a FN. Indeed, several researchers assert that a FN

should have a finite support and could not be unbounded. Moreover, a FN without boundaries is useless in practical application because it has infinite fuzziness.

For this reasons, a new membership function has been proposed; this tool approximates the normal distribution and is conceived to be cut and bounded in a range. In the next section, there will be also presented a link to the normal curve that allows to choose the right level of fuzziness and the point of the alpha-cut.

4 Bell-Shaped Fuzzy Numbers Associated with the Normal Curve

We introduce a particular kind of membership function such that, for increasing alpha level, its values close to the center are better than the corresponding values in a triangular fuzzy number. We consider a family of functions defined for a certain $k > 0$.

$$F_{\mu,\sigma,k}(x) = \begin{cases} e^{-\frac{(x-\mu)^2}{2\sigma^2}} & \text{in} \quad [\mu - k\sigma, \mu + k\sigma] \\ 0 \quad \text{otherwise} \end{cases} \tag{6}$$

Figure 4 shows the fuzzy number and its support. For $k > 1$ we have two inflection points, while for $k \leq 1$ the inflection points are out of the considered support.

The function $F_{\mu,\sigma,k}$ satisfies the conditions for the definition of a fuzzy number. In fact it is a function having as domain the set of real numbers and with values in [0, 1],

$$F_{\mu,\sigma,k} : \mathbb{R} \to [0, 1] \tag{7}$$

and complies with the conditions [15]:

- Bounded support: there are two real numbers $a = \mu - k\sigma$ and $b = \mu + k\sigma$, with $a \leq b$, called the endpoints of $F_{\mu,\sigma,k}$, such that:

$$\begin{cases} F_{\mu,\sigma,k}(x) = 0 & \text{for} \quad x \notin [a, b] \\ F_{\mu,\sigma,k}(x) > 0 & \text{for} \quad x \in (a, b); \end{cases} \tag{8}$$

- Normality: there is a real numbers μ, such that:

$$F_{\mu,\sigma,k} = 1 \tag{9}$$

- Convexity: $F_{\mu,\sigma,k}$ is a function increasing in the interval $[a, \mu]$ and decreasing in the interval $[\mu, b]$;
- Compactness: for every $\alpha \in (0, 1)$, the set $\{x \in \mathbb{R} : F_{\mu,\sigma,k} \geq \alpha\}$ is a closed interval.

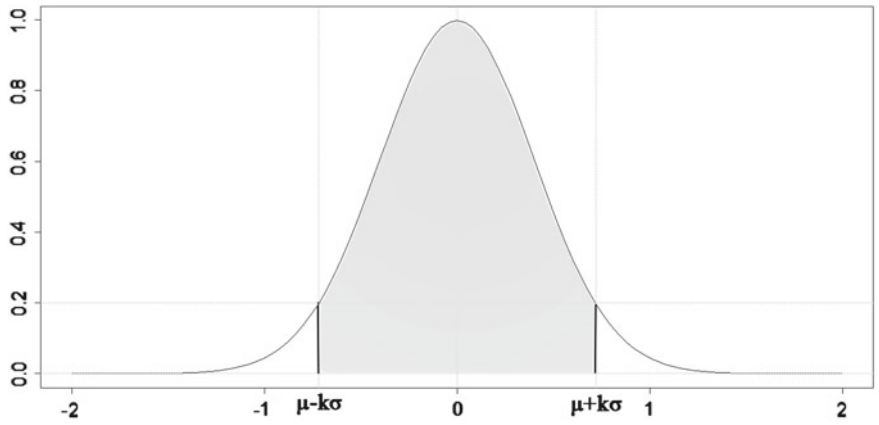


Fig. 4 Fuzzy number associated to G

Let $G_{\mu,\sigma}$ be the density function of the normal distribution with mean μ and variance σ^2. For every $x \in [\mu - k\sigma, \mu + k\sigma]$ we have:

$$F_{\mu,\sigma,k}(x) = \sigma\sqrt{2\pi}G_{\mu,\sigma}(x) \tag{10}$$

Then, what differs between the fuzzy number $F_{\mu,\sigma,k}$ and this density function $G_{\mu,\sigma}$ are the height and so the area defined by the function.

The function $F_{\mu,\sigma,k}(x)$ could be called "fuzzy number associated to $G_{\mu,\sigma}$".

We indicates with $\Phi(z)$ and $\varphi(z)$ the cumulative distribution function and the probability density function for the standard normal distribution.

We know that

$$\varphi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$$

We can prove that

$$F_{\mu,\sigma,k}(\mu + \sigma z) = \sqrt{2\pi}\varphi(z) \qquad \forall z \epsilon [-k, +k] \tag{11}$$

then in the endpoints of the support

$$F_{\mu,\sigma,k}(\mu - \sigma k) = \sqrt{2\pi}\varphi(-k) \tag{12}$$

$$F_{\mu,\sigma,k}(\mu + \sigma k) = \sqrt{2\pi}\varphi(k) \tag{13}$$

The endpoint are corresponding to α-cut with coordinates $(\mu - k\sigma, \sqrt{2\pi}\varphi(k))$ and $(\mu + k\sigma, \sqrt{2\pi}\varphi(k))$ and so the heights (ordinates of the α-cut points) depend only from k and are independent from μ and σ from a mathematical point of view.

The grey area is given by the equation

$$A_{\mu,\sigma,\kappa} = \int_{\mu-k\sigma}^{\mu+k\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma\sqrt{2\pi}[\Phi(k) - \Phi(-k)] \tag{14}$$

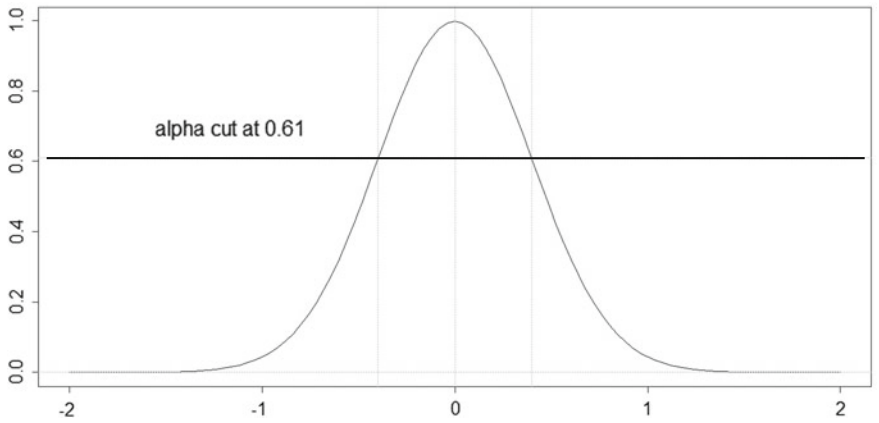


Fig. 5 Alpha-cut at 0.61 for k = 1

If we choose $k = 1$ then

$$\sqrt{2\pi}\varphi(k) \sim 0.61$$

and

$$\Phi(k) - \Phi(-k) \sim 0.683$$

We can note that it is an α-cut too much high because the area of the associate normal standard distribution area is almost 68 %.

We know that $\Phi(k) - \Phi(-k)$ is the area under the normal standard curve. Thus, the corresponding area of the fuzzy number is equal in percentage. For calculating the area of the associate fuzzy number's membership function we need multiply the area of the associate standard normal distribution with $\sigma\sqrt{2\pi}$. For this reason, the case with $k \leq 1$ is lesser interested because it covers less than 70 % of the corresponding area of the normal distribution (Fig. 5).

For $k = 2$ (Fig. 6),

$$\sqrt{2\pi}\varphi(k) \sim 0.36$$

and α-cut is lower while

$$\Phi(k) - \Phi(-k) \sim 0.954$$

We can note that the support of the area of the fuzzy number starts to increase.

For $k = 3$,

$$\sqrt{2\pi}\varphi(k) \sim 0.011$$

and α-cut is lower while

$$\Phi(k) - \Phi(-k) \sim 0.9974$$

Figure 7 shows that this alpha-cut choice corresponds to a very high value of the normal curve area.

The area of the fuzzy number coincides with the area of the associate normal distribution if $\sigma\sqrt{2\pi} = 1$, it is lower if $\sigma\sqrt{2\pi} < 1$ and it is greater if $\sigma\sqrt{2\pi} > 1$.

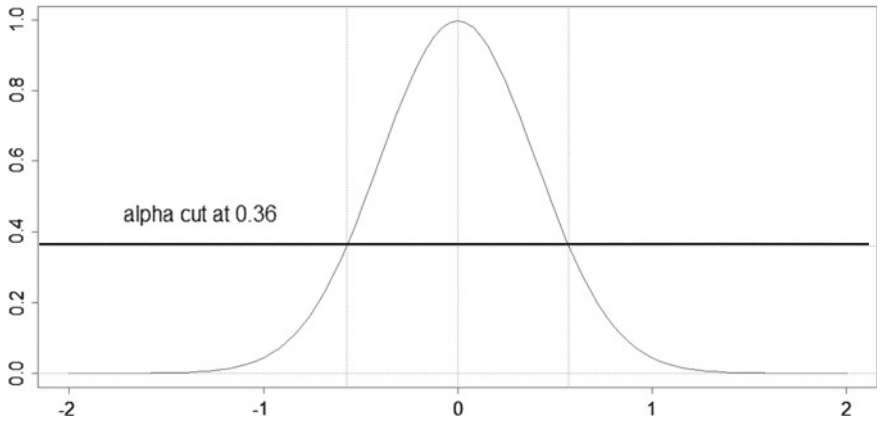


Fig. 6 Alpha-cut at 0.36 for k = 2

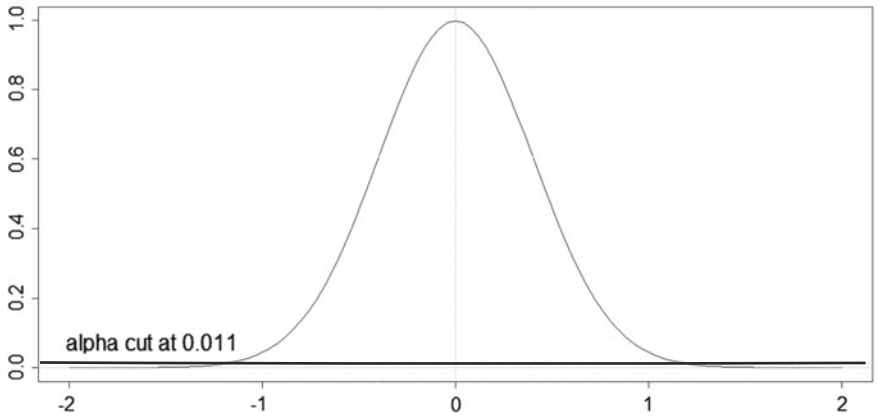


Fig. 7 Alpha-cut at 0.011 for k = 3

In fact, we obtain:

$$\begin{aligned} \sigma\sqrt{2\pi} = 1 \qquad & A_{\mu,\sigma,k} \equiv \Phi(k) - \Phi(-k) \\ \sigma\sqrt{2\pi} > 1 \qquad & A_{\mu,\sigma,k} > \Phi(k) - \Phi(-k) \\ \sigma\sqrt{2\pi} < 1 \qquad & A_{\mu,\sigma,k} < \Phi(k) - \Phi(-k) \end{aligned} \tag{15}$$

We could build an index making the relationship between area and spread as follows:

$$I_{\mu,\sigma,k} = \frac{A_{\mu,\sigma,k}}{S_{\mu,\sigma,k}} \tag{16}$$

where $S_{\mu,\sigma,k}$ is the total spread of the fuzzy number.

We can write:

$$I_{\mu,\sigma,k} = \frac{\sigma\sqrt{2\pi}[\Phi(k) - \Phi(-k)]}{2k\sigma} \tag{17}$$

then

$$I_{\mu,\sigma,k} = \frac{\sqrt{\pi}}{\sqrt{2}} \frac{\Phi(k) - \Phi(-k)}{k} \tag{18}$$

We can conclude that $I_{\mu,\sigma,k}$ is independent from μ and σ and depends only from k; thus we can write only I_k.

This result shows that, for every fixed k, the family of fuzzy numbers $F_{\mu,\sigma,k}$ is an homogeneous class of fuzzy numbers as the set of triangular fuzzy numbers. This circumstance justifies the treatment of fuzzy data replacing fuzzy triangular numbers with fuzzy numbers $F_{\mu,\sigma,k}$, with a fixed k.

Calculating this index for $k = 3$ we get:

$$\begin{aligned} I_3 &= \frac{A(k)}{S(k)} = \frac{\sigma\sqrt{2\pi}}{2k\sigma}[\phi(k) - \phi(-k)] = \\ &= \frac{\sqrt{2\pi}}{2k} 0.9974 = 0.4166 \end{aligned} \tag{19}$$

It is a nice discovery that confirms our conjecture. In fact, this relationship is always constant and it is influenced only by k. This is a very good tool to compare different fuzzy numbers.

For example, if we calculate the same index on a square we obtain a value of 1; it means that with a square we have maximum uncertainty. If we solve the same equation for a triangle we obtain a value of 0.5. Thus we can say that this index is an indicator of uncertainty. It can range from zero to one. It is 1 when we have a square and it assumes the value 0 for a scalar; in fact we should have maximum certainty.

Therefore, a value of 0.416 shows that we gain almost 20 % in relative terms compared to the triangle.

Calculating I_k for $k = 2$ we obtain $I_2 \sim 0.6$. This demonstrates that $K = 3$ is a good choice for this function.

5 Conclusions

Gaussian and bell membership functions are popular methods for specifying fuzzy sets. Both of these curves have the advantage of being smooth and nonzero at all points. Moreover, several researchers demonstrated that, dealing with real phenomena, a good estimation of input uncertainty is often given by bell-shaped fuzzy numbers. For these reasons, in some real cases, BSFNs are better than TFN because they better reflect the input variables. It is true that these functions are more difficult to treat than those triangular, but in recent decades literature has moved significantly on this field to search for practical applications. It is widely shared by researchers that there is a direct, although rarely explored, relation between uncertainty of input data and fuzziness expressed by membership functions. To fulfill this gap, this paper proposes a direct link between BSFNs and the normal distribution and, provides an index of fuzziness to help in the alpha-cut choice. Of course, various other BSFNs may

be considered, and their membership functions found, but, although the Gaussian membership functions and the bell membership functions achieve smoothness, the main drawback is that they are unable to specify asymmetric membership functions, which are important in certain applications.

References

1. Bojadziev, G., Bojadziev, M.: Fuzzy Set, Fuzzy Logic, Applications. World Scientific Publishing, New York (1995)
2. Dubois, D., Prade, H.: Fuzzy Set and Systems. Academic Press, New York (1980)
3. Duch, W.: Uncertainty of data, fuzzy membership functions, and multilayer perceptrons. IEEE Trans. Neural Netw. **16**(1), 10–23 (2005)
4. Ferrraro, M., Colubi, A., Gonzales-Rodriguez, A., Coppi, R.: A linear regression model for imprecise response. Int. J. Approx. Reason. **21**, 759–770 (2010)
5. Ferrraro, M., Colubi, A., Gonzales-Rodriguez, A., Coppi, R.: A determination coefficient for a linear regression model with imprecise response. Environmetrics **22**, 516–529 (2011)
6. Ghosh, S., Dubey, S.K.: Comparative analysis of K-means and fuzzy c-means algorithms. Int. J. Adv. Comput. Sci. Appl. **4**(4), 35–39 (2013)
7. Kasabov, N.K.: Foundations of Neural Networks, Fuzzy Systems and Knowledge Engineering. MIT Press Cambridge, London (1996)
8. Kecman, V.: Learning and Soft Computing. MIT Press, Cambridge (2001)
9. Klir, G.J.: Uncertainty and Information: Foundations of Generalized Information Theory. Wiley, New York (2006)
10. Kosko, B.: Fuzziness vs probability. Int. J. Gen. Syst. **17**(2–3), 211–240 (1990)
11. Kosko, B.: Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence. Engelwood Cliffs, New Jork (1992)
12. Kosko, B.: Fuzzy Thinking: The New Science of Fuzzy Logic. Hyperion, New York (1993)
13. Maturo, A.: On some structures of fuzzy numbers. Iran. J. Fuzzy Syst. **6**, 49–59 (2009)
14. Maturo, A., Maturo, F.: Fuzzy events, fuzzy probability and applications in economic and social sciences. In: Maturo, A., Hošková-Mayerová, Š., Soitu, D.-T., Kacprzyk, J. (eds.) Recent Trends in Social Systems: Quantitative Theories and Quantitative Models. Studies in Systems, Decision and Control, vol. 66, pp. 236–247. Springer International Publishing (2016). doi:10.1007/978-3-319-40585-8_20
15. Maturo, A., Maturo, F.: Research in Social Sciences: Fuzzy Regression and Causal Complexity. In: Ventre, A. G. S., Maturo, A., Hošková-Mayerová, Š., Kacprzyk, J. (eds.) Multicriteria and Multiagent Decision Making with Applications to Economic and Social Sciences. Studies in Fuzziness and Soft Computing, vol. 305, pp. 237–249. Springer Berlin Heidelberg (2013). doi:10.1007/978-3-642-35635-3_18
16. Maturo, F., Hošková-Mayerová, Š.: Fuzzy regression models and alternative operations for economic and social sciences. In: Maturo, A., Hošková-Mayerová, Š., Soitu, D.-T., Kacprzyk, J. (eds.) Recent Trends in Social Systems: Quantitative Theories and Quantitative Models. Studies in Systems, Decision and Control, vol. 66, pp. 236–247. Springer International Publishing (2016). doi:10.1007/978-3-319-40585-8_21
17. Mendel, J.: Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions. Prentice Hall, New Jersey (2001)
18. Nather, W.: Regression with fuzzy random data. Comput. Stat. Data Anal. **51**, 235–252 (2006)
19. Pal, S.K., Mitra, S.: Neuro-Fuzzy Pattern Recognition. Wiley, New York (1999)

20. Ramos-Guajardo, A., Colubi, A., Gonzales-Rodriguez, A.: One-sample tests for a generalized Fréchet variance of a fuzzy random variable. Metrika **71**, 185–202 (2010)
21. Shapiro, A.F.: Fuzzy regression models. ARC (2005)
22. Ventre, A.G.S.: Imprecisione e sfocatura (fuzziness). In: AA.VV.: Insiemi sfocati e decisioni, E.S.I., Napoli, pp. 11–21 (1983)
23. Yang, M.S., Ko, C.H.: On a class of fuzzy c-numbers clustering procedures for fuzzy data. Fuzzy Sets Syst. **84**(I), 49–60 (1996)
24. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
25. Zadeh, L.A.: Fuzzy algorithms. Inf. Control **12**, 94–102 (1968)
26. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning. Inf. Sci. **8**, 199–249 (1975)
27. Zimmerman, H.J.: Fuzzy Set Theory and Its Application. Kluwer Academic Publishers, Boston (1996)

Improving Co-authorship Network Structures by Combining Heterogeneous Data Sources

Vittorio Fuccella, Domenico De Stefano, Maria Prosperina Vitale and Susanna Zaccarin

Abstract

The present paper aims at describing the scientific collaboration patterns of the Italian academic statisticians by merging bibliographic data from heterogenous sources–ISI-WoS, Current Index to Statistics, and the database of nationally funded research projects, PRIN. To obtain a unified database, containing both top international as well as nationally oriented production, information were combined by identifying and linking duplicate records, i.e. record linkage. The unique co-authorship network was then used as basis for network analysis.

V. Fuccella (✉)
Department of Informatics, University of Salerno, Salerno, Italy
e-mail: vfuccella@unisa.it

D. De Stefano
Department of Economics, Ca' Foscari University of Venice, Venice, Italy
e-mail: domenico.destefano@unive.it

M.P. Vitale
Department of Economics and Statistics, University of Salerno, Salerno, Italy
e-mail: mvitale@unisa.it

S. Zaccarin
Department of Economics, Business, Mathematics and Statistics "B. de Finetti", University of Trieste, Trieste, Italy
e-mail: susanna.zaccarin@deams.units.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_14

1 Introduction

The recent interest in the analysis of collaboration networks lies in the fact that long-term trends in scientific work as well as scientist's productivity might closely depend on the topological features of disciplinary networks. Mainly thanks to the availability of international bibliographic archives, several seminal studies in various fields focused on the co-authorship relation as a proxy of scholars' collaborative skills (e.g., [1] for Physics, Mathematics and Neurosciences).

To the best of our knowledge, only few studies have been specifically devoted to the Statistics field. Reference [2] explored the properties of the network generated by the editorial policies of the journals classified as "Statistics and Probability" in the Journal of Citation Report by ISI-Thomson. Reference [7] analysed the co-authorship networks of the 792 Italian statisticians — as recorded in the MIUR database in March 2010 — derived from three bibliographic archives: WoS, CIS, and bibliographic information retrieved from the database of nationally funded research projects, PRIN. The authors discovered distinct collaboration patterns among statisticians as well as distinct effects of scientist network positions on scientific performance, by both Statistics subfield and data source. These results were in line with the findings of [6] on the publication style of Italian statisticians in which they recognized that the use of a single data source can led to biased and partial results.

In this study, we aimed at merging bibliographic data of the three archives exploited in [7] to obtain a complete unified archive, containing both top-international as well as nationally oriented scientific production, as a new basis for network analysis.

To obtain a single co-authorship network, we first combined information from heterogeneous sources by identifying and linking duplicate records (record linkage). The record linkage of metadata in Digital Libraries (DLs) is a very sensitive issue. It refers to "the task of identifying records from disparate data sources that refer to the same entity" [9, p. 245], often used to define integrated information systems in statistical setting [14]. In recent years, computer-oriented record linkages methods are reported in the literature [4,8], which ensure a high efficiency and scalability on large data sets.

The remaining of the paper is organised as follows. Section 2 reports co-authorship network definition and the main network results of Italian academic statisticians as retrieved from heterogeneous data sources. Section 3 describes the merging procedure and Sect. 4 presents the network results for the unified archive. Finally, Sect. 5 concludes with a discussion of future lines of research.

2 Co-authorship Network Definition

Let $\mathscr{N} = \{1, 2, \ldots, n\}$ be the set of n authors and $\mathscr{P} = \{1, 2, \ldots, p\}$ be the set of the p publications observed on the n authors. A co-authorship network is derived from the matrix product $\mathbf{Y} = \mathbf{A}\mathbf{A}'$, where $\mathbf{A}$ is a $n \times p$ affiliation matrix, with elements

Table 1 Number of publications and author coverage rate in the three bibliographic archives

	Years	# of publications	Author coverage rate (%)
WoS	1989–2010	2289	60.7
CIS	1975–2010	3459	73.4
PRIN projects*	*2000–2008**	5054	70.2

*Years of the project

$a_{ik} = 1$ if $i \in \mathscr{N}$ authored the publication $k \in \mathscr{P}$, 0 otherwise. The matrix $\mathbf{Y}$ is the undirected and valued $n \times n$ adjacency matrix with element y_{ij} greater than 0 if $i, j \in \mathscr{N}$ co-authored one or more publications in $\mathscr{P}$, 0 otherwise. The binary version of Y, setting all entries in the valued adjacency matrix greater than zero to 1, was used in the analysis.

As discussed in [7], the specific features of each data source (WoS, CIS, PRIN) used to obtain bibliographic data of Italian statisticians affected the retrieved number of publications and the percentage of statisticians found in a data source out of 792 (author coverage rate, Table 1), as well as the resulting co-authorship patterns.

Summarizing, WoS appeared as the data source in which the average number of co-authors for each statistician was extremely high, being affected by the presence of few statisticians with a large number of co-authors. Patterns consistent with well-established network structures were found out in CIS database. CIS captured internationalization openness by research topics and publication style, while WoS mainly captured the tendency towards an interdisciplinary behavior. Finally, PRIN combined some of CIS and WoS characteristics, although referred only to the selected publications by project's managers and members [7, p. 380].

3 Record Linkage Procedure

To get a unique data set, information from heterogeneous sources were combined by identifying and linking duplicate records. Given the relatively small number of records in the three data sources under analysis, we opted for a semi-automatic method for record linkage because of the presence of errors and omissions in the original datasets (e.g., misspellings in the names of authors and titles, discrepancies in the name of the venue, lack or inaccuracy in the year of publication), especially in PRIN.

To perform the linkage, we proceeded with the commonly used approach of matching the sources in pairs and then performing a reconciliation of possible discrepancies [16]. In particular, we used the following distance functions on each of the key field:

- **Co-authors**: *Jaccard* distance [5] between the set of author surnames of the two records (d_A).

Table 2 Number and % of publications in the unified archive after record linkage by source (np = not present, p = present)

WoS	CIS	PRIN	# pubs	%
np	np	p	3816	43.7
np	p	np	2147	24.6
np	p	p	483	5.5
p	np	np	1139	13.0
p	p	p	321	3.7
p	p	np	395	4.5
p	p	p	434	5.0

- **Title**: error rate measure derived from the edit distance between the two compared strings t_1 and t_2. In particular, we defined the distance as

$$d_T = Ld(t_1, t_2)/max(|t_1|, |t_2|)$$

  where the numerator is the *Levenshtein* distance [13] between t_1 and t_2 and the denominator is the maximum length of the two compared titles.
- **Year**: absolute value of the difference between the years of publication (d_Y).

All strings were lower cased before any comparison. The overall distance was defined as a 3-tuple (d_A, d_T, d_Y), where each element was the distance calculated as described above on the three key fields. We automatically linked the pairs whose distances were below the following thresholds: $d_T < 10\,\%$, $d_A = 0$, and $d_Y = 0$. The couples having $d_T < 20\,\%$, and $d_A \leq 1$ (except those already automatically linked) were manually inspected to establish whether to link them. The choice of the threshold values appeared reasonable enough to avoid an hard manual checking.

The resulting unified archive contains 8735 publications, and its composition by source is shown in Table 2. The overlapping publications retrieved in all the three data sources were quite small. They represented only 5.0 % in the combined archive. Very similar percentages were found by couples of databases. More than 40 % of publications were retrieved only in PRIN (in which only 8.1 % out of 5890 papers were published before 1990), followed by 24.6 % of publications from CIS and 13.0 % from WOS. These results confirmed the high heterogeneity of scientific production of Italian statisticians (top international papers as well as nationally oriented scientific production) that was reflected by the specific features of the three databases.

4 Network Results

In the unified database, the author coverage rate for all statisticians was 85.5 %. Different values were obtained from the five Statistics subfields. Statistics, Statistics for Experimental and Technological research, and Social Statistics were well represented (90.1, 86.7, and 90.5 %, respectively), whereas the lowest authors coverage

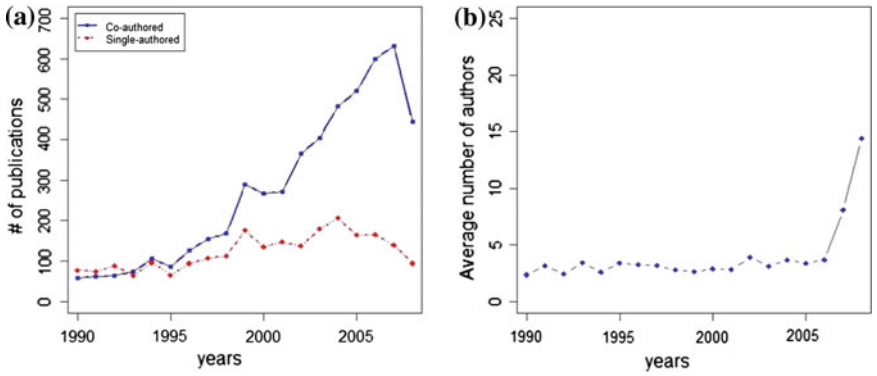


Fig. 1 **a** Co-authored and single-authored publications; **b** average number of authors

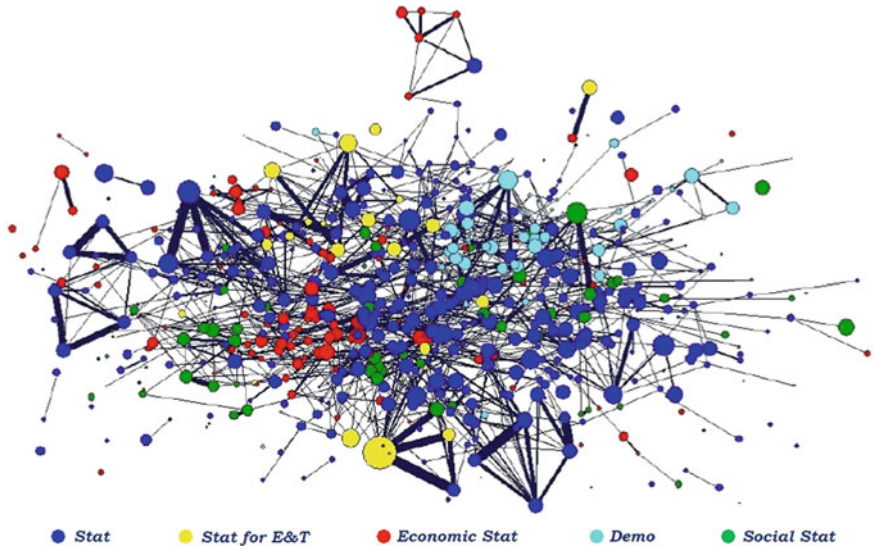


Fig. 2 Co-authorship network for statisticians. Statistics subfields: Statistics (*Stat*), Statistics for Experimental and Technological research (*Stat for E&T*), Economic Statistics (*Economic Stat*), Demography (*Demo*), and Social Statistics (*Social Stat*). *Node size*: # publications per author. *Edge size*: # pubs shared by pairs of authors

rates were observed for Economic Statistics (78.1 %), and Demography (70.6 %). The percentage of co-authored publications was 66.7 %.

In Fig. 1a, we considered the number of co-authored and single-authored publications in the period 1990–2008. The increasing of co-authorship behaviour for Italian statisticians was observed since the end of 1990 (with an exception for the year 2008 mainly due to update issues in the original archives). The average number of authors per publication was around 4 and considering only statisticians was around

Table 3 Network statistics, small-world and scale-free topology assessment for *All authors* and Statisticians

	All authors	Statisticians
#. of authors	7332	677
#. of authors per pub	4.31	2.53
St.Dev.	*29.38*	*2.47*
#. of pub per author	5.14	16.48
St.Dev.	*8.98*	*15.49*
#. of edges	474478	1197
Density	0.018	0.005
Average degree	129.43	3.54
Giant component (%)	97.64	81.24
Average Path Length (ℓ)	5.29	5.46
Clustering Coefficient (Γ)	0.85	0.39
E-I index	0.58	−0.43
	All authors	Statisticians
Small world		
$\ell(G)/\ell(ER)$	2.54	1.05
$\Gamma(G)/\Gamma(ER)$	49.82	49.33
Scale free[a]		
C	0.27	0.45
$\hat{\alpha}$	1.32	1.62

[a] Significant parameter at: $*p < .1$, $**p < .05$, $***p < .01$

3 (Fig. 1b and Table 3). It is worthy to note the high value of the average number of authors per publication for the past two years due to the presence of six and ten publications, respectively in 2007 and in 2008, with a number of authors greater than 100. More specifically, we found a total of 33 publications with this characteristic in WoS referred to nine statisticians. We decided to keep them in the analysis given their relevance in network studies for highlighting peculiar collaboration styles (e.g., interdisciplinary behavior and/or preferential attachment mechanisms [1]).

Finally, the average number of publications per author was around 5, but disregarding the external authors was 16 publications per each statistician (Table 3).

Taking into account both all authors and only statisticians, two adjacency data matrices were considered from the affiliation matrix retrieved from the unified database. The visualization of the co-authorship network for statisticians is shown in Fig. 2. Some network descriptive statistics [19] are reported in Table 3. The overall network cohesion measured by the number of observed ties among authors compared to the maximum possible number of ties in the network, given by $n(n-1)/2$, (*density*) was very low for both networks. The mean value of ties per author (*average degree*) was particularly high when we considered all authors (129.4) with respect to the

internal network defined considering only the co-authorship relations among statisticians (3.5). As noted before, the high value for all authors is due to the presence of those publications in WoS with more than 100 authors. Most of the authors are included in one large connected component (*giant component*) absorbing a considerable percentage of them (97.64, and 81.24 % for only statisticians). The extent of collaboration closure of all authors and statisticians was evaluated through the *E-I* index –a measure of the group embeddedness that compares the number of ties within groups (*Internal*) and between groups (*External*) [12]. The indexes highlighted both a high interdisciplinary behavior with outsider authors (*E-I* index $= .58$) and high internal collaboration among statisticians (*E-I* index $= -.43$), especially in Statistics and Demography subfields.

4.1 Network Topology

We assessed the consistency of the observed networks with the main topological structures emerging in co-authorship setting (i.e., small-world and scale-free topology). Small-world configuration [20] is characterized by: *(i)* small dense network regions revealed by high clustering coefficient Γ (average number of closed triangles out of the total number of triplets of actors), and *(ii)* short paths connecting any two authors revealed by low average path length ℓ (average number of ties along the shortest paths for all possible pairs of actors). A "scale free" network [1] implies the existence of a peculiar tie formation mechanism named preferential attachment, i.e., the tendency to collaborate with the best connected authors. If the degree distribution follows a power law, a scale-free network structure emerges.

Results in Table 3 suggested that statisticians appeared clustered into distinct groups—probably driven by subfields and by geographical/institutional affiliation—connected by few short-cuts, resembling a small-world configuration. This network structure allows statistical knowledge to flow easily among actors [15]. The scale-free results revealed no statistical evidence of preferential attachment mechanism in co-authorship style, although the presence of prominent statisticians could not be excluded, as discussed in [7, p. 378]. Figure 3 reports the Complementary Cumulative Function (CCF) of the observed degree distribution and the corresponding fitted power law distribution for all authors (Fig. 3a) and only statisticians (Fig. 3b). The plots show the strong departure of observed degree distribution from the power law distribution confirming the absence of a scale-free configuration.

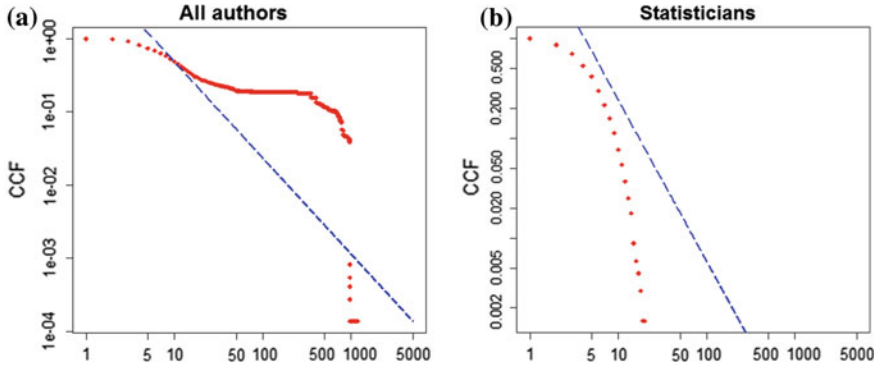


Fig. 3 Obs. Complementary Cumulative Function (CCF) *–dot line–* and fitted Power Law (PL) distributions *–dashed line*. Log–log scale plot: x axis = degree k. y axis = proportion of authors with $degree > k$. **a** all authors; **b** statisticians

5 Conclusion and Further Remarks

The co-authorship patterns of Italian statisticians have been explored by combining their heterogeneous scientific production retrieved in three distinct databases. To construct a unified archive, a record linkage procedure was adopted to correctly identify synonymous author names along with their publications.

A further issue is related to author name disambiguation in order to achieve better quality of network data. Specifically, it "occurs when one author can be correctly referred to by multiple name variations (synonyms) or when multiple authors have exactly the same name or share the same name variation (polysems)" [18, p. 680]. Reference [11, p. 85] pointed out that "One may argue that person name disambiguation inherently includes the problem of personal name matching, since there may exist many namesakes who have a variety of name variants." These two issues are usually treated in the specific literature as independent tasks with personal name matching preceding personal name disambiguation.

A myriad of recent studies were devoted to name disambiguation methods in bibliographic DLs (for a recent survey, see [10]) in computer science, sociological and linguistic setting by covering supervised, unsupervised or semi-supervised techniques. Due to the lack of training data, unsupervised methods [3], and especially the techniques described in [17], seem to be very promising to our case.

Once author disambiguation will be assessed, further analyses will be devoted to identify the characteristics of the emerging groups of statisticians, and to explore the presence of other configurations in co-authorship.

References

1. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**, 47–97 (2002)
2. Baccini, A., Barabesi, L., Marcheselli, M.: How are statistical journals linked? A network analysis. Chance **22**, 35–45 (2009)
3. Carvalho, A.P., Ferreira, A.A., Laender, A.H.F., Goncalves, M.A.: Incremental unsupervised name disambiguation in cleaned digital libraries. J. Inf. Data Manag. **2**, 289–304 (2011)
4. Christen, P.: A survey of indexing techniques for scalable record linkage and deduplication. IEEE Trans. Knowl. Data Eng. **24**, 1537–1555 (2012)
5. Cohen, W., Ravikumar, P., Fienberg, S.: A comparison of string metrics for matching names and records. In: KDD Workshop on Data Cleaning and Object Consolidation, vol. 3, pp. 73–78 (2003)
6. De Battisti, F., Salini, S.: Robust analysis of bibliometric data. Stat. Methods Appl. **22**, 269–283 (2013)
7. De Stefano, D., Fuccella, V., Vitale, M.P., Zaccarin, S.: The use of different data sources in the analysis of co-authorship networks and scientific performance. Soc. Netw. **35**, 370–381 (2013)
8. Domingo-Ferrer, J., Torra, V.: Disclosure risk assessment in statistical microdata protection via advanced record linkage. Stat. Comput. **13**, 343–354 (2003)
9. Durham, E., Xue, Y., Kantarcioglu, M., Malin, B.: Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. Inf. Fusion **13**, 245–259 (2012)
10. Ferreira, A.A., Goncalves, M.A., Laender, A.H.F.: A brief survey of automatic methods for author name disambiguation. SIGMOD Rec. **41**, 15–26 (2012)
11. Kang, I.S., Na, S.H., Lee, S., Jung, H., Kim, P., Sung, W.K., Lee, J.H.: On co-authorship for author disambiguation. Inf. Process. Manag. **45**, 84–97 (2009)
12. Krackhardt, D., Stern, R.N.: Informal networks and organizational crises: an experimental simulation. Soc. Psychol. Q. **51**, 123–140 (1988)
13. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions and reversals. Sov. Phys.-Dokl. **10**, 707–710 (1966)
14. Liseo, B., Montanari, G.E., Torelli, N. (a cura di): Metodi statistici per l'integrazione di dati da fonti diverse. Franco Angeli, Milano (2006)
15. Moody, J.: The structure of a social science: disciplinary cohesion from 1963 to 1999. Am. Sociol. Rev. **69**, 213–238 (2004)
16. Sadinle, M., Hall, R., Fienberg, S.E.: Approaches to multiple record linkage. In: Proceedings of International Statistical Institute, vol. 260 (2011)
17. Strotmann, A., Zhao, D., Bubela, T.: Author name disambiguation for collaboration network analysis and visualization. Proc. Am. Soc. Inf. Sci. Technol. **46**, 1–20 (2009)
18. Veloso, A., Ferreira, A.A., Goncalves, M.A., Laender, A.H.F., Meira Jr., W.: Cost-effective on-demand associative author name disambiguation. Inf. Process. Manag. **48**, 680–697 (2012)
19. Wasserman, S.: Social Network Analysis: Methods and Applications, vol. 8. Cambridge University Press, Cambridge (1994)
20. Watts, D., Strogatz, S.: Collective dynamics of small world networks. Nature **393**, 440–442 (1998)

Statistical Issues in Bayesian Meta-Analysis

Elías Moreno

Abstract

In this paper, we present two problems in meta-analysis. One is the model uncertainty generated by the available heterogenous sampling information. We claim that this model uncertainty has to be incorporated into the meta-inference, and propose a Bayesian clustering procedure for doing that. A second problem is that of choosing the linking distribution that relates the experimental sampling model and the meta-model. We claim that the join distribution for the experimental parameters and the meta-parameter has to be a copula in order to ensure that the Bayesian experimental model and meta-model are coherent. A general copula is proposed. Illustrative examples with real data set are given.

1 Introduction

When a medical treatment is applied to patients with a given disease and samples are collected in k different healthcare centers, the meta-analysis tries to see what can be concluded about the fundamental question that each of the trials sought to address: the efficacy of the treatments [18]. DuMouchel and Waternaux [15] encouraged the use of meta-analysis in medicine and in controlled clinical trials of psychopharmacological agents by asserting that even when the study protocols are similar (dosage, length of treatment, control treatment) there is often considerable variation between studies.

E. Moreno (✉)
Department of Statistics, University of Granada, Granada, Spain
e-mail: emoreno@ugr.es

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_15

This implies that the sampling distribution might change across samples, and the dimension of joint sampling model is then larger than that of the individual model. The point is to know how much the dimension of the joint sampling model increases. We argue that to answer to this question yields recognizing an additional source of uncertainty in the statistical inference, the model uncertainty. In the meta-analysis literature, the heterogeneity analysis mainly consists of either estimating the variance of the linking normal distribution, the so-called heterogeneity parameter, or testing the null hypothesis that the samples are homogeneous *versus* the alternative that they are heterogeneous ([2,6,31] among others). We note that this heterogeneity analysis compares two extreme situations, either all samples come from the same distribution or the distribution of the samples are all different. We claim that this is a unrealistic simplification of the problem. Intermediate situations in which the number of distributions generating the samples is smaller than the original number of samples k are quite feasible a priori, and hence, a clustering analysis of the k samples seems to be the natural method for analyzing the between samples heterogeneity. This is an important point because the likelihood of the meta-parameters strongly depends on the cluster structure of the samples involved.

Therefore, we propose to carry out a meta-analysis for each cluster of the samples, and then pooling them. The final result is a mixture of the meta-analyses conditional on the clusters, where the mixing distribution is the posterior probability of the clusters (model averaging). The number of cluster models, the Bell number $\mathfrak{B}(k)$, is a huge number, even for moderate values of k, and it precludes the computation of all posterior model probabilities. However, the hope is that in practice only a small number of cluster models will have nonnegligible posterior probability values.

We remark that the presence of k heterogeneous samples yields new difficulties related with the consistency of the Bayesian model selection procedure. While for regular models the parameter uncertainty disappears as the sample size grows to infinity, the model uncertainty does not necessarily disappear as the number of samples k grows to infinity. The difficulty comes from the fact that the dimension of the models grows as the number of samples grows, and the Bayes factors for model selection are not necessarily consistent in this context [23,24]. In particular, the well-established BIC procedure [30] is inconsistent when the dimension of the model grows at the same rate of growing than the sample size.

On the other hand, we emphasize that the linking distribution, the distribution of the experimental parameters conditional on the meta-parameter, is of the utmost importance in meta-analysis as it is a necessary distribution for defining the likelihood of the meta-parameter. We remark that in order to ensure that the conditional and marginal distributions of the experimental parameter and the meta-parameter be coherent, some constraints on the linking distribution have to be imposed. A constrain is that the bivariate distribution of the experimental parameter and the meta-parameter be a copula, or equivalently it is belonging to the class of bivariate distributions with given marginals, the Frèchet class. Following [25], a general linking distribution satisfying this requirement is presented.

The rest of the paper is organized as follows. The Bayesian meta-model, the linking distribution, and the meta-inference are presented in Sect. 2. Section 4 presents a

Bayesian clustering procedure for detecting heterogenous samples based on that given by [10]. In Sect. 5 some examples with real data sets are given, and Sect. 6 contains some concluding remarks.

2 The Bayesian Meta-Model

Suppose that a clinical trial is carried in $k \geq 2$ centers that provide heterogenous samples $\{x_i, i = 1, \ldots, k\}$ with distribution $\{f(x_i|\theta_i), i = 1, \ldots, k\}$, where $\theta_i \in \Theta$ represents the effectiveness of the treatment conditional on center i. For simplicity, we assume that Θ is a one-dimensional space. Assuming that the prior information on θ_i is weak the objective Jeffreys' prior $\pi^J(\theta_i)$ is recommended, which might be an improper distribution. Then, we have the objective Bayesian experimental model

$$M_i : \left\{ f(x_i|\theta_i),\ \pi^J(\theta_i) \right\}, \tag{1}$$

for $i = 1, \ldots, k$.

We now introduce a latent variable X, the meta-variable, which is defined as the result we would obtain when the treatment with effectiveness θ is applied to a patient, and there is no between center variability. The conditional distribution $f(x|\theta)$ of this meta–variable X is assumed to belong to the same parametric family than that of the observable variables X_i, where the meta–parameter θ represents the unconditional treatment effectiveness.

Thus, the objective Bayesian meta-model M is given by

$$M : \left\{ f(x|\theta),\ \pi^J(\theta) \right\}. \tag{2}$$

A first quantity of interest in meta-analysis is the posterior distribution of meta-parameter θ, conditional on the observed samples $\{x_i, i = 1, \ldots, k\}$, that is,

$$\pi(\theta|x_1, \ldots, x_k) = \frac{f(x_1, \ldots, x_k|\theta)\pi^J(\theta)}{\int f(x_1, \ldots, x_k|\theta)\pi^J(\theta)d\theta}.$$

In this expression $f(x_1, \ldots, x_k|\theta)$ is the likelihood of θ for the experimental data $(x_1, \ldots, x_k)$, and we remark that for computing this likelihood $f(x_1, \ldots, x_k|\theta)$ a linking distribution $\pi(\theta_i|\theta)$ relating the experimental parameters and the meta-parameter has to considered, a topic that we deal with in the next section.

A second quantity of interest is the meta-predictive distribution of a new patient y, conditional on the available data, which is given by

$$f(y|x_1, \ldots, x_k) = \int f(y|\theta)\pi(\theta|x_1, ..., x_k)d\theta.$$

2.1 The Linking Distribution and the Likelihood of the Meta-Parameter

Assuming that the heterogenous samples $\{x_i, i = 1, \ldots, k\}$ are independent, conditional on θ_i, we have the joint distribution

$$f(x_1, \ldots, x_k|\theta_1, \ldots, \theta_k) = \prod_{i=1}^{k} f(x_i|\theta_i).$$

Assuming that for the linking distribution $\pi(\theta_i|\theta)$ the parameters $\{\theta_i, i = 1, \ldots, k\}$ are independent, conditional on θ, the likelihood of the meta-parameter θ can be written as

$$f(x_1, \ldots, x_k|\theta) = \prod_{i=1}^{k} \int f(x_i|\theta_i)\pi(\theta_i|\theta)d\theta_i. \tag{3}$$

As noted in [25], the linking distribution $\pi(\theta_i|\theta)$ has to be compatible with the marginal priors $\pi^J(\theta_i)$ and $\pi^J(\theta)$. This means that $\pi(\theta_i|\theta)$ has to be chosen such that the bivariate distribution $\pi(\theta_i, \theta) = \pi(\theta_i|\theta)\pi^J(\theta)$ satisfies the integral equations

$$\int_0^1 \pi(\theta_i, \theta)d\theta_i = \pi^J(\theta), \quad \int_0^1 \pi(\theta_i, \theta)d\theta = \pi^J(\theta_i). \tag{4}$$

While the first equation is satisfied for any probability distribution $\pi(\theta_i|\theta)$, the second equation is only satisfied for prior distributions in the so-called Frèchet bidimensional class with Jeffreys marginal. There are very many bidimensional distributions with given marginal, for instance that given by [13,14,17,26,28], among many others, and we can certainly make use of them (for a discussion see [25]).

An interesting general linking distribution is the intrinsic prior class $\{\pi^I(\theta_i|\theta, t), t = 1, 2, ...\}$, which arises from the model comparison of the meta-model M *versus* M_i [4,21,22]. The intrinsic prior $\pi^I(\theta_i|\theta, t)$, where t is the training sample size, is given by

$$\begin{aligned}\pi^I(\theta_i|\theta, t) &= \pi^J(\theta_i)E_{z_1,\ldots,z_t|\theta_i}\frac{f(z_1, \ldots, z_t|\theta)}{\int f(z_1, \ldots, z_t|\theta_i)\pi^J(\theta_i)d\theta_i} \\ &= \pi^J(\theta_i)\int \frac{f(z_1, \ldots, z_t|\theta)}{\int f(z_1, \ldots, z_t|\theta_i)\pi^J(\theta_i)d\theta_i} f(z_1, \ldots, z_t|\theta_i)dz_1...dz_t.\end{aligned} \tag{5}$$

The intrinsic prior $\pi^I(\theta_i|\theta, t)$ is, conditional on θ, a proper prior for any t, and the joint distribution $\pi^I(\theta_i, \theta|t) = \pi^I(\theta_i|\theta, t)\pi^J(\theta)$ satisfies Eq. (4) as shown in the next lemma.

Lemma 1 *The bidimensional distribution $\pi^I(\theta_i, \theta|t) = \pi^I(\theta_i|\theta, t)\pi^J(\theta)$ satisfies Eq. (4).*

Proof For, we first note that for any t we have

$$\int \pi^I(\theta_i|\theta, t)d\theta_i = 1,$$

and hence, the first equation is satisfied. Further, for any t we have

$$\int \pi^I(\theta_i|\theta,t)\pi^J(\theta)d\theta = \pi^J(\theta_i)E_{z_1,\ldots,z_t|\theta_i}\frac{\int f(z_1,\ldots,z_t|\theta)\pi^J(\theta)d\theta}{\int f(z_1,\ldots,z_t|\theta_i)\pi^J(\theta_i)d\theta_i} = \pi^J(\theta_i),$$

and this proves the assertion. □

The training sample size t controls the concentration degree of the probability distribution of θ_i around θ. Next Lemma 2 shows that, under mild conditions, as t tends to infinity the distribution $\pi^I(\theta_i|\theta,t)$ degenerates to a point mass on θ.

Lemma 2 *For any regular sampling models $f(x|\theta)$ we have that $\pi^I(\theta_i|\theta,t)$ degenerates to a point mass on θ as t tends to infinity. Further, if $\lim_{t\to\infty}\pi^I(\theta_i|\theta,t)$ is a probability density, we then have that*

$$\lim_{t\to\infty}\pi^I(\theta_i|\theta,t) = \delta_{\{\theta\}}(\theta_i),$$

where $\delta_{\{\theta\}}(\theta_i)$ represents the Dirac's delta.

Proof We note that $\pi^I(\theta_i|\theta,t)$ can be written as

$$\pi^I(\theta_i|\theta,t) = \pi^J(\theta_i)E_{z_1,\ldots,z_t|\theta_i}B_{01}(z_1,\ldots,z_t),$$

where $B_{01}(z_1,\ldots,z_k)$ is the Bayes factor to compare model

$$M_0 : f(x|\theta), \text{ for fixed } \theta,$$

versus model

$$M_1 : \{f(x|\theta_i), \pi^J(\theta_i)\},$$

for the sample $z_1,\ldots,z_t$. Using the consistency properties of the Bayes factor for nested models [11], it follows that the limit in probability when sampling from model M_1 is zero, that is,

$$\lim_{t\to\infty}B_{01}(z_1,\ldots,z_t) = 0, \ [M_1],$$

and hence, $E_{z_1,\ldots,z_t|\theta_i}B_{01}(z_1,\ldots,z_t)$, the expectation of the Bayes factor $B_{01}(z_1,\ldots,z_t)$ with respect to the alternative model $f(z_1,\ldots,z_t|\theta_i)$, goes to zero as $t\to\infty$. Thus, the distribution $\pi^I(\theta_i|\theta,t)$ degenerates to zero when $t\to\infty$ and $\theta_i \neq \theta$.

Further, when $\lim_{t\to\infty}\pi^I(\theta_i|\theta,t)$ is a probability density, it follows that

$$\lim_{t\to\infty}\pi^I(\theta_i|\theta,t) = \begin{cases} 0 \text{ for } \theta_i \neq \theta, \\ 1 \text{ for } \theta_i = \theta, \end{cases}$$

and this completes the proof of Lemma 2. □

2.2 Estimating the Meta-Parameter θ

For the intrinsic linking distribution (5), the likelihood of the meta-parameter θ for the samples $\{x_1, \ldots, x_k\}$ becomes

$$f(x_1, \ldots, x_k|\theta, t) = \prod_{i=1}^{k} \int f(x_i|\theta_i)\pi^I(\theta_i|\theta, t)d\theta_i. \tag{6}$$

Using this likelihood of the meta-parameter θ and the Jeffreys prior $\pi^J(\theta)$, the posterior distribution of θ, conditional on t, is given by

$$\pi(\theta|x_1, \ldots, x_k, t) = \frac{f(x_1, \ldots, x_k|\theta, t)\pi^J(\theta)}{\int f(x_1, \ldots, x_k|\theta, t)\pi^J(\theta)d\theta}. \tag{7}$$

This posterior distribution depends on the training sample size t. The value of t is usually taken as smaller as possible [4], although there is no reason for doing that, and hence, there are several alternative ways for dealing with t. For instance, an estimator of θ could be the posterior expectation

$$E(\theta|x_1, \ldots, x_k, t) = \int \theta\pi(\theta|x_1, \ldots, x_k, t)d\theta$$

and if we let the training sample size t vary in the set of integers, posterior robustness of the posterior expectation with respect to the intrinsic prior class can be assessed [12]. We could also set t equal to the minimum of the actual sizes of the samples $\{x_1, \ldots, x_k\}$ so that the concentration of the linking distribution does not exceed to that of the likelihood [8,9]. We can also integrate out the training sample size t in the linking distribution $\pi^I(\theta_i|\theta, t)$ in (5) with respect to a prior $\pi(t)$ to obtain the likelihood of the meta-parameter as

$$f(x_1, \ldots, x_k|\theta) = \prod_{i=1}^{k} \sum_{t=1}^{\infty} \pi(t) \int f(x_i|\theta_i)\pi^I(\theta_i|\theta, t)d\theta_i.$$

For a derivation of $\pi(t)$ see [25].

2.3 Testing the Equality of Treatments Effectiveness

Treatment effectiveness comparison based on multiple studies is among the most important chapter in meta–analysis. Most of the Bayesian meta–analyses for comparing two treatments estimate either the difference of the meta–effectiveness parameters or their odds ratio (see, for instance, [7,19,29]; among many others). Typically, the 95 % HPD region of the posterior distribution of the difference of the meta–parameters is computed, and the equality of meta–effectiveness of the treatments is accepted if the "singular" point 0 is contained in the region, and rejected otherwise. When the odds ratio is used instead, the equality is accepted if the "singular" point 1 is contained in the 95 % HPD of the posterior odds distribution.

We note that by doing so the same statistical evidence in favor of the null is obtained whatever is the position of the "singular" point in the HPD region. We also note that this Bayesian procedure mimics the frequentist testing methodology that uses confidence intervals, and it dates back to [20].

The difficulty with this methodology is that the null hypothesis does not play any role in the construction of the HPD region, and hence, the prior uncertainty on the null and the alternative models are not taken into account.

It was [16] who strongly recommended to separate testing problems from estimation problems. He advocated using the Bayesian testing methodology based on Bayes factors. Sensible motivations for using model selection for testing problems were also given by [5] and references there in. We here provide a Bayesian meta-test that follows the standard model selection methodology.

Let us consider, two treatments T_j, $j = 1, 2$, and assume that for each treatment there are available k_j independent samples $\mathbf{x}_j = \{x_{ji}, i = 1, \ldots, k_j\}$ on their effectiveness. The likelihood $f(\mathbf{x}_j|\xi_j)$ of the meta-parameter ξ_j of treatment T_j is given by

$$f(\mathbf{x}_j|\xi_j, t) = \prod_{i=1}^{k_j} \int f(x_{ji}|\theta_{ji})\pi(\theta_{ji}|\xi_j, t)d\theta_i,\ j = 1, 2, \tag{8}$$

where $\pi(\theta_{ji}|\xi_j)$ is the linking distribution.

The interest now is on testing the null hypothesis $H_0 : \xi_1 = \xi_2$ *versus* the alternative unrestricted hypothesis $H_1 : (\xi_1, \xi_2) \in (0, 1)^2$. A simpler objective Bayesian solution of this problem is the model comparison between model

$$M_0 : \{f(\mathbf{x}_1|\xi, t)\Pr(\mathbf{x}_2|\xi, t),\ \pi^J(\xi)\},$$

and model

$$M_1 : \{f(\mathbf{x}_1|\xi_1, t)f(\mathbf{x}_2|\xi_2, t),\ \pi^J(\xi_1)\pi^J(\xi_2)\},$$

where $\pi^J(\xi_j)$ is the Jeffreys prior for the model $f(x_j|\xi_j)$.

Assuming a uniform model prior $\Pr(M_0) = \Pr(M_1) = 1/2$, the posterior probability of the null M_0 is given by

$$\Pr(M_0|\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{1 + B_{10}(\mathbf{x}_1, \mathbf{x}_2)}, \tag{9}$$

where the Bayes factor $B_{10}(\mathbf{x}_1, \mathbf{x}_2, t)$ is

$$B_{10}(\mathbf{x}_1, \mathbf{x}_2, t) = \frac{\prod_{j=1}^2 \int f(\mathbf{x}_j|\xi_j, t)\pi^I(\xi_j)d\xi_j}{\int \left(\prod_{j=1}^2 f(\mathbf{x}_j|\xi, t)\pi^J(\xi)\right) d\xi}. \tag{10}$$

The decision rule is that of choosing model M_0 if $\Pr(M_0|\mathbf{x}_1, \mathbf{x}_2) \geq 1/2$, and M_1 otherwise.

3 Accounting for the Between Sample Heterogeneity: Clustering

The likelihood $f(x_1, \ldots, x_k|\theta) = \prod_{i=1}^{k} \int f(x_i|\theta_i)\pi(\theta_i|\theta)d\theta_i$ considered so far is based on the assumption that there are k different clusters in the samples $\{x_i, i = 1, \ldots, k\}$. However, alternative clusters are a priori possible, and hence, we consider the problem of clustering the k samples. The quantity of interest is now the posterior distribution of the cluster models. This distribution will be the mixing distribution for averaging the meta-inference conditional on the cluster models.

In the next section, we briefly describe the method we use for clustering the k samples.

3.1 The Cluster Models

For clustering the binomial experimental samples $\{x_i, i = 1, \ldots, k\}$ we follow the product partition model approach introduced by [3] and further studied in [10].

Let us first introduce some notation. For an integer p, $1 \leq p \leq k$, we denote a partition of the samples $\{x_i, i = 1, \ldots, k\}$ into p clusters by a vector $\mathbf{r}_p = (r_1, \ldots, r_k)$, where r_i is an integer between 1 and p denoting the cluster to which x_i is assigned. By $\mathfrak{R}_p$ we denote the set of partitions of the samples into p clusters, the number of which is the Stirling number of the second kind $S(k, p)$. Hence, the set of all possible partitions of the samples is $\mathfrak{R} = \cup_{p=1}^{k} \mathfrak{R}_p$, the number of which is given by the Bell number $\mathfrak{B}_k = \sum_{p=1}^{k} S(k, p)$.

Given a partition $\mathbf{r}_p = (r_1, \ldots, r_k)$, the sampling distribution of $\mathbf{x}$ is the product partition model

$$f(\mathbf{x}|p, \mathbf{r}_p, \theta_p, k) = \prod_{j=1}^{p} \prod_{i:r_i=j} f(x_i|\theta_i) = \prod_{i=1}^{k} f(x_i|\theta_{r_i}), \tag{11}$$

where $\theta_p = (\theta_{r_1}, \ldots, \theta_{r_k})$ is an unknown parameter in the space Θ^p, and the component θ_{r_i} indicates the distribution of the sample x_i. We remark that the partition $\mathbf{r}_p$ defines the sampling model, so that partition and model are equivalent words in this context.

The partition $\mathbf{r}_p = (1, .., 1)$ corresponds to the singular case where the samples are grouped in only one cluster. Its corresponding likelihood function is given by

$$f(\mathbf{x}|1, \mathbf{r}_1, \theta, k) = \prod_{i=1}^{k} f(x_i|\theta_i).$$

We note that under this model, there is no variation between samples, and hence, the meta-parameter and the experimental parameters coincide, that is, $\theta_1 = \cdots = \theta_k = \theta$. Under this model there is no sampling information for formulating the meta-analysis.

3.2 Prior Distributions

To complete the specification of the Bayesian cluster model, we need a prior distribution for the models $(p, \mathbf{r}_p)$ and for the model parameter parameters θ_p, conditional on k. A natural decomposition of the joint prior distribution $\pi(p, \mathbf{r}_p, \theta_p|k)$ is

$$\pi(p, \mathbf{r}_p, \theta_p|k) = \pi(\theta_p|p, \mathbf{r}_p, k)\pi(\mathbf{r}_p|p, k)\pi(p|k).$$

Let us specify the three priors $\pi(\theta_p|p, \mathbf{r}_p, k)$, $\pi(\mathbf{r}_p|p, k)$, and $\pi(p|k)$.

1. The prior for θ_p, conditional on the partition $\mathbf{r}_p$, is assumed to be the intrinsic prior arising from the model comparison between the one cluster model againts the cluster model $(p, \mathbf{r}_p)$ [4,22]. This prior turns out to be

$$\pi^I(\theta_p|\theta, p, \mathbf{r}_p, k) = \prod_{j=1}^{p} \pi^I(\theta_j|\theta).$$

2. To assign $\pi(\mathbf{r}_p|p, k)$ we decompose the class of partitions $\mathfrak{R}_p$ as follows. Let k_i be the number of samples assigned to the *ith* cluster, $i = 1, \ldots, p$. Then the class $\mathfrak{R}_p$ as be expressed as

$$\mathfrak{R}_p = \cup_{\substack{1\le k_1\le\cdots\le k_p \\ k_1+\cdots+k_p=k}} \mathfrak{R}_{p;k_1,\ldots,k_p}.$$

The subclass $\mathfrak{R}_{p;k_1,\ldots,k_p}$ is called a configuration class, that is, a class of partitions in $\mathfrak{R}_p$ having the same configuration $(k_1, \ldots, k_p)$. Using this decomposition of $\mathfrak{R}_p$ we decompose the prior $\pi(\mathbf{r}_p|p, k)$ as

$$\pi(\mathbf{r}_p|p, k) = \pi(\mathbf{r}_p|\mathfrak{R}_{p;k_1,\ldots,k_p}, k)\pi(\mathfrak{R}_{p;k_1,\ldots,k_p}|p, k).$$

Since the labels of the clusters are irrelevant, the number of partitions in $\mathfrak{R}_{p;k_1,\ldots,k_p}$ can be written as

$$\binom{k}{k_1 \cdots k_p} \frac{1}{R(k_1, \ldots, k_p)},$$

where $\binom{k}{k_1\cdots k_p}$ is the multinomial coefficient, and $R(k_1, \ldots, k_p) = \prod_{i=1}^{k} \left(\sum_{j=1}^{p} \mathbf{1}_{(k_j=i)}\right)!$ corrects the count by considering the redundant strings corresponding to the vector $(k_1, \ldots, k_p)$. For instance, for the vector $(k_1, \ldots, k_p)$ such that $k_1 = \cdots = k_{p-4} < k_{p-3} = k_{p-2} < k_{p-1} = k_p$, we have that $R(k_1, \ldots, k_p) = (p-4)!2!2!$. Thus, the number of partition in $\mathfrak{R}_p$ is given by the Stirling number of second kind, which we write as

$$S(k, p) = \sum_{\substack{1\le k_1\le\cdots\le k_p \\ k_1+\cdots+k_p=k}} \binom{k}{k_1 \cdots k_p} \frac{1}{R(k_1, \ldots, k_p)}.$$

Since that the partitions $\mathbf{r}_p$ in $\mathfrak{R}_{p;k_1,\ldots,k_p}$ are exchangeable, it seems reasonable to assign a uniform prior to them, that is,

$$\pi(\mathbf{r}_p|\mathfrak{R}_{p;k_1,\ldots,k_p}, k) = \binom{k}{k_1 \cdots k_p}^{-1} R(k_1, \ldots, k_p).$$

3. Further, since the configuration classes $\{\Re_{p;k_1,\ldots,k_p},\ 1 \le k_1 \le \cdots \le k_p,\ k_1 + \cdots + k_p = k\}$ in $\Re_p$ contain models of the same complexity, it seems reasonable to assign to these classes a uniform prior. For doing that we need to count the number of configuration classes in $\Re_p$. We note that this number is also the number of ways the integer k can be partitioned into p integer parts, which we denote by $b(k, p)$. This number does not seem to have a closed form expression as a function of p and k. However, it can be shown that $b(k, p)$ satisfies the recursive equation

$$b(k, p) = b(k-1, p-1) + b(k-p, p),\ 1 \le p \le k,$$

with

$$b(k, 1) = b(k, k) = 1.$$

Therefore,

$$\pi(\Re_{p;k_1,\ldots,k_p}|p, k) = \frac{1}{b(k, p)},\ \Re_{p;k_1,\ldots,k_p} \in \Re_p.$$

4. In the meta-analysis scenario there is no reason to penalize a priori large number of clusters, so that the prior on the number of clusters $\pi(p|k)$ is assumed to be the uniform distribution

$$\pi(p|k) = \frac{1}{k},\ p = 1, \ldots, k.$$

Thus, we finally have the prior

$$\pi(p, \mathbf{r}_p|k) = \binom{k}{k_1 \cdots k_p}^{-1} R(k_1, \ldots, k_p) \frac{1}{b(k, p)} \frac{1}{k}, \mathbf{r}_p \in \Re_p. \tag{12}$$

Sampling properties of this Bayesian procedure has been explored in [10].

3.3 Posterior Distribution of the Cluster Models

Using the likelihood (11) and prior (12), the posterior probability of an arbitrary model $M_{\mathbf{r}_p}$ is given by

$$\pi(p, \mathbf{r}_p|\mathbf{x}, k) = \frac{m(\mathbf{x}|p, \mathbf{r}_p, k)\pi(p, \mathbf{r}_p|k)}{\sum_{p=1}^{k} \sum_{\mathbf{r}_p \in \Re_p} m(\mathbf{x}|p, \mathbf{r}_p, k)\pi(p, \mathbf{r}_p|k)},\ M_{\mathbf{r}_p} \in \Re, \tag{13}$$

where $m(\mathbf{x}|p, \mathbf{r}_p, k)$, the marginal of the data, conditional on model $M_{\mathbf{r}_p}$, is given by

$$m(\mathbf{x}|p, \mathbf{r}_p, k) = \int_0^1 \cdots \int_0^1 f(\mathbf{x}|p, \mathbf{r}_p, \theta_p, k)\pi(\theta_p|p, \mathbf{r}_p, k)d\theta_p.$$

Each model $M_{\mathbf{r}_p} \in \Re$ indicates a different heterogeneity structure of the samples $\mathbf{x} = \{x_i, i = 1, \ldots, k\}$, and the probability vector $\{\pi(M_{\mathbf{r}_p}|\mathbf{x}),\ M_{\mathbf{r}_p} \in \Re\}$ gives us a measure of the uncertainty on these structures. Hence, the starting point for meta-analysis are the models $M_{\mathbf{r}_p} \in \Re$.

4 Examples

The following Example 1 considers multicenter binomial samples to test whether the effectiveness of new drug-eluting stents (DES) is equal to that of the bare-metal stents (BMS), and shows that the posterior inference on the meta-parameter heavily depends on how the samples are clustered. Therefore, the usual assumption that the clinical trials are all heterogenuous might lead to serious misleading meta-inference, and, consequently, it illustrates the need for a cluster analysis previous to the meta-analysis.

Assuming conditional independence between patients in trial i, the natural model is the binomial $Bin(x_i|n_i, \theta_i)$, where the probability of major myocardial infarction θ_i is assigned a uniform prior $\pi(\theta_i) = 1_{(0,1)}(\theta_i)$. The linking distribution (5) for the binomial model turns out to be

$$\pi^I(\theta_i|\theta, t) = (1-\theta)^t(1-\theta_i)^t(t+1)_2F_1\left[-t, -t, 1, \frac{\theta\theta_i}{(\theta_i-1)(\theta-1)}\right], \quad (14)$$

where ${}_2F_1$ denotes the hypergeometric function. The likelihood function of the meta-parameter θ for the data $\mathbf{x} = \{(x_i, n_i), i = 1, \ldots, 11\}$ becomes

$$f(\mathbf{r}_p|p, \theta, t) = \prod_{i=1}^{p}\left\{(t+1)(1-\theta)^t Beta(1+n_{ip}+t-x_{ip}, 1+x_{ip}) \right.$$
$$\left. {}_3F_2\left(a, b, \frac{\theta}{\theta-1}\right)\right\},$$

where ${}_3F_2$ denotes the generalized hypergeometric function with $a = (-t, -t, 1 + x_i)$, $b = (1, -n_{ip} - t + x_{ip})$, and (n_{ip}, x_{ip}) denotes the number of patients and number of major myocardial infarction grouped in p clusters according to the partition $\mathbf{r}_p$.

Following [25] the prior distribution on the hyperparameter t is

$$\pi(t) = \frac{3}{(t+2)(t+3)}, \; t = 1, 2, \ldots$$

so that the likelihood of the meta-parameter θ for the partition $\mathbf{r}_p$ is

$$f(\mathbf{r}_p|p, \theta) = \sum_{t=1}^{\infty} f(\mathbf{r}_p|p, \theta, t)\pi(t). \quad (15)$$

Thus, the posterior expectation of the meta-parameter θ is

$$E(\theta|\mathbf{x}) = \sum_{p=1}^{k}\pi(p)\sum_{\mathbf{r}_p\in\mathfrak{R}_p} E(\theta|p, \mathbf{r}_p)\pi(\mathbf{r}_p|p), \quad (16)$$

where

$$E(\theta|p, \mathbf{r}_p, k) = \frac{\int_0^1 \theta f(\mathbf{r}_p|p, \theta)d\theta}{\int_0^1 f(\mathbf{r}_p|p, \theta)d\theta} \quad (17)$$

Table 1 Posterior rate of major myocardial infarction for DES and for BMS, conditional on some cluster models

Cluster $\mathbf{r}_p$	Post. rate for DES	Post. rate for BMS
$\mathbf{r}_{11} =$ (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)	0.010	0.011
$\mathbf{r}_5 =$ (2, 3, 4, 1, 5, 2, 2, 3, 4, 4, 5)	0.021	0.019
$\mathbf{r}_3 =$ (2, 2, 2, 1, 3, 2, 2, 2, 3, 3, 3)	0.054	0.048
$\mathbf{r}_2 =$ (2, 2, 1, 2, 1, 1, 1, 1, 2, 2, 2)	0.103	0.106
$\mathbf{r}_1 =$ (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)	0.027	0.030

Example 1 Let us consider, the data from 11 randomized clinical trials to compare the effectiveness of new DES versus BMS given in Fig. 3 in [1]. The vectors $\boldsymbol{n}_1 = (120, 533, 50, 175, 31, 260, 662, 117, 152, 517, 24)$ and $\mathbf{x}_1 = (4, 15, 1, 8, 0, 8, 23, 3, 2, 7, 0)$ are the number of patients and number of major myocardial infarction, respectively, for the clinical trials for DES, and $\boldsymbol{n}_2 = (118, 525, 50, 177, 30, 263, 652, 58, 38, 512, 26)$ and $\mathbf{x}_2 = (5, 17, 2, 4, 0, 14, 24, 1, 0, 5, 0)$ for BMS.

In Table 1 we show the posterior rate of major myocardial infarction computed using formula (17), conditional on some clusters $\mathbf{r}_p$ of the samples for DES and BMS.

Table 1 shows that the posterior rate of myocardial infarction for DES and for BMS dramatically vary across clusters. For instance, the posterior rate of major myocardial infarction for DES conditional on the eleven cluster model $\mathbf{r}_{11}$ (all the samples are heterogenous) is 0.01, while the rate conditional on the two cluster model $\mathbf{r}_2 = (2, 2, 1, 2, 1, 1, 1, 1, 2, 2, 2)$ (that corresponds to grouping in one cluster the trials $\{3, 5, 6, 7, 8\}$ and in another cluster the trials $\{1, 2, 4, 9, 10, 11\}$), is as large as 0.103 which is 10 times the former rate. Something similar can be said about the rate of major myocardial infarction for BMS.

The following example analyzes a multicenter clinical trial carried out in the seventies to assess the efficacy of a given daily dose of aspirin to reduce the mortality rate in postmyocardial infarction patients. The clinical trials involve six center in Europe and United States with a total of 10,816 patients. The distribution of the mortality data for each of the trials is assumed to be binomial with θ_i the unknown probability of mortality, conditional on trial i.

The linking distribution is given in (16) and the expression of the likelihood for the meta-parameter is given in (17).

Table 2 Aspirin *versus* placebo trials. n_i is the number of patients in trial i, and x_i the number of deaths

Trial	Aspirin		Placebo	
	n_i	x_i	n_i	x_i
UK–1	615	49	624	67
CDPA	758	44	771	64
GAMS	317	27	309	32
UK–2	832	102	850	126
PRIS	810	85	406	52
AMIS	2267	246	2257	219

Table 3 Top cluster models, their posterior mortality rate under aspirin (top panel), and under placebo (bottom panel)

Aspirin		
Clusters	Post. prob.	Post. expectation
{1, 2, 3}, {4, 5, 6}	0.54	0.15
{1, 2}, {3, 4, 5, 6}	0.23	0.15
{2}, {1, 3, 4, 5, 6}	0.18	0.14
{1, 3}, {2}, {4, 5, 6}	0.05	0.09
Placebo		
Clusters	Post. prob.	Post. expectation
{1, 2, 3, 6}, {4, 5}	0.57	0.17
{1, 2, 3, 5, 6}, {4}	0.39	0.18
{1, 3, 4, 5}, {2, 6}	0.04	0.12

Example 2 Table 2 reports all causes of mortality for six major randomized multicenter clinical trials of aspirin and placebo during the period 1970–79 in postmyocardial infarction patients [6]. The centers were first United Kingdom trial (UK–1), Coronary Drug Project Aspirin trial (CDPA), German–Austrian Multicenter Study (GAMS), second United Kingdom trial (UK–2), Persantine-Aspirin Reinfarction study (PRIS), and Aspirin Myocardial Infarction Study (AMIS). The question to be answered is whether the rate of mortality is significantly reduced by the use of aspirin in postmyocardial infractions patients.

Morris and Normand [27] discussed the meta–analysis of this data using a hierarchical normal model for the logit transformation of the data, and suggested investigating how the assessment of the effectiveness of aspirin is altered by the use of long-tailed prior distribution for the reparametrized logit effects. In Carlin [7] this suggestion were developed using $t-$distributions for the logit; the test consists of computing the 95 % HPD region of posterior distribution of the difference of the meta–effectiveness of the aspirin and placebo. He concluded that this difference

does not significantly differ from zero so that the rate of mortality is not reduced significantly by the use of aspirin (however, he recommended that a more complete analysis of the data might employ beta distributions on the original survival proportions in each group). No heterogeneity test were considered; the heterogeneity of the six data sets for aspirin and for placebo were taken for granted.

Reference [6] studied the heterogeneity of the six trials asserting that for the five clinical trial listed -all but AMIS- "The test for homogeneity of logodds ratio gives $\chi^2_H = 0.63\,(p - value = 0.96)$, which confirms the high consistency among the five odds ratios noted visually. Adding the sixth trial, AMIS, changes the picture dramatically and cancels out a substantial proportion of the apparent beneficial effect." He reported a $p-$value of 0.08, and concludes that it is "a borderline suggestion of heterogeneity results among the six trials".

However, our clustering analysis exhibit a quite different picture. Our analysis for the aspirin data shows that the posterior probability of the two clusters model defined by the partition $\mathbf{r}_p = (1, 1, 1, 1, 1, 2)$, that is, the first five trials are in a cluster and the sixth in a second cluster, is negligible. Table 3 displays the top cluster models for aspirin and placebo according to their posterior probabilities. The last column displays the posterior expectation of the meta-parameter, conditional on the clusters.

Table 3 shows that, after clustering, the number of samples under aspirin has been reduced from six four, and from six to three for placebo. Using this clustering structure, it follows that the unconditional posterior expected mortality rate under aspirin turns out to be 0.14, and under placebo 0.17. We remark that the uncertainty on these estimates is quite large. For instance, the 95 % HPD region of the meta-parameter under aspirin is (0, 0.42), and under placebo (0, 0.46).

Using the cluster models in Table 3 we test the null hypothesis that the mortality rate under aspirin and placebo are equal. The posterior probability of the equality turns out to be 0.75, which leads accepting the equality of the effects with only substantial empirical evidence, according to the Jeffreys's evidence scale. We remark that assuming that all the trials are heterogenous the posterior probability that the mortatiliy rate under aspirin and placebo are equal turns out to be 0.85.

5 Asymptotic

We point out that the asymptotic in meta-analysis is a delicate subject since the dimension of the models involved might grow as the number of samples grow, and this might yield inconsistency [23]. We illustrate these difficulties in the next theorem on a very particular and simple nested model comparison for Bernoulli samples.

Theorem 1 *To compare the null model*

$$M_0 : f((x_1, \ldots, x_k)|\theta_0) = \prod_{i=1}^{k} \theta_0^{x_i} (1 - \theta_0)^{1-x_i},\ x_i = 0, 1,$$

against the alternative model

$$M_1 : \left\{ f((x_1, \ldots, x_k)|\theta) = \prod_{i=1}^{k} \theta_i^{x_i} (1-\theta_i)^{1-x_i},\ \pi(\theta) = \prod_{i=1}^{k} 1_{(0,1)}(\theta_i) \right\},$$

where θ_0 *is an arbitrary but fixed point, and* $\theta = (\theta_1, \ldots, \theta_k) \in (0,1)^k$*, the Bayes factor*

$$B_{10}(x_1, \ldots, x_k) = \frac{\prod_{i=1}^{k} \int_0^1 \theta_i^{x_i} (1-\theta_i)^{1-x_i} d\theta_i}{\prod_{i=1}^{k} \theta_0^{x_i} (1-\theta_0)^{1-x_i}}$$

is consistent as $k \to \infty$ *when sampling from the null model* $f(\mathbf{x}|\theta_0)$*, but it is inconsistent when sampling from an alternative model* $f(\mathbf{x}|\theta)$ *such that* $\lim_{k\to\infty} \sum_{j=1}^{k} \theta_j/k = \theta \in A(\theta_0)$*, where*

$$A(\theta_0) = \{\theta : 2\theta_0^{\theta}(1-\theta_0)^{1-\theta} \geq 1\}.$$

Proof The marginal of the random variables $(X_1, \ldots, X_k)$ under the alternative model $f(\mathbf{x}|\theta_1, \ldots, \theta_k)$ is given by

$$m_1(x_1, \ldots, x_k) = \prod_{i=1}^{k} \int_0^1 \theta_i^{x_i} (1-\theta_i)^{1-x_i} d\theta_i = \frac{1}{2^k},$$

and hence, the Bayes factor $B_{10}(x_1, \ldots, x_k)$ turns out to be

$$B_{10}(x_1, \ldots, x_k) = \left(\frac{1}{2\ \theta_0^{\bar{x}_k} (1-\theta_0)^{1-\bar{x}_k}} \right)^k,$$

where $\bar{x}_k = \sum_{j=1}^{k} x_j/k$. Since $\lim_{k\to\infty} \sum_{j=1}^{k} X_j/k = \theta_0$, $[P_{\theta_0}]$, and

$$2\ \theta_0^{\theta_0}(1-\theta_0)^{1-\theta_0} \geq 1,$$

we have that

$$\lim_{k\to\infty} B_{10}(x_1, \ldots, x_k) = 0,\ [P_{\theta_0}].$$

Likewise, when sampling from any alternative model P_θ, we have that $\lim_{k\to\infty} \sum_{j=1}^{k} x_j/k = \sum_{j=1}^{k} \theta_j/k = \theta$, $[P_\theta]$. Then, for $\theta \in A(\theta_0)$ it follows that

$$\lim_{k\to\infty} B_{10}(x_1, \ldots, x_k) = 0\,,\ [P_\theta],$$

and this proves the assertion. □

Theorem 1 asserts that the Bayes factor $B_{10}(x_1, \ldots, x_k)$ is consistent under the null model M_0 but there is a nonempty region $A(\theta_0)$ in the alternative parameter space for which the Bayes factor is inconsistent under any model in the region.

Corollary 1 *If we replace the uniform prior in model* M_1 *with an arbitrary unimodal prior* $\pi(\theta_i)$*, still there exists an inconsistency region* $A'(\theta_0)$*.*

Proof The proof follows from the fact that any unimodal prior $\pi(\theta_i)$ can be written as a convex combination of uniform priors. □

From the Corollary it follows that the inconsistency of the Bayes factor $B_{10}(x_1, \ldots, x_k)$ under some alternative models persists if we use the intrinsic linking prior $\pi^I(\theta_i|\theta_0)$. Thus, the asymptotic behavior of the meta-inference is an open problem that deserves more research.

6 Concluding Remarks

Two main difficulties arise in meta-analysis, the analysis of the heterogeneity of the samples involved, which generates a clustering problem, and the construction of the linking distribution, the distribution of the experimental parameters conditional on the meta-parameter.

We have claimed that clustering the samples using a Bayesian approach is a natural way of investigating the heterogeneity of the samples. Further, this approach allows us incorporating the model uncertainty in the meta-inference. Example 1 showed that the inference on the meta-parameter dramatically changes across clusters, and, consequently, to take for granted that all samples are heterogenous might give serious misleading results. Thus, the recommendation is to analyze the cluster structure of the samples and then proceed with the meta-analysis, conditional on the cluster models. The final inference is then obtained as a mixture of the inferences conditional on the clusters, using the posterior distribution of the clusters as the mixing distribution.

We have also claimed that the joint prior of the experimental parameter and the meta-parameter has be chosen in the class of bidimensional priors with given marginals, the so-called Frèchet class. This condition ensures that the experimental model and the meta-model are coherent. We have shown that the intrinsic prior, which is obtained from the model comparison between the experimental model and the meta-model, is a conditional prior that satisfies the requirements for a reasonable linking distribution. For a discussion on the linking distribution for binomial samples see [24].

Acknowledgments This paper has been supported by Ministerio de Ciencia y Tecnología, Grant MTM2011-28945.

References

1. Babapulle, M.N., Joseph, L., Bélisle, P., Brophy, J.M., Eisenberg, M.J.: A hierarchical Bayesian meta-analysis of randomized clinical trials of drug-eluting stents. Lancet **364**, 583–591 (2004)
2. Bhaumik, D.K., Amatya, A., Normand, S.T., Greenhouse, J., Kaizar, E., Neelon, B., Gibbons, R.: Meta-analysis of rare binary adverse event data. J. Am. Stat. Assoc. **107**, 555–567 (2012)

3. Barry, D., Hartigan, J.A.: Product partition models for change point problems. Ann. Stat. **20**, 260–279 (1992)
4. Berger, J.O., Pericchi, L.R.: The intrinsic Bayes factor for model selection and prediction. J. Am. Stat. Assoc. **91**, 109–122 (1996)
5. Berger, J.O., Pericchi, L.R.: Objective Bayesian model selection: introduction and comparisons. In: Lahiri, P. (eds.) Model Selection. Lectures Notes of the Institute of Mathematical Statistics, pp. 135–207 (2001)
6. Canner, P.L.: An overview of six clinical trials of aspirin in coronary heart disease. Stat. Med. **6**, 255–263 (1987)
7. Carlin, J.B.: Meta-analysis for 2×2 tables: a Bayesian approach. Stat. Med. **11**, 141–159 (1992)
8. Casella, G., Moreno, E.: Intrinsic meta-analysis of contingency tables. Stat. Med. **24**, 583–604 (2005)
9. Casella, G., Moreno, E.: Assessing robustness of intrinsic test of independence in two-way contingency tables. J. Am. Stat. Assoc. **104**, 1261–1271 (2009)
10. Casella, G., Moreno, E., Girón, F.J.: Cluster analysis, model selection, and prior distributions on models. Bayesian Anal. **9**(3), 613–658 (2014)
11. Casella, G., Girón, F.J., Martinez, M.J., Moreno, E.: Consistency of Bayesian procedures for variable selection. Ann. Stat. **37**, 1207–1228 (2009)
12. Consonni, G., Moreno, E., Venturini, S.: Testing Hardy–Weinberg equilibrium: an objective Bayesian analysis. Stat. Med. **30**(1), 62–74 (2011)
13. Cuadras, C.M.: Probability distributions with given multivariate marginals and given dependence structure. J. Multivar. Anal. **42**, 51–66 (1992)
14. Cuadras, C.M.: Constructing copula functions with weighted geometric means. J. Stat. Plan. Inference **139**, 3766–3772 (2009)
15. DuMouchel, W., Waternaux, C.: Hierarchical models for combining information and for meta-analyses (with discussion). In: Bernando, J.M., Berger, J., Dawid, A., Smith, A. (eds.) Bayesian Statistics 4, pp. 338–341. Clarendon Press, Oxford (1992)
16. Jeffreys, H.: Theory of Probability, 3rd edn. Clarendon Press, Oxford (1961)
17. Joe, H.: Multivariate Models and Dependence Concepts. Monographs in Statistics and Probability, vol. 73. Chapman and Hall, New York (1997)
18. Kadane, J.K.: Principles of Uncertainty. Chapman & Hall/CRC Texts in Statistical Science, Boca Raton (2011)
19. Larose, D.R., Dey, D.K.: Grouped random effects models for Bayesian meta-analysis. Stat. Med. **16**, 1817–1829 (1997)
20. Lindley, D.V.: Introduction to Probability and Statistics from a Bayesian Viewpoint. Cambridge Unvierstiy Press, Cambridge (1965)
21. Moreno, E.: Bayes factors for intrinsic and fractional priors in nested models: Bayesian robustness. In: Dodge, Y. (eds.) IMS Lectures Notes-Monograph Series, vol. 31, pp. 257–270 (1997)
22. Moreno, E., Bertolino, F., Racugno, W.: An intrinsic limiting procedure for model selection and hypothesis testing. J. Am. Stat. Assoc. **93**, 1451–1460 (1998)
23. Moreno, E., Girón, F.J., Casella, G.: Consistency of objective bayes factors as the model dimension grows. Ann. Stat. **38**, 1937–1952 (2010)
24. Moreno, E., Girón, F.J., Casella, G.: Posterior model consistency in variable selection as the model dimension grows. Stat. Sci. **30**(2), 228–241 (2015)
25. Moreno, E., Vázquez-Polo, F.J., Negrín, M.A.: Objective Bayesian meta-analysis for sparse discrete data. Stat. Med. **33**(21), 3676–3692 (2014)
26. Morgenstern, D.: Einfache Beispiele zweidimensionaler Verteilungen. Mitteilungsblatt für Mathematische Statistik **8**, 234–235 (1956)
27. Morris, C.N., Normand, S.L.: Hierarchical models for combining information and for meta analyses. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) Bayesian Statistics 4, pp. 321–344. Clarendon Press, Oxford (1992)

28. Sarmanov, O.V.: Generalized normal correlated and two-dimensional Fréchet classes. Doklady (Soviet Mathematics) **168**, 596–599 (1966)
29. Schömig, A., Mehilli, J., Waha, A.D., Seyfarth, M., Pache, J., Kastrati, A.: A meta-analysis of 17 randomized trials of a percutaneous coronary intervention-based strategy in patients with stable coronary artery disease. J. Am. College Cardiol. **52**, 894–904 (2008)
30. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
31. Sutton, A.J., Higgins, J.P.: Recent developments in meta-analysis. Stat. Med. **27**, 625–650 (2008)

Statistical Evaluation of Forensic DNA Mixtures from Multiple Traces

Julia Mortera

Abstract

A statistical model for the quantitative peak information obtained from a forensic DNA mixture sample is illustrated on a real case example. We use the combined information from two DNA traces: to find likelihood ratios to quantify the strength of evidence; to deconvolve the mixtures for the purpose of finding likely profiles of unknown contributors to the traces; and to analyse the artefacts that might be present in the mixture after DNA amplification.

1 Introduction

DNA is now routinely used in criminal investigations and court cases, although DNA samples taken at crime scenes vary in quality and thus present challenging problems for their interpretation. The identification of the DNA composition of mixed samples gives rise to a wide range of challenging statistical questions, some associated with uncertainties and artefacts in the measurement processes and some associated with population genetic variations. A new statistical model for the peak heights of a DNA mixture is presented in [4]. The simplifications of the model combined with an efficient Bayesian network representation enables fast computation and permits analysis of complex mixtures, allowing for simultaneous analysis of the evidence from several DNA samples and the identification of the artefacts.

J. Mortera (✉)
Department of Economics, Università Roma Tre, Roma, Italy
e-mail: julia.mortera@uniroma3.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_16

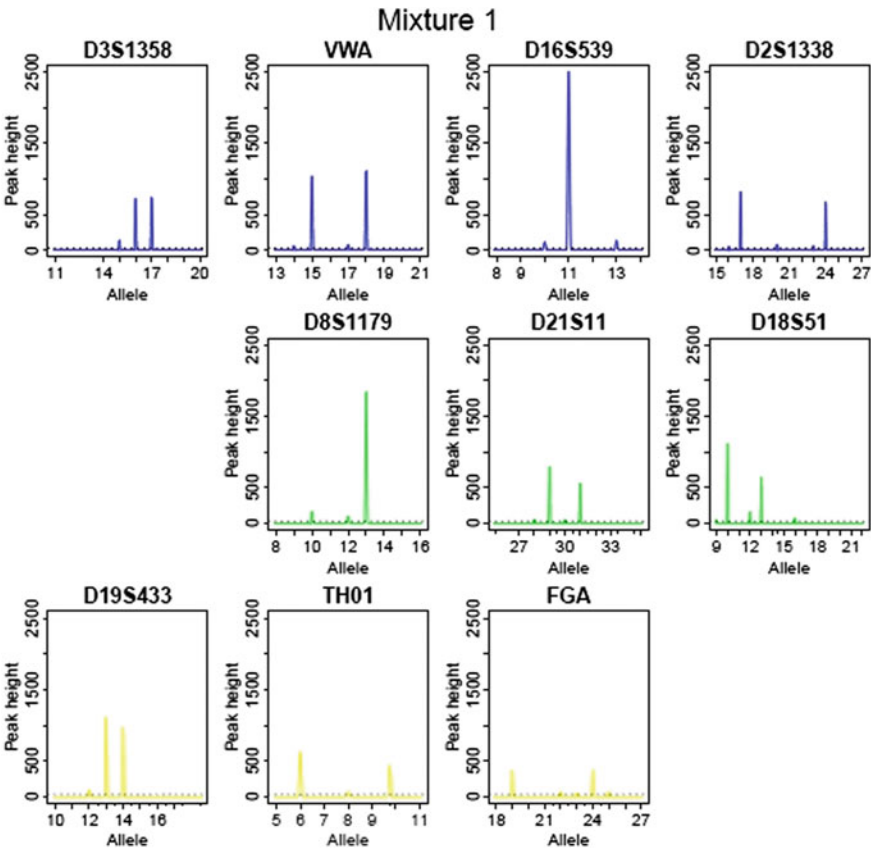


Fig. 1 Electropherogram (EPG) showing the peak heights at the different alleles in DNA mixture for the first trace $T1$ on 10 markers

As a motivating example we consider a DNA mixture case from [10]. The electropherogram (EPG) showing the DNA mixture for the first trace $T1$, is given in Fig. 1.

Table 1 shows the alleles a in the mixture, the corresponding peak heights H_a for traces T1 and T2, and genotypes $\mathbf{g}$ of two typed individuals K_1 and K_2, assumed to have contributed to the mixture, for an excerpt of the 10 markers.

Note, that if the two traces T1 and T2, consist of DNA from K_1 and K_2 alone, the peaks at alleles 14 and 17 of marker VWA in both traces, and for T2 the peak at allele 12 of marker D16 would need to be due to an artefact termed stutter, as neither K_1 nor K_2 have these alleles. Also K_2's allele 32.2 on marker D21 would have dropped out of the trace T1 (see Sect. 2 for definitions of stutter and dropout). Note that, taking into account these artefacts, the entire profile would be consistent with only DNA from K_1 and K_2 being present in the traces.

Table 1 Alleles a, peak heights in traces, T1 and T2, and genotypes $\mathbf{g}$ of individuals K_1 and K_2 for an excerpt of the 10 markers

Marker	Allele	T1	T2	K_1	K_2
D3	15	132	549		15
	16	719	646	16	
	17	736	1131	17	17
VWA	14	56	61		
	15	1033	1365	15	15
	17	71	80		
	18	1113	1216	18	18
D16	10	110	96		
	11	2496	2312	11	11
	12	0	49		
	13	129	536		13
D21	28	43	44		
	29	794	969	29	29
	30	39	49		
	31	561	686	31	
	32.2	0	444		32.2

The objective of an analysis of a DNA profile can be a quantification of the *strength of evidence* for a given hypothesis over another, or it may be a *deconvolution* of the profile, *i.e.*, to identify likely genotypes of contributors. The *evidence* E, in Table 1 consists of the peak heights T1 and T2 and the genotypes of the known individuals K_1 and K_2, $E = \{\text{T1, T2, } K_1, K_2\}$. To quantify the strength of the evidence against the defendants K_1 and K_2, two competing hypotheses are typically specified. The *prosecution hypothesis*, $\mathscr{H}_p$, *e.g.*, which the contributors are K_1 and K_2, which is then compared to a *defence hypothesis*, say, $\mathscr{H}_d : U_1 \& U_2$ that the contributors are two unknown individuals assumed to be chosen randomly and independently from a reference population with known allele frequencies.

The strength of the evidence (see [1,9]) is normally represented by the *likelihood ratio*:

$$LR = \Pr(E \mid \mathscr{H}_p)\}/\Pr(E \mid \mathscr{H}_d).$$

Following [1] we report the *weight of evidence* as

$$\text{WoE} = \log_{10} LR$$

in the unit *ban* introduced by Alan Turing [6].

Another objective of the analysis is the *deconvolution* or separation of DNA mixtures to identify the combined genotypes across all markers of each unknown contributor to the mixture and give a list of potential genotypes of a perpetrator to use, for example, in a database search. For example, we could wish to calculate

the predictive probability $\Pr\{U_1, U_2 \mid E, \mathscr{H}\}$, where U_1, U_2 represent genotypes of unknown contributors under an investigative hypothesis $\mathscr{H}$.

2 A Gamma Model with Artefacts

We now briefly describe the gamma model for peak heights which is based on [2–4]. We consider I potential contributors to a DNA mixture. Let there be M markers used in the analysis of the mixture with marker m having A_m allelic types, $m = 1, \ldots, M$. Let ϕ_i denote the *proportion* of DNA from individual i prior to PCR amplification, with $\phi = (\phi_1, \phi_2, \ldots, \phi_I)$, $\phi_i \geq 0$ and $\sum_{i=1}^{I} \phi_i = 1$.

The peak height H_{ia} is roughly proportional to the amount of DNA of type a contributed by individual i and, for fixed ϕ, we assume it has a gamma distribution

$$H_{ia} \sim \Gamma(\rho \phi_i n_{ia}, \eta),$$

where ρ is proportional to the total amount of DNA prior to amplification; n_{ia} is the number of alleles of type a carried by individual i; and η determines scale.

The individual peak heights H_{ia} are not observable, but ignoring artefacts we observe the aggregates $H_a = \sum_{i \in I} H_{ia}$, which also have a gamma distribution $H_a \sim \Gamma\{\rho B_a(\phi, \mathbf{n}), \eta\}$, where $B_a(\phi, \mathbf{n}) = \sum_i \phi_i n_{ia}$ and $\mathbf{n} = (n_{ia}, i = 1, \ldots, I; a = 1, \ldots, A_m)$. Then — in a trace with only one heterozygous diploid contributor and no artefacts — $\mu = \rho\eta$ is the *mean peak height* and $\sigma = 1/\sqrt{\rho}$ the *coefficient of variation* for peak heights.

Incorporating artefacts. One important artefact associated with the PCR amplification process is known as *stutter*, *i.e.*, small proportion of DNA molecules tend to lose one repeat number in the amplification process so some of the alleles at a show up in position $a - 1$. The peak height at allele a is thus

$$Y_a = H_a - S_a H_a + S_{a+1} H_{a+1},$$

S_a being the proportion of the ideal peak height H_a that has been lost to the allele $a - 1$ with

$$S_a \sim \text{Beta}(\xi \rho B_a(\phi, \mathbf{n}), (1 - \xi)\rho B_a(\phi, \mathbf{n}))$$

where ξ denotes the mean stutter proportion.

In mixtures with small amounts of DNA, some alleles present in the mixture *dropout*, *i.e.*, are not amplified and so no corresponding peak can be observed above a threshold C. This implies that we are not observing Y_a, but rather Z_a, having a gamma cdf, where

$$Z_a = \begin{cases} Y_a & \text{if } Y_a > C \\ 0 & \text{otherwise.} \end{cases}$$

Likelihood function For given mixture composition, given genotypes of the contributors $\mathbf{n}$, given proportions ϕ, and fixed values of parameters (ρ, ξ, η), all observed

peak heights are independent. Thus the conditional likelihood function based on the observations $\mathbf{z} = \{z_{ma}\}_{m\in M, a\in A_m}$ for all markers m and alleles a is

$$L(\rho, \xi, \phi, \eta \mid \mathbf{z}, \mathbf{n}) = \prod_m \prod_a L_{ma}(z_{ma}).$$

For a given hypothesis $\mathscr{H}$, the full likelihood is obtained by summing over all possible combinations of genotypes with probabilities associated with the hypothesis to give

$$L(\mathscr{H}) = P(E \mid \mathscr{H}) = \sum_{\mathbf{n}} L(\rho, \xi, \phi, \eta \mid \mathbf{z}, \mathbf{n}) P(\mathbf{n} \mid \mathscr{H}).$$

This sum is astronomical in size for any hypothesis which potentially involves unknown contributors to the mixture, but can be calculated efficiently by appropriate use of *Bayesian network techniques*.

3 Results

Table 2 shows the maximum likelihood estimates of the parameters and the corresponding standard error for the DNA data given in Table 1. These are obtained using software `DNAmixtures` [7].

Note that the parameter estimates are almost the same under $\mathscr{H}_p : K_1\&K_2$ and $\mathscr{H}_d : U_1\&U_2$. Furthermore, the proportion of DNA in T1 is unbalanced, K_1 having contributed 88 % of the DNA, $\phi_{K_1} = 0.88$, whereas, in T2 the proportion is balanced, having $\phi_{K_1} = 0.54$.

The weight of evidence for $\mathscr{H}_p : K_1\&K_2$ versus $\mathscr{H}_d : U_1\&U_2$ is $WoE = -211.3 - (-236.8) = 25.6$ whereas, if the defence hypothesis were $\mathscr{H}_d : K_1\&U_1$ the $WoE = 12.4$ which is less incriminating for K_2. We emphasize that when using only trace T1 the weight of evidence is $WoE = 20.7$, and when using only the data in T2 the weight of evidence is $WoE = 21.6$ yielding, in both cases, lower evidential strength than using the simultaneous information in both traces.

Table 2 Maximum likelihood estimates when information in the traces T1 and T2 is combined

	$\mathscr{H}_p : K_1\&K2$					$\mathscr{H}_d : U_1\&U_2$			
	T1		T2			T1		T2	
Par.	Est.	SE	Est.	SE	Par.	Est.	SE	Est.	SE
μ	902	49.4	1133	49.5	μ	901	49.3	1133	49.6
σ	0.245	0.027	0.195	0.022	σ	0.244	0.027	0.196	0.023
ξ	0.061	0.009	0.061	0.009	ξ	0.060	0.009	0.060	0.009
ϕ_{K_1}	0.884	0.024	0.544	0.032	ϕ_{U_1}	0.884	0.024	0.545	0.032
ϕ_{K_2}	0.117	0.024	0.456	0.032	ϕ_{U_2}	0.116	0.024	0.455	0.032
$\log_{10} \hat{L}(\mathscr{H})$				−211.3	$\log_{10} \hat{L}(\mathscr{H})$				−236.8

Table 3 Most probable genotypes of U_1 and U_2

	D3	VWA	D16	D21
U_1	(16, 17)	(15, 18)	(11, 11)	(29, 31)
U_2	(15, 17)	(15, 18)	(11, 13)	(29, 32.2)
Probability	0.999	0.998	1	0.994

Mixture deconvolution. For mixture deconvolution we consider the traces jointly under the hypothesis $\mathscr{H}_d : U_1 \& U_2$ and identify the most probable genotypes of the unknown individuals U_1 and U_2. For all markers, the predictive genotypes of the two unknown individuals share the profile of K_1 and K_2. Table 3 gives an extract of the deconvolved genotype together with the predictive probability.

The overall probability that the profile of the unknown contributors is the one predicted is 0.95.

Interpreting artefacts. Our model does not impose at the outset that a specific peak or allele is due to stutter, or has dropped out. One of the features of software `DNAmixtures` [7] is to produce Bayesian networks for each marker with peak height evidence propagated. Thanks to the flexibility of Bayesian networks we can then elaborate these so as to identify the possibility that artefacts might be present in some of the markers. For this purpose, we use object-oriented Bayesian networks (OOBN) [5]. We introduce networks that represent the presence of stutter and dropout. Instances of the networks are then combined with the networks produced by `DNAmixtures`. In this way, we can answer queries like: what is the probability that an allele is due to stutter? that an allele has dropped out?

Figure 2 shows the global OOBN for marker D21 used for artefact identification. In this network the output nodes are: `n_1_a` and `n_2_a` representing the total allele counts for a specific allele a for U_1 and U_2, and the Boolean nodes `O_a` indicating whether allele a is observed or not. The networks for stutter and dropout are shown in Fig. 3.

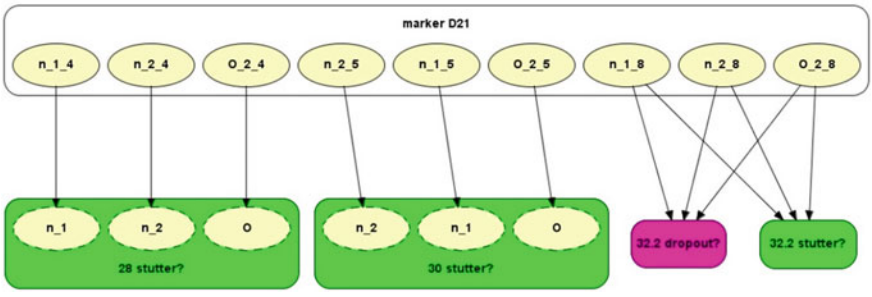


Fig. 2 OOBN for artefact identification in marker D21

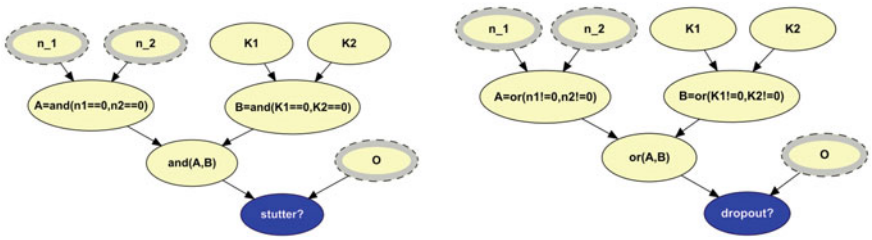


Fig. 3 Networks for stutter and dropout

Table 4 Posterior probabilities for alleles to have dropped out and to be due only to stutter for trace T2

Marker(a)	P(dropout)	P(stutter)
VWA(14)		1
VWA(17)		0.999
D21(28)		1
D21(30)		0.998
D21(32.2)	0.999	0.001

Using the parameter estimates in Table 2 under the defence hypothesis $\mathscr{H}_d$: $K_1 \& U_1$ and using the networks described above, we obtain the posterior probabilities for selected alleles to have dropped out and to be due to stutter only as shown in Table 4. Alleles 14 and 17 for marker VWA and alleles 28 and 30 for markers D21 have a very high probability of being due to stutter alone. Allele 32.2 of marker D21 has a probability around 0.999 of having dropped out of trace T2.

4 Concluding Remarks

The combined analysis of two traces gives more informative results than what emerges from separate analyses of the two traces, not only for evidential and deconvolution purposes, but also for the analysis of artefacts. We refrain from reporting the separate analyses here.

Here we have treated the allele frequencies as fixed and known, although in principle they also are parameters which should be estimated. In future work, we will incorporate both this uncertainty, identity by descent (IBD) and subpopulation issues following [8].

References

1. Balding, D.: Evaluation of mixed-source, low-template DNA profiles in forensic science. Proc. Natl Acad. Sci. USA **110**(30), 12241–12246 (2013)
2. Cowell, R.G., Lauritzen, S.L., Mortera, J.: A gamma model for DNA mixture analyses. Bayesian Anal. **2**(2), 333–348 (2007)
3. Cowell, R.G., Lauritzen, S.L., Mortera, J.: Probabilistic expert systems for handling artifacts in complex DNA mixtures. Forensic Sci. Int.: Genet. **5**(3), 202–209 (2011)
4. Cowell, R.G., Graversen, T., Lauritzen, S., Mortera, J.: Analysis of DNA mixtures with artefacts. J. R. Stat. Soc. Ser. C (with discussion) **64**, 1–48 (2015)
5. Dawid, A.P., Mortera, J., Vicard, P.: Object-oriented Bayesian networks for complex forensic DNA profiling problems. Forensic Sci. Int. **169**, 195–205 (2007)
6. Good, I.J.: Studies in the history of probability and statistics. XXXVII A. M. Turing's statistical work in World War II. Biometrika **66**(2), 393–396 (1979)
7. Graversen, T.: DNAmixtures: statistical inference for mixed samples of DNA. R package version 0.1-3 (2013). http://dnamixtures.r-forge.r-project.org
8. Green, P.J., Mortera, J.: Sensitivity of inferences in forensic genetics to assumptions about founder genes. Ann. Appl. Stat. **3**(2), 731–763 (2009)
9. Lindley, D.V.: A problem in forensic science. Biometrika **64**(2), 207–213 (1977)
10. Puch-Solis, R., Rodgers, L., Mazumder, A., Pope, S., Evett, I., Curran, J., Balding, D.: Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters. Forensic Sci. Int.: Genet. **7**(5), 555–563 (2013)

A Note on Semivariogram

Giovanni Pistone and Grazia Vicario

Abstract

(Semi)Variograms are usually discussed in the framework of stationary or intrinsically stationary processes. We retell here this piece of theory in the setting of generic Gaussian vectors and of Gaussian vectors with constant variance. We show how to reparametrize the distribution as a function of the variogram and how to characterise all the Gaussian distribution with a given variogram.

1 Introduction

If $(Y_t)_{t\in T}$ is a Gaussian random field, its variogram is the mapping from 2-sets $\{s,t\}$, $s,t\in T$, to $\mathrm{Var}\,(Y_s-Y_t)\,/2$. In some applied fields, such as Geostatistics or Metrology, such a multivariate parameter is considered more telling than the process correlation $\{s,t\}\mapsto \mathrm{Cor}\,(Y_s,Y_t)$. For example, if a meaningful distance $\{s,t\}\mapsto d(s,t)$ is available, one would like the variogram values to increase with distance to model the larger randomness at far away locations.

In this paper we discuss some general topics on variogram that was originally discussed by Matheron [8] under assumptions of stationarity and homogeneity. Modern expositions are to be found in [2, Chap. 2] [4, Chap. 2], [5, Chap. 1], [6]. The present

G. Pistone (✉)
de Castro Statistics, Collegio Carlo Alberto, Moncalieri, Italy
e-mail: giovanni.pistone@carloalberto.org

G. Vicario
DISMA Giuseppe Luigi Lagrange, Politecnico di Torino, Torino, Italy
e-mail: grazia.vicario@polito.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_17

piece of research work was prompted by the need of a clear and sound methodology in the occasion of previous applied research [9,10].

Our goal now is to rework the basic mathematics in order to prepare for a future better treatment of a number of items of interest e.g.,

- simulation of a Gaussian random field with given variogram;
- geometry of the Gaussian model based on the use of variograms as parameters, in the sense of [1,12];
- parsimonious models, e.g., graphical models [7], parametrized by variograms;
- Bayes approach to Kriging, especially nonparametric Bayes.

In Sect. 2 we formally discuss the case of a generic Gaussian vector and of the special case of a constant variance. In Sect. 3 we briefly discuss the connection with the case of Gaussian stationary random fields. A few conclusions are discussed in the final section.

2 Variogram of a Normal Vector

We first consider a generic Gaussian vector and we plan to specialise our assumptions later on.

Definition 1 Assume $Y \sim \mathrm{N}_n(\boldsymbol{\mu}, \Sigma), \boldsymbol{\mu} = (\mu_i : i = 1, \dots, n), \Sigma = [\sigma_{ij}]_{i,j=1}^n$. The *(semi)variogram* of Y is the $n \times n$ matrix $\Gamma = [\gamma_{ij}]_{i,j=1}^n$ with

$$2\gamma_{ij} = \mathrm{Var}\left(Y_i - Y_j\right) = (\boldsymbol{e}_i - \boldsymbol{e}_j)' \Sigma (\boldsymbol{e}_i - \boldsymbol{e}_j) = \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}.$$

The matrix Γ can be written

$$\Gamma = \frac{1}{2}\left(\boldsymbol{vdiag}\,(\Sigma)\,\mathbf{1}' + \mathbf{1}\boldsymbol{vdiag}\,(\Sigma)'\right) - \Sigma = \frac{1}{2}\left(\mathrm{diag}\,(\Sigma)\,\mathbf{1}\mathbf{1}' + \mathbf{1}\mathbf{1}'\,\mathrm{diag}\,(\Sigma)\right) - \Sigma$$

where $\mathbf{1}$ is the unit column vector and $\boldsymbol{vdiag}\,(\Sigma) = \mathrm{diag}\,(\Sigma)\,\mathbf{1}$ is the diagonal of Σ as a column vector. Recall that $\frac{1}{n}\mathbf{1}\mathbf{1}'$ is the orthogonal projector on Span $(\mathbf{1})$.

Let us recall the basic properties of the variogram matrix.

Proposition 1 *The variogram Γ is symmetric, with zero diagonal, and it is* conditionally negative definite.

Proof The quadratic form of

$$\Sigma = \frac{1}{2}\left(\boldsymbol{vdiag}\,(\Sigma)\,\mathbf{1}' + \mathbf{1}\boldsymbol{vdiag}\,(\Sigma)'\right) - \Gamma$$

at $\boldsymbol{\alpha} \in \mathbb{R}^n$ is

$$\boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha} = (\boldsymbol{\alpha} \cdot \mathbf{1})(\boldsymbol{\alpha} \cdot \boldsymbol{vdiag}\,(\Sigma)) - \boldsymbol{\alpha}'\Gamma\boldsymbol{\alpha},$$

hence $\boldsymbol{\alpha} \cdot \mathbf{1} = 0$ implies $\boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha} = -\boldsymbol{\alpha}'\Gamma\boldsymbol{\alpha}$, in particular Γ is negative definite conditionally to $\sum_j \alpha_j = 0$.

Definition 2 A nonzero symmetric matrix which has zero diagonal and is conditionally negative definite will be called a *variogram matrix*.

Proposition 2 *Let Γ be a variogram matrix. There exist nonnegative $\mu_1, \ldots, \mu_{n-1}$ and orthonormal vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{n-1}$ in* Span $(\mathbf{1})^{\perp}$ *such that*

$$\Gamma = \frac{\sum_{j=1}^{n-1} \mu_j}{n} \mathbf{1} \otimes \mathbf{1} - \sum_{j=1}^{n-1} \mu_j \boldsymbol{w}_j \otimes \boldsymbol{w}_j \tag{1}$$

Proof For each matrix $U = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_{n-1}] \in \mathbb{R}^{n \times (n-1)}$ such that $U^T U = I_{n-1}$ and $\mathbf{1}^T U = 0$, the matrix $\Sigma_0 = -U^T \Gamma U \in \mathbb{R}^{n \times (n-1)}$ is nonnegative definite. It follows that $V^T \Sigma_0 V = \text{diag}\,(\mu_i : j - 1, \ldots, n-1)$ for some $V \in \mathbb{O}_{n-1}$, $\mu_i \geq 0$, $i = 1, \ldots, n-1$, hence $(UV)^T \Gamma (UV) = \text{diag}\,(\mu_i : j - 1, \ldots, n-1)$. If $W = UV \in \mathbb{R}^{n \times (n-1)}$, then $W^T W = V^T U^T U V = V^T V = I_{n-1}$ and $\mathbf{1}^T W = 0$. If $W = [\boldsymbol{w}_1 \cdots \boldsymbol{w}_{n-1}]$, then $(w_j, -\mu_j)$, $j = 1, \ldots, n-1$ are couples of eigenvectors and eigenvalues of $-\Gamma$. As Γ has zero trace, then the n-eigenvalue of Γ is $\sum_{j=1}^{n-1} \mu_j > 0$. Its eigen space must be orthogonal to all $\boldsymbol{w}_j$', hence it contains Span $(\mathbf{1})$.

Computation of the parameters suggests that the variogram matrix carries $n(n-1)/2$ degrees of freedom, while the diagonal of Σ carries n df. Together, Λ and Γ carry as many degrees of freedom as Σ, i.e., $n(n-1)/2 + n = (n+1)n/2$. More precisely, we have the following re-parametrization of Σ.

Proposition 3 *1. The mapping from a positive definite Σ to a positive diagonal Λ and a variogram matrix Γ defined by*

$$\Sigma \mapsto \left(\text{diag}\,(\Sigma)\,, \frac{1}{2}(\boldsymbol{v}\boldsymbol{diag}\,(\Sigma)\,\mathbf{1}' + \mathbf{1}\boldsymbol{v}\boldsymbol{diag}\,(\Sigma)') - \Sigma \right) = (\Lambda, \Gamma) \tag{2}$$

is the restriction to the cone of positive definite matrices of a linear map on $n \times n$ real matrices. It is injective with inverse

$$(\Lambda, \Gamma) \mapsto \frac{1}{2}(\Lambda \mathbf{1}\mathbf{1}' + \mathbf{1}\mathbf{1}'\Lambda) - \Gamma = \frac{1}{2}(\mathbf{vec}\,(\Lambda)\,\mathbf{1}' + \mathbf{1}\mathbf{vec}\,(\Lambda)') - \Gamma. \tag{3}$$

2. The range of the mapping (1) consists of all Λ, Γ positive diagonal and symmetric, respectively, and satisfying

$$(\boldsymbol{\beta} \cdot \mathbf{1})(\boldsymbol{\beta} \cdot \mathbf{vec}\,(\Lambda)) \geq \boldsymbol{\beta}' \Gamma \boldsymbol{\beta}, \quad \boldsymbol{\beta} \in \mathbb{R}^n\,. \tag{4}$$

3. In particular, Γ is conditionally negative definite and

$$n\,\text{Tr}\,(\Lambda) \geq \mathbf{1}'\Gamma\mathbf{1} = \sum_{i,j=1}^{n} \gamma_{ij}\,. \tag{5}$$

4. If the spectral decomposition (1) holds, then the condition (5) on Λ becomes $\text{Tr}\,(\Lambda) \geq \sum_j \mu_j$.

Proof 1. If $\Sigma_i \mapsto (\Lambda_i, \Gamma_i)$, $i = 1, 2$, and $(\Lambda_1, \Gamma_1) = (\Lambda_2, \Gamma_2)$, then $\operatorname{diag}(\Sigma_1) = \operatorname{diag}(\Sigma_2)$ and $\Sigma_1 = \Sigma_2$ follows from $\Gamma_1 = \Gamma_2$.

2. Let Λ and Γ be generic positive diagonal and conditionally negative definite, respectively. Then for a generic $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0 + \overline{\alpha}\mathbf{1}$, with $\boldsymbol{\alpha}_0 \cdot \mathbf{1} = 0$ and $\bar{\alpha} = \frac{1}{n}\boldsymbol{\alpha} \cdot \mathbf{1}$, we have

$$\boldsymbol{\alpha}' \left[\frac{1}{2}(\Lambda \mathbf{1}\mathbf{1}' + \mathbf{1}\mathbf{1}'\Lambda) - \Gamma \right] \boldsymbol{\alpha} = n\overline{\alpha}\, \boldsymbol{\alpha} \cdot \mathbf{vec}(\Lambda) - \boldsymbol{\alpha}'\Gamma\boldsymbol{\alpha}$$
$$= \begin{cases} -\boldsymbol{\alpha}_0 \Gamma \boldsymbol{\alpha}_0 \geq 0 & \text{if } \bar{\alpha} = 0\,, \\ \boldsymbol{\alpha} \cdot \mathbf{vec}(\Lambda) - \boldsymbol{\alpha}'\Gamma\boldsymbol{\alpha} & \text{if } n\bar{\alpha} = 1\,. \end{cases}$$

Finally, we take $\boldsymbol{\alpha} = (\boldsymbol{\beta} \cdot \mathbf{1})^{-1}\boldsymbol{\beta}$ to obtain (4).

3. Equation (4) implies a conditionally negative definite Γ if $\boldsymbol{\beta} \cdot \mathbf{1} = 0$. Otherwise, if $\boldsymbol{\beta} = \mathbf{1}$ the inequality becomes (5).

4. If the spectral decomposition holds, then $\mathbf{1}^T \Gamma \mathbf{1} = n \sum_{j=1}^{n-1} \mu_J$.

Remark 1 If $\det(\Sigma) \neq 0$, similar formulæ are obtained by considering the correlation matrix

$$R = (\operatorname{diag}\Sigma)^{-1/2}\Sigma(\operatorname{diag}\Sigma)^{-1/2},$$

viz

$$\Gamma = \frac{1}{2}(\boldsymbol{vdiag}(\Sigma)\,\mathbf{1}' + \mathbf{1}\boldsymbol{vdiag}(\Sigma)') - (\operatorname{diag}\Sigma)^{1/2} R (\operatorname{diag}\Sigma)^{1/2}$$
$$= \Lambda^{1/2}\left(\frac{1}{2}\left(\Lambda^{1/2}\mathbf{1}\mathbf{1}'\Lambda^{-1/2} + \Lambda^{-1/2}\mathbf{1}\mathbf{1}'\Lambda^{1/2}\right) - R\right)\Lambda^{1/2}\,. \tag{6}$$

where $\operatorname{diag}\Sigma = \Lambda$.

This formula is sometimes preferred in the applied literature because both Γ and R carry the same number of degrees of freedom and they are thought as being equivalent assignments. However, it is important to consider that the imputation of the a coherent diagonal Λ depends on Γ.

Given a variogram matrix Γ, Eq. (5) is a linear bound on Λ. Here, we do not discuss it in full generality, but we move to consider the case where the variance is constant. Such an assumption is of interest in applications where a minimum of stationarity must be assumed.

Proposition 4 *Assume that the variance is constant,* $\operatorname{diag}(\Sigma) = \lambda I_n$.

1. *Equations (2) and (3) become*

$$\Sigma \mapsto \left(\lambda, \lambda\mathbf{1}\mathbf{1}' - \Sigma\right) = (\lambda, \Gamma) \tag{7}$$

and

$$(\lambda, \Gamma) \mapsto \lambda\mathbf{1}\mathbf{1}' - \Gamma = \Sigma, \tag{8}$$

respectively.

2. *The existence condition (5) on* λ *becomes*

$$n^2\lambda \geq \sum_{i,j=1}^{n} \gamma_{ij} \,. \tag{9}$$

3. *The correlation Eq. (6) becomes*

$$\Gamma = \lambda(\mathbf{1}\mathbf{1}^T - R) \,.$$

4. *If* $n\lambda > \mathbf{1}'\Gamma\mathbf{1}$, *then* $\det\Gamma \neq 0$ *then* Σ *is invertible and, in such a case,*

$$\Gamma^{-1} = -\Sigma^{-1} - \lambda(1 - \lambda\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}\Sigma^{-1}\mathbf{1}\mathbf{1}'\Sigma^{-1} \,, \tag{10}$$

$$\Sigma^{-1} = -\Gamma^{-1} - \lambda(1 - \lambda\mathbf{1}'\Gamma^{-1}\mathbf{1})^{-1}\Gamma^{-1}\mathbf{1}\mathbf{1}'\Gamma^{-1} \,. \tag{11}$$

Proof Everything but the last item is a special case of Proposition 3. If the matrices Σ and Γ are both invertible, from the Sherman–Morrison formula we obtain Eqs. (10) and (11). Assume $\det\Gamma \neq 0$ and $n^2\lambda > \mathbf{1}'\Gamma\mathbf{1}$. From the spectral representation of Γ in Eq. (1), we derive $\mathbf{1}'\gamma^{-1}\mathbf{1} = n\left(\sum_{j=1}^{n-1}\right)^{-1}$. It follows from the assumption that $1 - \lambda\mathbf{1}'\Gamma^{-1}\mathbf{1} \neq 0$, so that the Sherman–Morrison formulæ hold.

Remark 2 The knowledge of the support of the parametrization with λ and Γ is crucial in the choice of a coherent apriori distribution.

Let us discuss first the case $n = 2$. We have

$$\begin{bmatrix} \sigma & \sigma_{1,2} \\ \sigma_{1,2} & \sigma \end{bmatrix} = \begin{bmatrix} \lambda & \lambda - \gamma \\ \lambda - \gamma & \lambda \end{bmatrix}, \quad \gamma = \sigma > 0,$$

and we need the sign of

$$\det\begin{bmatrix} \lambda & \lambda - \gamma \\ \lambda - \gamma & \lambda \end{bmatrix} = \lambda^2 - (\lambda - \gamma)^2 = \gamma(2\lambda - \gamma),$$

which is positive if $\lambda \geq \gamma/2$. This shows existence and shows that there is a restriction on λ which is worthwhile to investigate further. The condition in (9) involves the lower bound $\max_\alpha 2\alpha(1-\alpha)\gamma = \gamma/2$. If the lower bound is reached with $\lambda = \gamma/2$, hence $\det\Gamma = 0$.

Assume now $n = 3$, that is,

$$\begin{bmatrix} \sigma & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{1,2} & \sigma & \sigma_{2,3} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma \end{bmatrix} = \begin{bmatrix} \lambda & \lambda - \gamma_{12} & \lambda - \gamma_{13} \\ \lambda - \gamma_{12} & \lambda & \lambda - \gamma_{23} \\ \lambda - \gamma_{13} & \lambda - \gamma_{23} & \lambda \end{bmatrix},$$

with Γ conditionally negative definite. We have to assume $\lambda > \gamma_{12}/2$ and moreover we need the sign of

$$\det\begin{bmatrix} \lambda & \lambda - \gamma_{12} & \lambda - \gamma_{13} \\ \lambda - \gamma_{12} & \lambda & \lambda - \gamma_{23} \\ \lambda - \gamma_{13} & \lambda - \gamma_{23} & \lambda, \end{bmatrix} =$$

$$-2\gamma_{12}\gamma_{13}\gamma_{23} + \lambda\left(-\gamma_{12}^2 + 2\gamma_{12}\gamma_{13} - \gamma_{13}^2 + 2\gamma_{12}\gamma_{23} + 2\gamma_{13}\gamma_{23} - \gamma_{23}^2\right) \geq 0.$$

The solution of such algebraic inequalities is difficult in general, but we see that the admissible values of λ form a semi-infinite interval. In this and other similar cases, we can use a symbolic software such as Sage [13] to help with the algebra.

We now change our point of view to consider the same problem from a different angle. We can associate the variogram with the state-space description of the Gaussian vector. This is of use, for example, when a simulation is required. The following proposition is similar to Proposition 2.

Proposition 5 *1. The matrix Γ is a variogram matrix if, and only if, the matrix*

$$\Sigma_0 = -\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)' \Gamma \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) \tag{12}$$

is symmetric and positive definite. In such a case, the variogram of Σ_0 is Γ.

2. If $Y_0 \sim \mathrm{N}_n(0, \Sigma_0)$, then its variogram is Γ and it is supported by Span $(\mathbf{1})^\perp$.

Proof 1. If Γ is a variogram matrix, then the matrix Σ_0 of Eq. (12) is symmetric and positive definite. In fact, for a generic vector $\boldsymbol{\alpha}$ the vector $(I - \frac{1}{n}\mathbf{1}\mathbf{1}')\boldsymbol{\alpha}$ is orthogonal to $\mathbf{1}$, hence

$$\boldsymbol{\alpha}'\Sigma_0\boldsymbol{\alpha} = -\left(\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\boldsymbol{\alpha}\right)' \Gamma \left(\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\boldsymbol{\alpha}\right) \geq 0.$$

Viceversa, assume Σ_0 is a covariance matrix. As $\boldsymbol{e}_i - \boldsymbol{e}_j \in$ Span $(\mathbf{1})^\perp$, the variogram of Σ_0 has elements

$$(\boldsymbol{e}_i - \boldsymbol{e}_j)'\Sigma_0(\boldsymbol{e}_i - \boldsymbol{e}_j) =$$

$$(\boldsymbol{e}_i - \boldsymbol{e}_j)' \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)' (-\Gamma) \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) (\boldsymbol{e}_i - \boldsymbol{e}_j) =$$

$$-(\boldsymbol{e}_i - \boldsymbol{e}_j)'\Gamma(\boldsymbol{e}_i - \boldsymbol{e}_j) = -\gamma_{ii} - \gamma_{jj} + 2\gamma_{ij} = 2\gamma_{ij}.$$

2. As $\mathbf{1}'(\boldsymbol{e}_i - \boldsymbol{e}_j) = 0$, then $\mathbf{1}' \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)' (-\Gamma) \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) \mathbf{1} = 0$, hence the distribution of Y_0 is supported by the space Span $(\mathbf{1})^\perp$.

Remark 3 Let us derive some other equivalent expression for Σ_0. The h-th element of diag (Σ_0) is

$$\boldsymbol{e}_h'\Sigma_0\boldsymbol{e}_h = -\boldsymbol{e}_h' \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)' \Gamma \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) \boldsymbol{e}_h = -(\boldsymbol{e}_h - \frac{1}{n}\mathbf{1})'\Gamma(\boldsymbol{e}_h - \frac{1}{n}\mathbf{1})$$

hence

$$\text{diag}\,(\Sigma_0) = -\sum_h \boldsymbol{e}_h\boldsymbol{e}_h' \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)' \Gamma \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) \boldsymbol{e}_h\boldsymbol{e}_h'$$

and

$$\operatorname{diag}(\Sigma_0)\,\mathbf{1}\mathbf{1}' = -\sum_h \boldsymbol{e}_h\boldsymbol{e}_h'\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)'\Gamma\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)\boldsymbol{e}_h\mathbf{1}'$$

$$\mathbf{1}\mathbf{1}'\operatorname{diag}(\Sigma_0) = -\sum_h \mathbf{1}\boldsymbol{e}_h'\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)'\Gamma\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)\boldsymbol{e}_h\boldsymbol{e}_h'.$$

Let us compute the (i, j) element.

$$\boldsymbol{e}_i'\operatorname{diag}(\Sigma_0)\,\mathbf{1}\mathbf{1}'\boldsymbol{e}_j = -\boldsymbol{e}_i'\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)'\Gamma\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)\boldsymbol{e}_j$$

$$\boldsymbol{e}_i'\mathbf{1}\mathbf{1}'\operatorname{diag}(\Sigma_0)\,\boldsymbol{e}_j = -\boldsymbol{e}_j'\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)'\Gamma\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)\boldsymbol{e}_j.$$

The previous computation are of use in the analysis of the variogram of Σ_0, because in this case

$$\gamma = \frac{1}{2}(\operatorname{diag}(\Sigma_0)\,\mathbf{1}\mathbf{1}' + \mathbf{1}\mathbf{1}'\operatorname{diag}(\Sigma_0)) + \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)'\Gamma\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right).$$

In the following Proposition, we derive an additive decomposition of a generic Gaussian vector into a term whose variance is that obtained in Proposition 5 and a Gaussian vector proportional to the unit vector $\mathbf{1}$.

Proposition 6 *Let $Y \sim \mathrm{N}_n(\boldsymbol{\mu}, \Sigma)$ with variogram Γ. Let $Y_0 = \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)Y$ be the projection of Y onto $\operatorname{Span}(\mathbf{1})^{\perp}$ so that we can write $Y = Y_0 + \overline{Y}$, where each component of $\overline{Y}$ is the empirical mean $\frac{1}{n}\mathbf{1}'Y$.*

1. *The distribution of Y_0 is $\mathrm{N}_n(\boldsymbol{\mu}-\frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{\mu}, \Sigma_0)$, with $\Sigma_0 = -\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)\Gamma\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)$, that is, it depends on the mean and the variogram only.*
2. *The distribution of $\frac{1}{n}\mathbf{1}'Y$, conditionally to Y_0, is Gaussian with mean*

$$\frac{1}{n}\mathbf{1}'\boldsymbol{\mu} + \boldsymbol{l}'\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)(Y-\boldsymbol{\mu})$$

and variance

$$\frac{\sum_j \lambda_j}{n} - \frac{\sum_{i,j}\gamma_{ij}}{n^2} + \boldsymbol{l}'\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)\Gamma\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)\boldsymbol{l},$$

where $\boldsymbol{l}$ is a vector such that

$$\frac{1}{n}\mathbf{1}'\boldsymbol{\mu} + \boldsymbol{l}'\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}'\right)(Y-\boldsymbol{\mu}) = \mathbb{E}\left(\frac{1}{n}\mathbf{1}\mathbf{1}'Y\,\middle|\,Y_0\right).$$

Proof 1. The variance of Y_0 is

$$\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\Sigma\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) = \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\left(\frac{1}{2}(\Lambda\mathbf{1}\mathbf{1}' + \mathbf{1}\mathbf{1}'\Lambda) - \Gamma\right)\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) = -\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\Gamma\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right).$$

2. We have $\mathbb{E}\left(\frac{1}{n}\mathbf{1}'Y \middle| Y_0\right) = \frac{1}{n}\mathbf{1}'\boldsymbol{\mu} + \boldsymbol{l}'\left(Y_0 - \mathbb{E}\left(Y_0\right)\right)$, if the vector $\boldsymbol{l} \in \mathbb{R}^n$ is such that $\operatorname{Cov}\left(\frac{1}{n}\mathbf{1}'Y, Y_o\right) = \boldsymbol{l}' \operatorname{Var}\left(Y_o\right)$, that is,

$$\frac{1}{n}\mathbf{1}'\Sigma\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) = \boldsymbol{l}'\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\Sigma\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right),$$

or, in terms of $\Lambda = \operatorname{diag}\Sigma$ and the variogram Γ,

$$\frac{1}{2}\mathbf{1}'\Lambda\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) = \frac{1}{n}\mathbf{1}'\Gamma\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) - \boldsymbol{l}'\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\Gamma\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right),$$

The variance of $\frac{1}{n}\mathbf{1}'Y$ is

$$\frac{1}{n^2}\mathbf{1}'\Sigma\mathbf{1} = \frac{1}{n^2}\left(n\mathbf{1}'\Lambda\mathbf{1} - \mathbf{1}'\Gamma\mathbf{1}\right) = \frac{\sum_j \lambda_j}{n} - \frac{\sum_{i,j}\gamma_{ij}}{n^2},$$

and the variance of $\boldsymbol{l}'Y_0$ is

$$\boldsymbol{l}'\Sigma_0\boldsymbol{l} = -\boldsymbol{l}'\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\Gamma\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\boldsymbol{l}.$$

The conclusion follows from the conditioning formula for Gaussian vectors.

Remark 4 The previous proposition suggest an algorithm for the simulation when the variogram is given, by generating first the deviations from the general mean by using the covariance Σ_0, then from the conditional distribution of the general mean, given the deviations. It should be noted that in case of a stationary variance $\lambda = \lambda_j$, $j = 1, \ldots, n$ the vector $\boldsymbol{l}$ depends on Γ only.

Remark 5 Geostatistical applications of the Gaussian model parametrized with the variogram require the computation of the expression of the density of $Y \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, $\det\Sigma \neq 0$ and of the conditional expectation. This is not done here, but see some partial results in [3].

3 Stationarity

Let G be an additive topological locally compact group, e.g., $\mathbb{Z}$ or $\mathbb{R}$ with the ordinary sum $x + y$. A centered Gaussian random process $(Y(x))_{x\in G}$ is *stationary* if $\mathrm{Cov}\,(Y(x), Y(y)) = \mathrm{Cov}\,(Y(x-y), Y(0)) = C(x-y)$. The autocovariance function C is positive definite, that is, $\sum_{ij=0}^{n} \alpha_i\alpha_j C(x_i - x_j) \geq 0$, $n \in \mathbb{N}$, $x_1, \ldots, x_n \in G$, $\alpha \in \mathbb{R}^n$. The process is *intrinsically stationary* if $\mathrm{Var}\,(Y(x) - Y(y)) = \mathrm{Var}\,(Y(x-y) - Y(0)) = 2\gamma(x-y)$. The variogram function γ is conditionally negative definite, i.e., the matrix $\Gamma = [\gamma(x_i - x_j)]_{i,j=1}^{n}$, $n \in \mathbb{N}$, $x_1, \ldots, x_n \in G$, is conditionally negative definite, as in Proposition 1.

We plan to discuss, in a paper currently in progress, the existence of an intrinsically stationary process Y given a conditionally negative definite function and we want to characterise specific classes of variogram functions, e.g., those which are increasing (if an order is available) and bounded as $x \to \infty$. Increasing and bounded variograms are considered especially adapted to Geostatistics. In fact, D.G. Krige himself assumed that the variance of the difference between values measured in two locations is increasing with the distance between the locations, while the covariance vanishes. In the stationary case, these assumptions are still valid; therefore, we can use the results of the previous section, together with a further characterisation of variograms, which is based on the following theorem.

Proposition 7 [11, Theorem 6.1.8] *Let $\gamma\colon G$ and $f(0) \geq 0$. Then γ is conditionally negative definite if, and only if, for all finite sequence $x_1, \ldots, x_n$, the matrix $A = [\gamma(x_i - x_j) - \gamma(x_i) - \gamma(-x_j)]_{i,j=1}^{n}$ is negative definite.*

Proof If the matrix A is negative definite and $\sum_i \alpha_i = 0$, then

$$0 \geq \sum_{i,j=1}^{n} \alpha_i\alpha_j(\gamma(x_i - x_j) - \gamma(x_i) - \gamma(-x_j)) = \sum_{ij=1}^{n} \alpha_i\alpha_j\gamma(x_i - x_j)$$

Viceversa, from generic $x_1, \ldots, x_n$, $\alpha_1, \ldots, \alpha_n$, define $x_{n+1} = 0$ and $\alpha_{n+1} = -\sum_i \alpha_i$, then write the condition for conditional negativity.

Finally, in this setting one must take advantage of the harmonic representation of positive definite functions.

4 Conclusions and Future Developments

When dealing with Kriging meta-models, it is mandatory to provide a description of how the responses are correlated, since the goodness of the Kriging predictions in untried points strongly depends on the Gaussian random field. The correlation quantifies the smoothness of the response function and there are two approaches in the literature. The first one is the use of the spatial correlation function, SCF, (typical

of the Design and Analysis of Computer Experiments); the second one, proposed by Matheron, is based on the use of the variogram. In this paper, the equivalence between variogram and SPC is proved for stationary and intrinsically stationary processes. The use of the variogram is favourite because it does not require a parametric approach as the correlation estimation does.

Further developments to be published later and to be presented in forthcoming conferences, concern the use of the variogram for detecting technological signature in manufactured parts and a benchmark of different approaches (parametric and not parametric approach of the variogram and the Artificial Neuronal Networks) in the capability evaluation of the turbine features in order to maximise the performances (minimisation of the fuel consumption).

Acknowledgments G. Pistone is supported by the de Castro Statistics Initiative, Collegio Carlo Alberto, and is a member of GNAMPA–INdAM.

References

1. Amari, S., Nagaoka, H.: Methods of Information Geometry. American Mathematical Society. Providence (2000)
2. Chilès, J.P., Delfiner, P.: Geostatistics, 2nd edn. Wiley Series in Probability and Statistics. Wiley, Hoboken (2012)
3. Craparotta, G.: Metamodelli per il progetto di turbine di bassa pressione. Master's thesis, Politecnico di Torino (2014)
4. Cressie, N.A.C.: Statistics for spatial data. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley, New York (1993)
5. Gaetan, C., Guyon, X.: Spatial statistics and modeling. Springer Series in Statistics. Springer, New York (2010)
6. Gneiting, T., Sasvári, Z., Schlather, M.: Analogies and correspondences between variograms and covariance functions. Adv. Appl. Probab. **33**(3), 617–630 (2001)
7. Lauritzen, S.L.: Graphical models. The Clarendon Press, Oxford University Press, New York (1996)
8. Matheron, G.: Traité de géostatistique appliqué. **14** In: Mem. Bur. Rech. Geog. Minieres. Editions Technip (1962)
9. Pistone, G., Vicario, G.: Comparing and generating Latin Hypercube designs in Kriging models. ASTA Adv. Stat. Anal. **94**, 353–366 (2010)
10. Pistone, G., Vicario, G.: Kriging prediction from a circular grid: application to wafer diffusion. Appl. Stoch. Model. Bus. Ind. **29**(4), 350–361 (2013)
11. Sasvári, Z.: Positive definite and definitizable functions. Mathematical Topics, vol. 2. Akademie Verlag, Berlin (1994)
12. Skovgaard, L.T.: A Riemannian geometry of the multivariate normal model. Scand. J. Stat. **11**(4), 211–223 (1984)
13. Stein, W., et al.: Sage Mathematics Software (Version 5.9). The Sage Development Team. http://www.sagemath.org (2013)

Geographically Weighted Regression Analysis of Cardiovascular Diseases: Evidence from Canada Health Data

Anna Lina Sarra and Eugenia Nissi

Abstract

This paper aims to present an exploratory spatial analysis for ascertaining Canadian regional variations in the relationships between cardiovascular diseases prevalence and some well-established risk factors. Since the geographic variation in risk factors for cardiovascular diseases is too complex to be captured by a single set of regression coefficients, a local regression technique is employed. In particular, in this study, we make use of Geographically Weighted Regression (GWR) models with a ridge regression parameter to condense model complications related to the occurrences of local collinearity in the weighted explanatory variables. Local regression coefficients and associated statistics for both traditional GWR and GWR where a ridge regression parameter has been integrated are compared to evaluate their relative abilities in modelling the heterogeneous impact of risk factors on cardiovascular diseases across space.

1 Introduction

Cardiovascular diseases (CVDs) are one of the major worldwide health concerns, with severe implications for health and life expectancy. The latest available data from WHO indicate that CVDs account for 17.5 million deaths in the world, which

A.L. Sarra (✉) · E. Nissi
Department of Economics, University G. d'Annunzio of Chieti-Pescara, Chieti and Pescara, Italy
e-mail: asarra@unich.it

E. Nissi
e-mail: eugenia.nissi@unich.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_18

represents 30 % of all deceases. The highest percentage of global deaths from CVDs and their consequences occur in low and middle-income countries. Notwithstanding, CVDs remains the leading cause of premature death also in Europe and North America. CVDs, such as heart, stroke, hypertension, are the result of multi-interacting risk factors that can increase the likelihood for developing the disease. It is well established that the main contributors to the CVDs disability and deaths in populations are smoking, high level of alcohol consumption, inappropriate diet, life stress, sedentary lifestyle, high level of blood cholesterol, diabetes, high blood pressure (hypertension). CVDs and related conditions are unevenly distributed within regional territories. Understanding substantial geographic variations in the strength of the relationship between CVDs and associated risk factors can provide policymakers with important information to target predictors of CVDs at the local level, in an effort to reduce or eliminate heart disease risks and achieve an increase in life expectancy. When spatial heterogeneity is suspected to exist, a localized model should be more conveniently calibrated. Several statistical techniques have been developed to model nonstationarity of relations across space, mainly via some regression adaptations [1,4,6,15]. An effective way to address the spatial heterogeneity and accurately describe relationships among variables is represented by the Geographically Weighted Regression (GWR) [2,8]. Local specific results, achieved by GWR, may provide a more detailed perspective on underlying relationships, allowing refinements in the global specification. The present analysis has as its primary intention in assessing the extent to which the association between hypertension prevalence in the Canadian Health regions and some well-established risk factors vary spatially. We adopt the GWR to explore if the effects of some cardiovascular disease predictors are heterogeneous across space and hence vary from place to place. However, some caution should be used in interpreting the spatial patterns of local GWR coefficients. In comparison to a classical Ordinary Least Square (OLS) regression, GWR will produce higher correlation in the weighted explanatory variables, which increases the variances in the estimated regression coefficients and potentially invalidates conclusions about the relationships under study. For this reason, as an alternative to a standard GWR, in this study we also chose to use a GWR model where a ridge regression parameter has been incorporated [3] to reduce model complications arising from collinearity. The rest of the paper is organized as follows. In Sect. 2 we describe the methodological issues (e.g. the basics of GWR and GWR with a local ridge compensation). Section 3 considers the available data. The results of the local regression analysis are presented in Sect. 4. Finally, in Sect. 5 there are some concluding remarks.

2 GWR Modelling

The GWR is a local estimation procedure accounting for spatially changing relationships, deemed a complement tool to global modelling. The main appeal of method is that GWR relaxes the assumption in traditional OLS models that the relationships (regression coefficients) between dependent and independent variables being mod-

elled is constant across a study area, by allowing local rather than glob-al parameters to be estimated. A typical model of GWR can be written as

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i)x_{ik} + \varepsilon_i \tag{1}$$

where $\beta_k(u_i, v_i)$ $(k = 1, 2, \ldots.., M)$ are the regression coefficients for each location i and each variable k, (u_i, v_i) denotes the coordinates of i-th point in space and ε_i are the error terms. The local parameters $\beta_k(u_i, v_i)$ are estimated by weighted least-squares estimator, given by

$$\hat{\beta}_i = (\mathbf{X}^{'}\mathbf{W}((u_i, v_i)\mathbf{X}))^{-1}\mathbf{X}^{'}\mathbf{W}(i)\mathbf{Y} \tag{2}$$

In (2) the weight matrix $\mathbf{W}(u_i, v_i)$ is no longer constant but varies according to the location of point i. In GWR an observation is weighted in accordance with its proximity to regression point i: data points near to location (u_i, v_i) will be assigned higher weights in the model than data points farther away. Accordingly, an essential step in GWR modelling is the building of the weight matrix which involves the selection of the distance function and the definition of a finest bandwidth [9]. In this study we adopt the bi-square nearest neighbour formulation of the weighting function. The kernel shape is defined by the following equation which takes into account only the n^{th} nearest neighbours

$$w_{ij} = \begin{cases} [1 - (\frac{d_{ij}}{h_i})^2]^2 & if \quad d_{ij} < h_i \\ 0 & otherwise \end{cases} \tag{3}$$

where h_i is the n^{th} nearest neighbour distance from i.

This equation yields "spatially adaptive kernels". As a result, the calibration of the model involves also the choice of n, the number of data point to be included in the estimation of local parameters. Different methods are traditionally used to define the finest bandwidth value or the appropriate value of n. Among them, there are the AICc [14] and the cross-validation score (CV) procedure [5]. Here, we rely on the AICc method because it has the advantage of taking into account differences in model complexity, that is the varying degrees of freedom of models centred on different observations. One of the main advantages of the GWR-based technique regards the possibility to graphically display the spatial changes in the magnitude of the parameter estimates across the study region, indicating the locally changing influence of a predictor on the dependent variable. Mapping local variation and looking at the local performance of predictor variables might also shed light on the question of model specification and/or support the identification of missing important predictors. In addition to the local parameter estimates, GWR technique enables to plot local statistics (pseudo t-values, pseudo R^2, local residuals) useful in exploring and interpreting spatial nonstationarity. However, there is a growing body of literature that has argued some problems with using GWR for statistical inference on regression relationships. Recent studies show that this technique is not able to estimate the regression coefficients accurately (see, among others [10,16,19]). One important issue is related to the occurrence of local collinearity in weighted explanatory variables. Collinearity can be an important issue as it increases the variances

in estimated regression coefficients and potentially invalidates conclusions about the estimated relationships [17]. As pointed out in many works [17,18], the lack of collinearity in global regression model is not a guarantee for the absence of that trouble in GWR models. The adverse effects of collinearity in the predictors of a linear model are more pronounced in the localized framework of GWR. In that context, the smaller samples used to calibrate each regression along with the potential spatially heterogeneous correlation structures of the data exacerbate the detrimental influence of collinearity among the predictor variables. Alternative methods, proven to outperform standard GWR in estimating true regression effect over space, have been proposed [7,17]. One way to constrain and stabilize the regression coefficients is to calibrate a ridge regression [11,12]. The initial adoption of the ridge regression in the context of GWR is due to [17]. Later [3] introduce a GWR with a locally compensated ridge term (GWR LCR). A distinguishing feature of that approach is basically the possibility to fit local ridge regressions where the ridge parameter is allowed to vary across space. Besides, to achieve an indepth understanding of where the undesirable influences of local collinearity might occur, the GWR LCR calibrates the ridge regressions only at locations where some diagnostics (for instance the condition number) is above a user's specified threshold. The GWR LCR relies on the existence of a link between the definition of the condition number for the design matrix and the ridge parameter. Given that for a generic square symmetrical matrix $\mathbf{M}$, the condition number is defined to be the ratio of the largest (e_1) to the smallest (e_m) eigenvalues of that matrix, it follows that in a ridge-adjusted matrix this measure will be $e_1 + \lambda / e_m + \lambda$. Accordingly, it will be possible to specify the ridge parameter such that it will be below a chosen threshold. In a geographically weighted regression there will be a condition number associated with every point in the study area at which GWR coefficients are estimated. Formally, the estimator for this locally compensated ridge regression model is

$$\hat{\beta}_i = (\mathbf{X}^{'}\mathbf{W}(u_i, v_i)\mathbf{X} + \lambda \mathbf{I}(u_i, v_i)^{-1}\mathbf{X}^{'}(u_i, v_i)\mathbf{Y} \tag{4}$$

where $\lambda \mathbf{I}((u_i, v_i))$ is the locally compensated value of the ridge term λ at location i. The effect to introduce a ridge term in Eq. (4) is that to increase the difference between the diagonal element of the design matrix and the off-diagonal elements. Technically, the addition of a displacement to the leading diagonal of the geographically weighted cross-product matrix allows to overcome the difficulty to numerically inverted this matrix since it has a denominator close to zero in presence of collinearity. As in the standard GWR regression, the cross-validation procedure can be adopted to estimate the optimal bandwidth. More technical details underlying GWR LCR are described in [3].

3 Data: The Canadian Community Health Survey

Data employed in this study arise from the Canadian Community Health Survey (CCHS) which is a series of national cross-sectional surveys that have been carried

Table 1 Definitions of explanatory variables (*Source* Statistics Canada)

Variable name	Variable description
Dietary practices (DIETS)	Population aged 12 and over, by the average number of times per day that they consume fruit and vegetables
Weight (BMI)	Overweight (BMI of 25.0–29.9), by number of persons 18 and over, excluding pregnant women, both sexes
Physical activity (PHYSIN)	Household population 12 and over, physically inactive
Smoking status (SMOK)	Household population 12 and over, daily smoker
Diabetes (DIAB)	Population aged 12 and over who report that they have been diagnosed by a health professional as having diabetes

out by Statistics Canada since 2001. This survey was designed to produce regular and timely cross-sectional estimates of health determinants, health status and health system utilization at provincial and sub-provincial levels (health region or combination of health regions). The base population for each health region provides the option of normalizing the data. In all the empirical models of this study the dependent variable of interest is percent of people aged 12 and over who report that they have been diagnosed by a health professional as having high blood pressure (HBP). Albeit the exact causes of high blood pressure are usually unknown, there are several factors that have been highly associated with this condition. In the current study is acknowledged that life style behaviours, specifically weight loss, regular increase in physical activity, avoiding tobacco smoking and a healthy diet, rich in the consume of fruits and vegetables per day, can effectively lower blood pressure. In addition research from the WHO [20] highlights the importance of raised diabetes as a risk factor for HBP. All the employed explanatory variables are summarized in Table 1 and are drawn from 2011 CCHS. Due to limited availability of the same data in all Canadian Health Regions, our analysis consider only 69 of them.

4 Results

We first carried out an OLS regression and the results are reported in Table 2.

The hypothesized relationship between the HBP and the explanatory variables are supported quite well by the data. Indeed, most of the predictor variables, except those related to dietary practices (Consume of Vegetables and Fruits) and Physical Activity, are statistically significant according to their t-values at the significance level of $\alpha = 0.05$. The adjusted coefficient of determination R^2 is 0.49. The residuals for the OLS results clear exhibit spatial patterning.

Visual inspection of residuals, mapped in Fig. 1, reveals that the model tends to underestimate HBP in the Northwest and Southwest of the region under study, while overestimate the outcome of interest in the North and South of the Health Regions

Table 2 OLS parameter summaries

	Coefficient	SE	t-statistic	p value	Sig.
Intercept	−1.06	6.24	−0.17	0.87	
Diabetes (DIAB)	0.57	0.23	2.42	0.02	*
Current smokers (SMOK)	−0.20	0.06	−3.52	0.00	***
Body mass index (OVERBMI)	0.35	0.08	4.59	0.00	***
Dietary practices (DIETS)	0.07	0.05	1.29	0.20	
Physical inactivity (PHYSIN)	−0.04	0.08	−0.55	0.58	

[a]Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

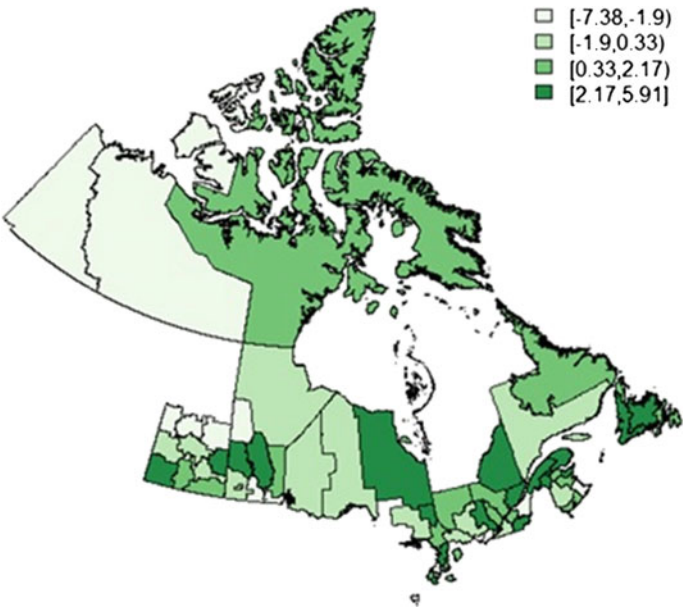


Fig. 1 Residuals for OLS estimates

Table 3 Goodness-of-fit test for improvement in model fit of GWR over global model (OLS)

Source	SS	DF	MS	F
OLS residuals	480.2	5		
GWR improvement	107.7	9.74	11.0	
GWR residuals	372.5	54.2	6.86	1.61
Diagnostic information	OLS	GWR		
R^2	0.53			
Adjusted R^2	0.49	0.63		

[a]SS = sum of squares; DF = degree of freedom; MS = residual mean square; F = F-statistic

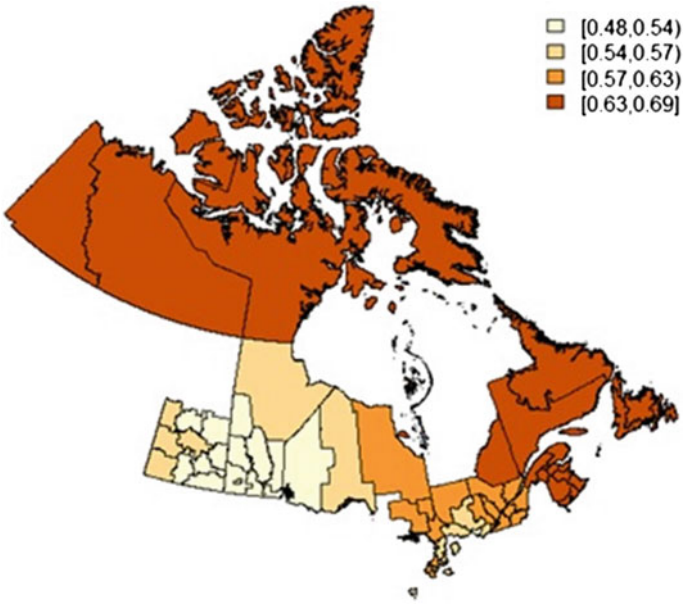


Fig. 2 Local R-squared values for GWR

included in our analysis. The GWR estimation improves the overall explanatory power of the HBP model by raising the adjusted R^2 values from 0.49 to 0.63 (Table 3), indicating that are spatially circumstances influencing the percent of people with high blood pressure. Local R^2 values range from 0.48 to 0.69, with the highest R^2 values located in the North and in the South-Eastern of the area under study (Fig. 2).

For each coefficient of interest, it is possible to test the null hypothesis of same variation over space in the estimated local values of the parameter against the alternative hypothesis that local parameter estimates exhibits spatial variability. A Monte

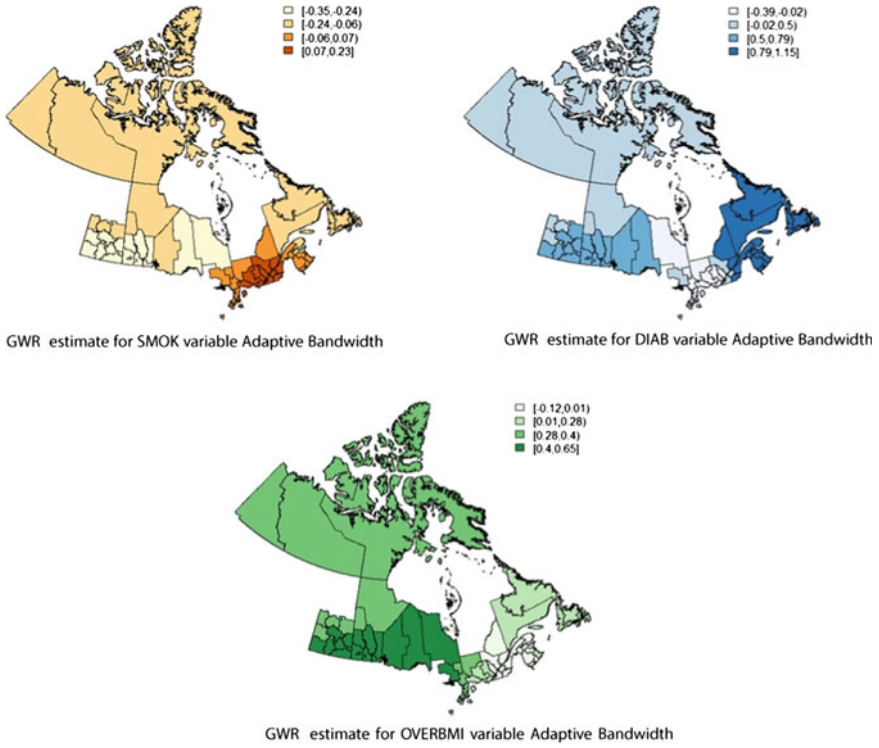


Fig. 3 Spatial mapping of the coefficients from GWR modelling

Carlo approach is traditionally used to test for nonstationarity in individual parameters [9, 13].

In this case study, the Monte Carlo test results for spatially variability of parameters suggested that the association between HBP and almost the predictors included in the GWR model (OVERBMI, SMOK and DIAB) is nonstationary across space. The spatial variation of each explanatory variable can be better understood by looking at Fig. 3, displaying the local parameter estimates. More pronounced and diverse spatial nonstationarity is evident in the effect of Diabetes on HBP. The relationship is positive and strongest in the South whereas becomes weaker in the North. Besides, the GWR findings indicate that the large coefficients for OVERBMI and SMOK variables are also concentrated in the Southern Health regions. Further enlightening information is provided by maps of t-statistics, displayed in Fig. 4, obtained dividing each local estimate of the regression coefficients by its corresponding local standard error. Actually, there are some theoretical difficulties in interpreting these t-statistics values. As pointed out by Waller et al. 2007, the pseudo t-values should be used in an exploratory fashion since they do not represent a formal statistical estimate, resulting from a relatively ad hoc inference.Reference 'Waller et al. (2007)' is cited in text but not provided in the reference list. Please provide references in the list or delete these citations.

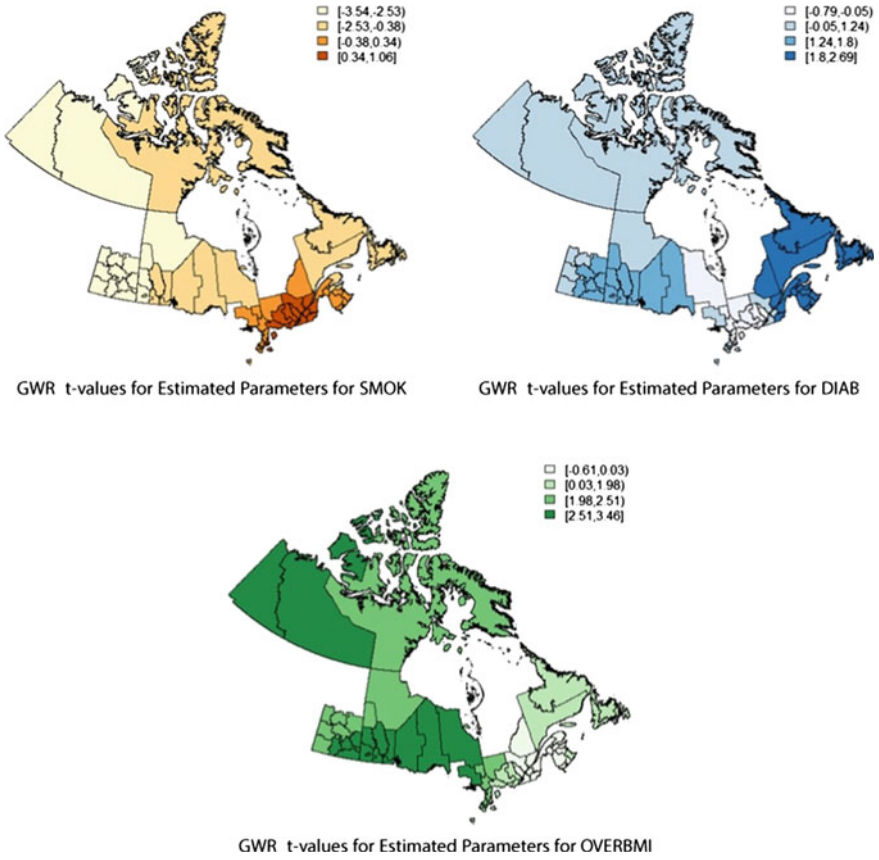


Fig. 4 Spatial mapping of t-values the coefficients from GWR modelling

To continue the analysis we also investigated the issue of collinearity using the approach of locally compensated ridge GWR, described in Sect. 2. To this end, we compare the coefficient estimates for the unadjusted basic GWR with those from a GWR LCR. As noted previously, a standard measure of the effects of collinearity is the condition number of the matrix ($\mathbf{X}'\mathbf{X}$), defined to be the ratio of the largest to the smallest eigenvalues of that matrix. Condition numbers above around 30 are regarded as potentially problematic, suggesting that the associated results may be unreliable. This diagnostic can be easily adapted to the local context by replacing the global design matrix ($\mathbf{X}'\mathbf{X}$) with the local weighted cross-products matrices ($\mathbf{X}'\mathbf{W}\mathbf{X}$). The local design matrix condition numbers are found at the same spatial scale as each local regression of the GWR model and can thus be mapped.

The local condition numbers for a basic GWR range from 50.51 to 105.79. On inspection of the map, shown in Fig. 5a, it is evident that they are large everywhere. Accordingly, we proceed to a more locally focused analysis. By calibrating a GWR

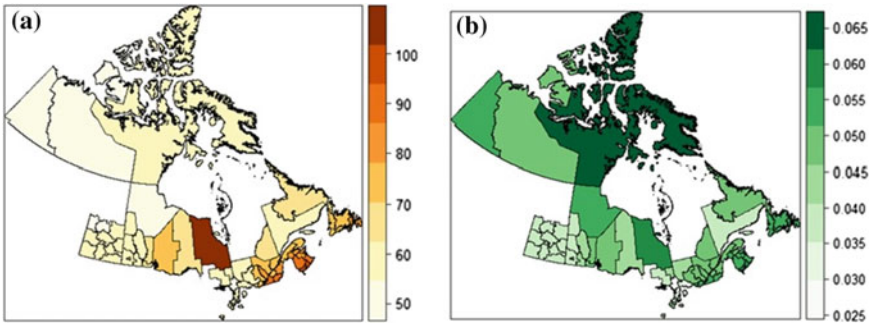


Fig. 5 Local condition numbers from basic GWR (**a**) and local ridge terms (**b**)

LCR, the local compensation, that is the ridge adjustment to each local weighted cross-products matrices, only takes place at locations displaying condition numbers higher than a given threshold. In this case we specify a threshold of 30 for the condition numbers. As a result, for this data set, a local ridge term will be found at all locations. Mapping the corresponding values it is possible to identify the areas where the major adjustments are necessary. The spatial pattern displayed in Fig. 5b suggest that the greatest adjustments were required in Northern health districts.

After performing different adjustments in relation to the varying levels of collinearity among the predictors, it would result that the larger coefficients for the basic GWR model are reduced in magnitude and the smaller coefficients are raised. The comparison of basic and locally compensated GWR estimates for Diabetes, Smoking status and overweight predictors, are given in Fig. 6. The relationship is nonlinear and a loess fit is shown in the plot.

5 Conclusions

Following the trend of the last three decades of exploring the geographical aspects of population health, this paper has presented a spatial analysis to investigate the association between cardiovascular disease (hypertension) and a set of related modifiable risk factors, over some Canadian Health districts. We have assumed that the relationship under study is not universal across the study area and addressed the issue of nonstationarity via GWR. That approach is currently a well-established technique, especially designed to allow the regression parameters and the strength of the relationship to vary over space. After accounting for place, we have found that hypertension-related contributors association fluctuates from negative to positive as a function of geographical location, confirming the spatial heterogeneity of the predictors that do not always have the same impact. Our study has also explicitly taken

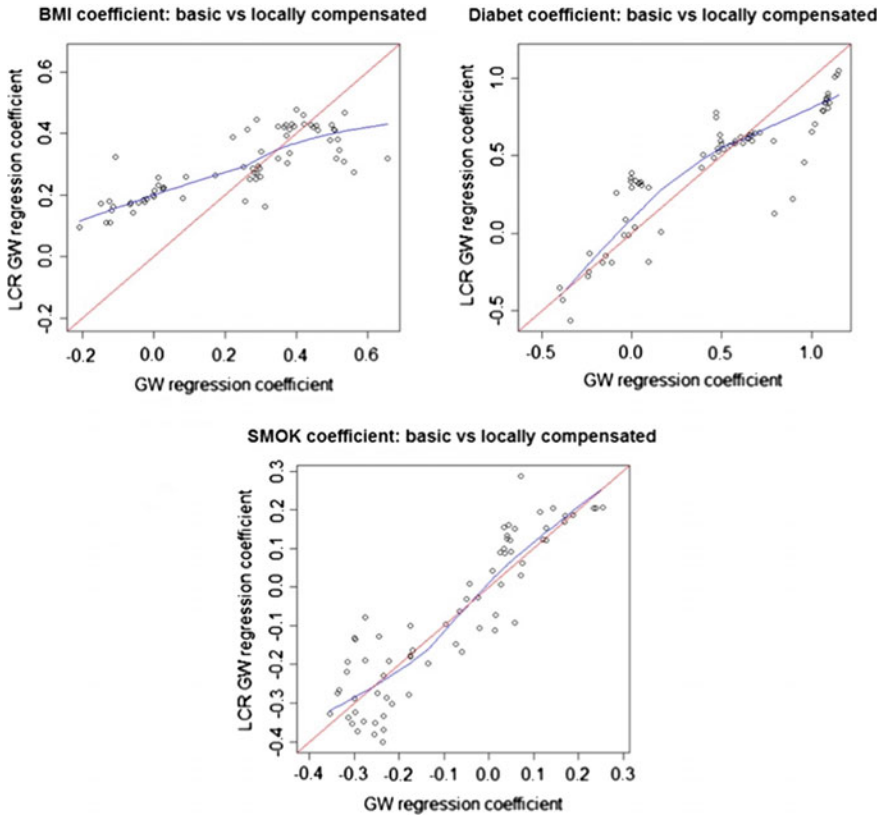


Fig. 6 Comparison of coefficient estimates

into account an important issue in the GWR modelling related to the occurrence of local collinearity in weighted explanatory variables that may produce unreliable estimates and misleading inference. To measure the degree of collinearity existing in our dataset and deal with local collinearity problems, we have adopted a GWR with a locally compensated ridge term. This local compensation approach has revealed that a worryingly collinearity within predictors variables exists over all the health areas analyzed and thus the related adjustments were performed everywhere. The main findings of our analysis suggest that a more robust computational framework for spatially varying coefficients, as that implied by the GWR LCR model, would be greatly beneficial for researchers interested in estimating true regression effects over space. In this specific case study, thank to the identification of reliable spatial variations between CVDs and their main predictors, the health institutions will be knowledgeable where resources for management and prevention of CVDs risk

factors should be allocated in an effort to increase population's life expectancy. It is worth noting that even if the paper focuses on a specific application, the method discussed herein can be deemed general and it results appropriate when the interest is in localized analyses and need-based interventions.

References

1. Assunc$\tilde{a}$o, A.: Space varying coefficient models for small area data. Environmetrics **14**, 453–473 (2003)
2. Brunsdon, C.F., Fotheringham, A.S., Charlton, M.E.: Geographically weighted regression: a method for exploring spatial non stationarity. Geogr. Anal. **28**, 281–298 (1996)
3. Brunsdon, C., Charlton, M., Harris, P.: Living with collinearity in local regression models. In: Proceedings of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Brasil (2012)
4. Casetti, E.: Generating models by expansion methods: application to geographical research. Geogr. Anal. **4**, 81–91 (1972)
5. Cleveland, W.: Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. **74**, 829–836 (1979)
6. Duncan, C., Jones, K.: Using multilevel models to model heterogeneity: potential and pitfalls. Geogr. Anal. **32**(4), 280–305 (2000)
7. Finley, A.O.: Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. Methods Ecol. Evol. **2**, 143–154 (2011)
8. Fotheringham, A.S., Charlton, M.E., Brunsdon, C.F.: Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. Environ. Plan. A **30**, 1905–1927 (1998)
9. Fotheringham, A.S., Brunsdon, C., Charlton, M.: Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Chichester (2002)
10. Griffith, D.A.: Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). Environ. Plan. A **40**, 2751–2769 (2008)
11. Hoerl, A.E.: Application of ridge analysis to regression problems. Chem. Eng. Prog. **58**, 54–59 (1962)
12. Hoerl, A.E., Kennard, R.W.: Ridge regression. Applications to non-orthogonal problems. Technometrics **12**, 69–82 (1970)
13. Hope, A.C.A.: A simplified Monte Carlo significance test procedure. J. R. Stat. Soc. B. **51**, 582–589 (1968)
14. Hurvich, C.M., Simonoff, J.S., Tsai, C.L.: Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J. R. Stat. Soc. B **60**, 271–293 (1998)
15. Jones, J.P., Casetti, E., Jones, J.P., Casetti, E.: Applications of the Expansion Methods. Routledge, London (1992)
16. Pàez, A., Farber, S., Wheeler, D.: SA simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. Environ. Plan. A **43**, 2992–3010 (2011)
17. Wheeler, D.C.: Diagnostic tools and a remedial method for collinearity in geographically weighted regression. Environ. Plan. A **39**, 2464–2481 (2007)

18. Wheeler, D.C., Tiefelsdorf, M.: Multicollinearity and correlation among local regression coefficients in geographically weighted regression. J. Geogr. Syst. **7**, 161–187 (2005)
19. Wheeler, D.C., Calder, C.A: An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. J. Geogr. Syst. **9**, 145–166 (2007)
20. World Health Organization: The Atlas of Heart Disease and Stroke (2012)

Pseudo-Likelihoods for Bayesian Inference

Laura Ventura and Walter Racugno

Abstract

The interplay between Bayesian and frequentist inference can play a remarkable role in order to address some theoretical and computational drawbacks, due to the complexity or misspecification of the model, or to the presence of many nuisance parameters. In this respect, in this paper we review the properties and applications of the so-called pseudo-posterior distributions, i.e., posterior distributions derived from the combination of a pseudo-likelihood function with suitable prior information. In particular, we illustrate the various notions of pseudo-likelihood highlighting their use in the Bayesian framework. Moreover, we show the simple but effective application of pseudo-posterior distributions in three challenging examples.

1 Introduction

In the presence of models with complicated dependence structures, of multidimensional nuisance parameters, or of model misspecifications, both frequentist and Bayesian inference may encounter some theoretical and computational difficulties. Indeed, in these situations the original likelihood function may be intractable or computationally cumbersome. In order to take into proper account of such difficulties,

L. Ventura (✉)
Department of Statistics, University of Padova, Padova, Italy
e-mail: ventura@stat.unipd.it

W. Racugno
Department of Mathematics and Informatics, University of Cagliari, Cagliari, Italy
e-mail: racugno@unica.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_19

it is possible to consider surrogates of the original likelihood, which produce the wide class of the so-called *pseudo-likelihoods*; see, for instance, [55, Chap. 4], [71, Chaps. 8 and 9], and [76], and references therein.

The aim of this paper is to review the properties and to illustrate some applications of the so-called pseudo-posterior distributions, i.e., distributions derived from the combination of a pseudo-likelihood function with suitable prior information. It is a Bayesian non-orthodox procedure widely used in the recent statistical literature and theoretically motivated in several papers; see, among others [4,11,12,17,19–21,30,34,36,40,46,51,58,60,63,67–69,73,77–79,81], and references therein.

The outline of the paper is as follows. Section 2 gives a brief review on pseudo-likelihood functions. Section 3 introduces the notion of pseudo-posterior distribution, discusses the choice of the prior and the validation of a pseudo-posterior distribution, also through first and higher-order asymptotic results. In Sect. 4 we illustrate the calculation of pseudo-posterior distributions using a one-way random effects model with heteroscedastic error variances, the Cox proportional hazards model, and a multilevel probit model. Finally, some concluding remarks close the paper.

2 Notion of Pseudo-Likelihood

Let $y = (y_1, \ldots, y_n)$ be a random sample of size n from a statistical model with parameter space Θ, not necessarily finite-dimensional. Let $\tau = \tau(\theta)$, with $\tau \in T \subseteq \mathbb{R}^k$, $k \geq 1$, be the parameter of interest. The more complex is the component complementary to τ in θ, then the more useful is the possibility of basing inference on a likelihood function which depends on τ only.

Let us denote with $L_{ps}(\tau) = L_{ps}(\tau; y)$ a pseudo-likelihood function for τ, that is a function of the data y which depends only on the parameter of interest τ and which behaves, in some respects, as it were a genuine likelihood. This means that, under mild regularity conditions, $L_{ps}(\tau)$ has unbiased score function, the pseudo-maximum likelihood estimator $\hat{\tau}_{ps}$ is consistent and asymptotically normal, and the pseudo-likelihood ratio test $W_{ps}(\tau) = 2(\ell_{ps}(\hat{\tau}_{ps}) - \ell_{ps}(\tau))$, with $\ell_{ps}(\tau) = \log L_{ps}(\tau)$, has null asymptotic χ^2_k distribution. Some well-known examples of pseudo-likelihood functions are the marginal, the conditional, the profile, the approximate conditional, the modified profile, the integrated, the partial, the quasi, the empirical, the weighted, the composite and the pairwise likelihood. For reviews on pseudo-likelihood functions see, e.g., [55, Chap. 4], [71, Chaps. 8 and 9], and [76], and references therein.

There are several reasons for introducing a pseudo-likelihood function for inference on τ. Here we propose a possible taxonomy of pseudo-likelihoods based on three main classes.

1. Elimination of nuisance parameters. Consider a parametric model with density function $p(y; \theta)$, $\theta \in \Theta \subseteq \mathbb{R}^p$, $p > 1$, and write $\theta = (\tau, \lambda)$, where the nuisance parameter λ is of dimension $p - k$. For inference on τ, pseudo-likelihoods based on a statistical model defined as a reduction of the original model are the *marginal* and the

conditional likelihoods [71, see Chap. 8]. However, they are available essentially only in exponential and in group families. Outside of these cases, one simple and general way of obtaining a pseudo-likelihood for τ is to replace the nuisance parameter λ with its maximum likelihood estimate (MLE) for fixed τ, i.e., $\hat{\lambda}_\tau$, in the original likelihood $L(\tau, \lambda)$. The corresponding function $L_p(\tau) = L(\tau, \hat{\lambda}_\tau)$ is the well-known *profile likelihood*. It is not a genuine likelihood and its behavior may not be entirely satisfactory, especially when the dimension of λ is large. Various modifications of $L_p(\tau)$ have been proposed, starting from the *approximate conditional likelihood* of [24], which is based on the choice of an orthogonal parameterization, to the various proposals of *modified profile likelihoods*, which require notions about higher-order asymptotic methods (see [71, Chap. 9]). All the available modifications of the profile likelihood are equivalent to the second order and share the common feature of reducing the score bias to $O(n^{-1})$ (see, e.g., [56]). A further approach that can be applied generally for the elimination of nuisance parameters is to average the likelihood function $L(\tau, \lambda)$ with respect to a "weight" function $\pi(\lambda)$ on λ, in order to define the *integrated likelihood function* $L_I(\tau) = \int L(\tau, \lambda)\, \pi(\lambda)\, d\lambda$ (see [71, Chap. 8], [10]).

2. Semi or nonparametric models. The *quasi-likelihood* (see [2,6,8,48]) is a pseudo-likelihood function associated to a semi parametric model specified in terms of first (and sometimes second) order moments of a particular unbiased estimating function. Instead, the *empirical likelihood* [54] was introduced to deal with inference problems on k-dimensional smooth functionals in nonparametric models. The study of these pseudo-likelihoods, when derived from M-estimators, has been investigated in [1,3,54]. When robustness with respect to influential observations or to model misspecifications is of interest, also the *weighted likelihood* can be considered (see, e.g., [38,47]), which is a pseudo-likelihood defined through a set of weights which are supposed to opportunely down-weight likelihood single term components.

3. Complex models. The class of *composite likelihoods* (see [76], and references therein) is useful when the fully specified likelihood is computationally cumbersome as well as when a fully specified model is out of reach. This class contains the ordinary likelihood, as well as many other interesting alternatives, such as the *Besag pseudo-likelihood* [13], the *m-order likelihood* for stationary processes [5], the *approximate likelihood* of [74], and the *composite marginal likelihood* and the *pairwise likelihood* [26], constructed from marginal densities. Also the *partial likelihood* [22,23], introduced for inference about the regression coefficients in the proportional hazards model, may be considered a member of this class.

Finally, we remark that since the 1970s numerous other pseudo-likelihoods have been considered. Some of these are: the pseudo-likelihood of [35], where nuisance parameters are eliminated by means of a simple plug-in estimate; the *bootstrap likelihood* [28,29], which is in the spirit of empirical likelihood; the *dual likelihood* [53], which associates a likelihood to a martingale estimating equation; the *projected likelihood* [49,82] for semi parametric models; the *penalized likelihood* [25,37] for an infinite-dimensional parameter of interest such as a density or a regression function;

the various instances of *predictive likelihood* [14,16]; the *h-likelihood* [44,45,50], that is a hierarchical likelihood, for inferences from random effect models.

3 Pseudo-Posterior Distributions

Assuming a prior distribution $\pi(\tau)$ on τ and treating $L_{ps}(\tau)$ as an ordinary likelihood, from a purely formal expression of Bayes' theorem we obtain

$$\pi_{ps}(\tau|y) \propto \pi(\tau)\ L_{ps}(\tau)\ . \tag{1}$$

The posterior distribution $\pi_{ps}(\tau|y)$ is obtained "miming" the Bayesian procedure and thus is called *pseudo-posterior*. In general, Bayesian inferential procedures based on pseudo-likelihoods are called *hybrid*, or *quasi* or *pseudo* Bayesian methods.

When basing inference on τ on the pseudo-posterior distribution $\pi_{ps}(\tau|y)$, three issues need to be addressed

(a) the choice of the suitable pseudo-likelihood $L_{ps}(\tau)$;
(b) the choice of the prior $\pi(\tau)$;
(c) the validation of inference based on $\pi_{ps}(\tau|y)$.

Section 3.1 focuses on the choice of the pseudo-likelihood to be used in (1), which depends on the model and the objectives of the analysis. Section 3.2 reviews the results on the choice of the prior. Finally, Sect. 3.3 discusses the validation of a pseudo-posterior distribution, both numerically and through asymptotic results.

3.1 Areas of Application of Pseudo-Posterior Distributions

Although (1) cannot always be considered as orthodox in a Bayesian setting, the use of alternative likelihoods is nowadays widely shared, and several papers focus on the Bayesian application of some well-known pseudo-likelihoods. Of course, the choice of the pseudo-likelihood to be used in (1) depends on the objectives of the analysis. A possible classification of the main areas of applications of the pseudo-posterior $\pi_{ps}(\tau|y)$ may be based on the following five classes.

Elimination of nuisance parameters. When $\theta = (\tau, \lambda)$ and only inference on τ is of interest, the marginal, the conditional, the modified profile, and the approximate conditional likelihoods can be used in (1). Note that the use of these pseudo-likelihoods in $\pi_{ps}(\tau|y)$ has the advantages of avoiding the elicitation on the nuisance parameter λ and of the computation of a multidimensional integral necessary to compute the marginal posterior distribution for τ. Moreover, these pseudo-likelihood functions $L_{ps}(\tau)$ have an orthodox Bayesian interpretation. This means that they are equivalent to a suitable integrated likelihood, of the form $L_I(\tau) = \int L(\tau, \lambda)\, \pi(\lambda|\tau)\, d\lambda$, for a specific conditional prior $\pi(\lambda|\tau)$ (see, e.g., [57,70]). As a further remark, note

that the pseudo-posterior distribution $\pi_{ps}(\tau|y)$ is a genuine posterior distribution when using in (1) the modified profile likelihood with the corresponding matching prior (see [77,81]) or in non-normal regression-scale models, in which there is no loss of information about τ when using a pseudo-posterior distribution derived from a marginal likelihood (see [60]). For Bayesian applications of the marginal, the conditional, the modified profile, and of the approximate conditional likelihoods see, among others [10–12,17,19,20,32,34,51,60,64,70,77,79–81], and references therein.

Semi or nonparametric models. When dealing with semi parametric or nonparametric statistical models, for Bayesian inference on τ the quasi and the empirical likelihoods can be used. Note that the use of these pseudo-likelihoods in $\pi_{ps}(\tau|y)$ has the advantages of requiring the elicitation of the prior only on the parameter of interest τ. For applications of these pseudo-likelihoods for Bayesian inference see [42,46,60,68,78], and references therein.

Robustness. When robustness with respect to outliers, influential observations or model misspecifications is required, the quasi, the empirical and the weighted likelihoods can be used to obtain resistant pseudo-posterior distributions. Indeed, the occurrence of anomalous values can seriously alter the shape of the ordinary likelihood function and then lead to ordinary posterior distributions far from those one would obtain without these data inadequacies, as illustrated in [4,36,78].

Complex models. The composite and pairwise likelihoods deal with complex statistical models, for which the ordinary likelihood and thus the ordinary posterior distribution are impractical to compute or even analytically unknown. The use of these pseudo-likelihood in Bayesian inference has been discussed in [58,63,65,67,73].

Proportional hazards model. In the Bayesian framework, the use of the partial likelihood to derive a posterior distribution on the regression parameters of the Cox model has the advantage of avoiding the specification of a prior process on the unknown baseline cumulative hazard function. For the use of this pseudo-likelihood in Bayesian inference, see, among others [21,39,40,67,69].

3.2 Choice of the Prior

The choice of the prior distribution on τ in (1) involves the same problems typical of the standard Bayesian perspective. In particular, this occurs both when the elicitation of a proper prior distribution is required and when using default prior distributions that are often improper. For instance, the choice of parametric priors in $\pi_{ps}(\tau|y)$ has been considered in several papers (see, e.g., [4,36,40,42,58,60,67,73]).

Non-informative priors have been considered by [21,58,60]. Ventura et al. [78] discuss how to modify the Jeffreys' prior to yield a default prior for τ to be used with a general pseudo-likelihood $L_{ps}(\tau)$. It is shown that the Jeffreys-type prior for τ associated to $L_{ps}(\tau)$ is given by

$$\pi^{J}_{ps}(\tau) \propto \sqrt{|i_{ps}(\tau)|}\,, \tag{2}$$

where $i_{ps}(\tau)$ is the pseudo-expected information matrix, i.e.,

$$i_{ps}(\tau) = E(-\partial^2 \ell_{ps}(\tau)/\partial\tau\partial\tau^{\top}) .$$

This means that a parametrization invariant prior distribution for τ, derived from a pseudo-likelihood function, is proportional to the square root of the determinant of the pseudo-expected information.

The other prominent studied default priors are the matching priors, designed to produce Bayesian credible sets which are optimal frequentist confidence sets in a certain asymptotic sense (see, e.g., [27]). The use of matching priors has been widely discussed in (1) with $L_{ps}(\tau)$ denoting a marginal, conditional or modified profile likelihood for a scalar parameter of interest τ; see, e.g., [17,19,51,61,64,77,79–81]. For instance, when using the modified profile likelihood, the corresponding matching prior is (see [77]),

$$\pi_{mp}(\tau) \propto i_{\tau\tau.\lambda}(\tau, \hat{\lambda}_{\tau})^{1/2} , \tag{3}$$

with $i_{\tau\tau.\lambda}(\tau,\lambda) = i_{\tau\tau}(\tau,\lambda) - i_{\tau\lambda}(\tau,\lambda) i_{\lambda\lambda}(\tau,\lambda)^{-1} i_{\lambda\tau}(\tau,\lambda)$ partial information, and $i_{\tau\tau}(\tau,\lambda)$, $i_{\tau\lambda}(\tau,\lambda)$, $i_{\lambda\lambda}(\tau,\lambda)$ and $i_{\lambda\tau}(\tau,\lambda)$ blocks of the expected Fisher information from the genuine likelihood $L(\tau,\lambda)$.

3.3 Validation of the Pseudo-Posterior Distribution

The pseudo-posterior distribution $\pi_{ps}(\tau|y)$ calls for its validation for Bayesian inference. At the current state, a general finite-sample theory for pseudo-posterior distributions is not available, and every single problem has to be examined.

For the pseudo-posterior distributions listed in Sect. 3.1, the validation may be based on asymptotic results. In particular, paralleling the results for the full posterior distribution and under standard regularity conditions, it can be shown that (see [36, 42,58])

$$\pi_{ps}(\tau|y) \dot{\sim} N_k\left(\hat{\tau}_{ps}, j_{ps}(\hat{\tau}_{ps})^{-1}\right) , \tag{4}$$

where $j_{ps}(\hat{\tau}_{ps})$ is the pseudo-observed information evaluated at the pseudo-MLE. An asymptotically equivalent normal approximation is $\pi_{ps}(\tau|y) \dot{\sim} N_k\left(\tilde{\tau}_{ps}, \tilde{j}_{ps}(\tilde{\tau}_{ps})^{-1}\right)$, where $\tilde{\tau}_{ps}$ is the pseudo-posterior mode and $\tilde{j}_{ps}(\tilde{\tau}_{ps}) = -(\partial \log L_{ps}(\tau))/(\partial\tau\partial\tau^{\top})|_{\tau=\tilde{\tau}_{ps}}$. Moreover, paralleling results for the full posterior distribution, also a higher-order tail area approximation can be derived for a scalar parameter of interest τ (see [67]). In particular, it holds

$$\int_{\tau_0}^{\infty} \pi_{ps}(\tau|y)\, d\tau \doteq \Phi(r^*_{ps}(\tau_0)) , \tag{5}$$

where $\Phi(\cdot)$ is the standard normal distribution function and

$$r^*_{ps}(\tau) = r_{ps}(\tau) + r_{ps}(\tau)^{-1} \log b(r_{ps}(\tau)) ,$$

with

$$r_{ps}(\tau) = sign(\hat{\tau}_{ps} - \tau)[2(\ell_{ps}(\hat{\tau}_{ps}) - \ell_{ps}(\tau))]^{1/2}$$

pseudo-signed likelihood root and

$$b(r_{ps}(\tau)) = j_{ps}(\hat{\tau}_{ps})^{1/2} \frac{r_{ps}(\tau)}{\ell'_{ps}(\tau)} \frac{\pi(\tau)}{\pi(\hat{\tau}_{ps})} .$$

The symbol "$\doteqdot$" in (5) indicates that the approximation holds with error of order $O(n^{-3/2})$. From a practical point of view, the tail area approximation (5) can be used to compute posterior quantiles of τ, or equi-tailed credible intervals as $\{\tau : |r^*_{ps}(\tau)| \leq z_{1-\alpha/2}\}$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. Moreover, it can be used to approximate posterior moments or highest posterior density (HPD) credible intervals when using the HOTA algorithm (see [64,67]). The HOTA algorithm is essentially an inverse transform sampling method, which gives independent samples from the pseudo-posterior distribution.

A numerical possibility for a finite-sample validation of Bayesian inference based on $\pi_{ps}(\tau|y)$ is to use the procedure by Mohanan-Boos (1992). These authors discuss a criterion for evaluating whether or not an alternative likelihood can be used for Bayesian inference and, to this end, they introduce a definition of validity, based on the coverage properties of posterior credible sets. In practice, they compute the statistic $H = \int_{-\infty}^{\tau} \pi_{ps}(t|y)\, dt$, which corresponds to posterior coverage set functions of the form $(-\infty, t^{\alpha}]$, where t^{α} is the αth percentile of the pseudo-posterior distribution. They assume that $\pi_{ps}(\tau|y)$ is valid by coverage if H is uniformly distributed in (0, 1). Validity of Bayesian inference for the empirical likelihood was assessed in [42], for the quasi-likelihood in [36], and for the weighted likelihood in [4].

4 Three Examples of Pseudo-Posterior Distributions

In this section we illustrate the calculation of pseudo-posterior distributions in three illustrative examples based on: the modified profile likelihood in a one-way random effects model with heteroscedastic error variances, the partial likelihood in the Cox proportional hazards model, and the composite likelihood in a multilevel probit model. It is argued that pseudo-posterior distributions have an important role to play in Bayesian statistics.

4.1 Elimination of Nuisance Parameters with Matching Priors

Let $\theta = (\tau, \lambda)$, with τ scalar parameter of interest and λ multidimensional nuisance parameter. Bayesian inference on τ is based on the marginal posterior distribution

$$\pi_m(\tau|y) = \int \pi(\theta|y)\, d\lambda = \frac{\int \pi(\tau, \lambda) L(\tau, \lambda)\, d\lambda}{\int \pi(\tau, \lambda) L(\tau, \lambda)\, d\lambda d\tau} . \tag{6}$$

The computation of (6) may present some difficulties. First of all, it requires the elicitation on both ψ and λ. Second, it requires a multidimensional numerical integration.

These drawbacks can be avoided when using the class of matching priors in $\pi_m(\tau|y)$. In this case, the marginal posterior distribution can be written as (see, e.g., [81], and references therein)

$$\pi_m(\tau|y) \propto \pi_{mp}(\tau) L_{mp}(\tau) \,, \tag{7}$$

where $\pi_{mp}(\tau)$ is the matching prior (3), and $L_{mp}(\tau) = L_p(\tau) M(\tau)$ is the modified profile likelihood for τ with $M(\tau)$ suitable defined correction term. The advantages of (7) are that: (1) no elicitation on the nuisance parameter λ is required; (2) no numerical integration or MCMC simulation is needed; (3) accurate Bayesian inference even for small sample sizes. Moreover, it can routinely be applied in practice using results from likelihood asymptotics and the `R` package bundle `hoa` (see [81]).

Accurate tail probabilities from (7) can be computed using the third-order approximation (5), which reduces to (see also [80])

$$\int_{\tau_0}^{\infty} \pi_m(\tau|y)\, d\tau \doteq \Phi(r_p^*(\tau_0)) \,, \tag{8}$$

where

$$r_{ps}^*(\tau) = r_{ps}(\tau) + r_{ps}(\tau)^{-1} \log \frac{q(\tau)}{r_p(\tau)}$$

is the modified directed profile likelihood of [7], with

$$q(\tau) = \ell_p'(\tau) \frac{i_{\tau\tau.\lambda}(\hat{\tau}, \hat{\lambda})^{1/2}}{i_{\tau\tau.\lambda}(\tau, \hat{\lambda}_\tau)^{1/2}} \frac{1}{M(\tau)} \,.$$

The prior $\pi_{mp}(\tau)$ is also a strong matching prior [33] since a frequentist p-value coincides with a Bayesian posterior survivor probability. Moreover, note that the equi-tailed credible interval $\{\mu : |r_p^*(\tau)| \le z_{1-\alpha/2}\}$ for τ derived from (8) coincides with an accurate higher-order likelihood-based confidence interval for τ with approximate level $(1-\alpha)$. Therefore, this credible interval is also a likelihood-based confidence interval for τ, with accurate frequentist coverage.

In order to illustrate the use of (7), consider inference for the consensus mean in inter-laboratory studies. The analysis of data from inter-laboratory studies has received attention over the past several years, and it deals with the one-way random effects model with heteroscedastic error variances; see, among others [72], and references therein. Let us assume that there are m laboratories, with n_j observations at the j-th laboratory, for $j = 1, \ldots, m$. The model is

$$y_{ij} = \tau + \tau_j + \varepsilon_{ij} \,, \qquad i = 1, \ldots, n_j \,, \quad j = 1, \ldots, m \,, \tag{9}$$

where y_{ij} denotes the i-th observation at the j-th laboratory, and τ_j and ε_{ij} are independent random variables with distribution $\tau_j \sim N(0, \sigma^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_j^2)$, respectively. The parameter of interest is the consensus mean τ, which is also the mean of the $y_{ij}, i = 1, \ldots, n_j$ and $j = 1, \ldots, m$. The remaining $(m+1)$ parameters of the model, i.e., within-laboratory variances $(\sigma_1^2, \ldots, \sigma_m^2)$ and between laboratory variability σ^2, are nuisance parameters. Consider the marginal posterior distribution for τ based on the matching prior $\pi_{mp}(\tau)$. With respect to a standard Bayesian approach (see, e.g., [75]), it does not require the elicitation on the nuisance parameter

$\lambda = (\sigma^2, \sigma_1^2, \ldots, \sigma_m^2)$ and it enables us to perform simple and accurate Bayesian inference also when m and/or the n_j, $j = 1, \ldots, m$, are small. The log likelihood function for τ and $\lambda = (\sigma^2, \sigma_1^2, \ldots, \sigma_m^2)$ from model (9) is given by

$$\ell(\tau, \lambda) = -\frac{1}{2} \sum_{j=1}^{m} \left((n_j - 1) \log \sigma_j^2 - \log \rho_j + \rho_j (\bar{y}_j - \tau)^2 + \frac{(n_j - 1) s_j^2}{\sigma_j^2} \right),$$

with $\rho_j = 1/(\sigma^2 + \sigma_j^2/n_j)$, $\bar{y}_j = \sum_{i=1}^{n_j} y_{ij}/n_j$ and $s_j^2 = \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2/(n_j - 1)$, for $j = 1, \ldots, m$. Starting from $\ell(\tau, \lambda)$, all the quantities involved in (7) are given in [72], which discuss higher-order frequentist confidence intervals for τ. In particular, the matching prior of τ is given by

$$\pi_{mp}(\tau) \propto \sqrt{\sum_{j=1}^{m} \frac{1}{\hat{\sigma}_\tau^2 + \hat{\sigma}_{j\tau}^2/n_j}},$$

with $\hat{\sigma}_\tau^2$ and $\hat{\sigma}_{j\tau}^2$ partial MLEs of σ^2 and σ_j^2, $j = 1, \ldots, m$, for fixed τ. Note that to compute (7), the HOTA simulation scheme can be used [64].

Let us consider the study involving nine laboratories carried out by the Nutrient Composition Laboratory of the US Department of Agriculture. The objective was to validate a proposed simple nonenzymatic gravimetric method for determining total dietary fiber in some foods. Six samples (apple, apricots, cabbage, carrots, onions, and soy fiber) were sent in blind duplicates to the participating laboratories. The data on fiber in apples were analyzed by [75], using non informative priors. For this example, $m = 9$ and the number of measurements n_j made by the jth laboratory is 2, for $j = 1, \ldots, 9$. The posterior distributions for τ are illustrated in Fig. 1, and the credible intervals for the consensus mean and some summary statistics are given in the following table:

	mean (sd)	median	0.95 equi-tailed	0.95 HPD
$\pi_{mp}(\tau \mid y)$	12.91 (0.27)	12.93	(12.35,13.46)	(12.33,13.43)
$\pi_m^{vr}(\tau \mid y)$	12.87 (0.66)	12.90	(12.19,13.61)	(12.19,13.61)
(4)	12.91 (0.22)	12.91	(12.47,13.34)	(12.47,13.34)

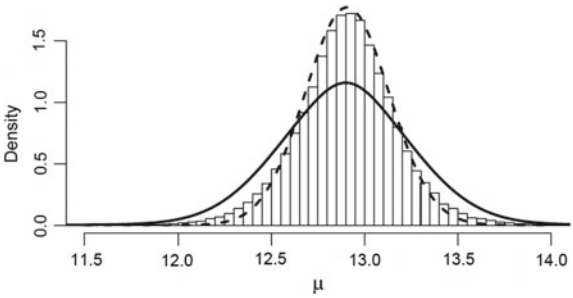


Fig. 1 HOTA posterior distribution (histogram), $\pi_m^{vr}(\tau \mid y)$ (*solid*) and first-order approximation (4) (*dashed*) for the mean dietary fiber in apples

The overall computation time was 1 s. The dashed curve in Fig. 1 is the first-order approximation (4), while the solid curve is the marginal posterior $\pi_m^{vr}(\tau|y)$ for τ discussed in [75]. This posterior is based on the independent priors $\pi(\tau) \propto 1$, $\pi(\sigma_j) \propto 1/\sigma_j$, $j = 1, \ldots, m$, and $\pi(\sigma) \propto 1$. Note that the first-order 95 % equi-tailed credible interval appears unsuitable since it is too short owing to a poor normal approximation to the posterior distribution (see also [15]).

4.2 Inference on the Cox Proportional Hazards Model

The Cox proportional hazards model [22,23] is widely used for semiparametric survival data modeling. In its simplest form the failure times $T_1, \ldots, T_n$, for n independent individuals, have hazard functions $h(t; x_i) = h_0(t) \exp\{x_i^T \beta\}$, where $\beta = (\beta_1, \ldots, \beta_p)$ is a vector of unknown regression parameters, x_i is a $(p \times 1)$ vector of covariates for the ith individual, $i = 1 \ldots, n$, and $h_0(t)$ is the baseline hazard function. Suppose that the failure time is subject to right-censoring by a mechanism independent of their values and uninformative about their distribution. The data are n pairs (t_i, δ_i), where t_i denotes the observed lifetimes for the ith individual and δ_i is an indicator of the survival status, with $d_i = 1$ if t_i is a failure time (uncensored) and $d_i = 0$ if t_i represents a right-censored value, that is if $T_i > t_i, i = 1, \ldots, n$. The partial likelihood for β is given by

$$L_P(\beta) = \prod_{i=1}^{m} \frac{e^{x_i^T \beta}}{\sum_{j \in \mathscr{R}(t_{(i)})} e^{x_j^T \beta}}, \tag{10}$$

where $t_{(i)}$ is the ordered failure time, $\mathscr{R}(t_{(i)})$ is the risk set comprising those individuals at risk at time $t_{(i)}$, $i = 1, \ldots, n$, and $m = \sum_i \delta_i$.

In the Bayesian framework prior opinion should be modeled through a prior process on the baseline cumulative hazard function and a prior density $\pi(\beta)$ on the regression parameters, since both $h_0(t)$ and β are unknown. To avoid issues related to the elicitation on $h_0(t)$, in practice the partial likelihood (10) can be used directly to derive the pseudo-posterior distribution

$$\tilde{\pi}_P(\beta|y) \propto \pi(\beta)\, L_P(\beta); \tag{11}$$

see [39,40,69], and references therein, for various Bayesian applications of (11). Suppose it is of interest to focus on the scalar parameter β_j, i.e., the jth component of β. Let then $\beta = (\psi, \lambda)$, with $\psi = \beta_j$ the parameter of interest and $\lambda = (\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)$ the $(p-1)$-dimensional nuisance parameter. Non-informative priors on β, such as $\pi(\beta) \propto 1$ (see, e.g., [21]) or vague normal priors (see, e.g., [40]), can be considered.

Let us consider a real dataset concerning a clinical study on malignant mesothelioma (MM) [31]; this example is discussed in [67]. The dataset reports censored survival times for $n = 109$ and the type of malignant mesothelioma, i.e., type epithelioid, biphasic, or sarcomatoid. The partial likelihood (10) is thus a function of $\beta = (\beta_1, \beta_2)$.

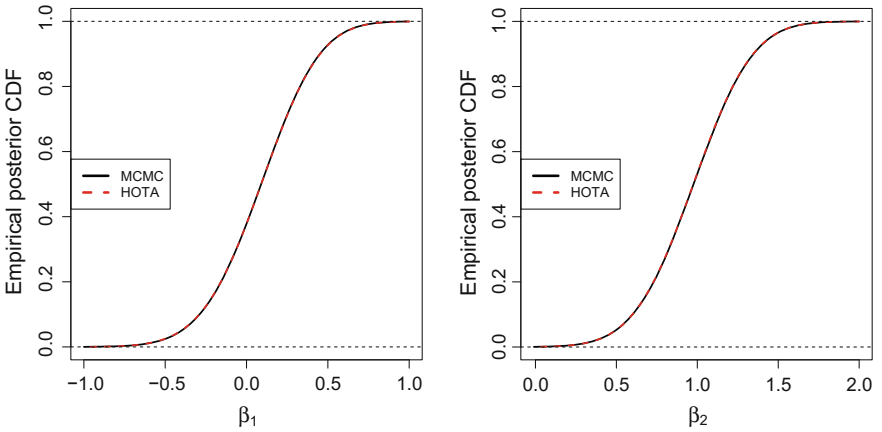


Fig. 2 Marginal posterior distributions for β_1 and β_2 computed with HOTA and MCMC for the Cox regression model

The marginal partial posterior distributions for β_1 and β_2 can be computed both using the HOTA algorithm based on higher-order approximations or with MCMC, both based on 10^4 simulations and a non-informative prior on β. A graphical comparison of the two cumulative distribution functions is given in Fig. 2, whereas numerical comparisons are reported in the following table:

Method		**Mean**	**Std. Dev**	$Q_{0.025}$	**Median**	$Q_{0.975}$	**0.95 HPD**
HOTA	β_1	0.084	0.291	−0.501	0.089	0.641	(−0.480, 0.656)
HOTA	β_2	0.974	0.291	0.396	0.976	1.540	(0.415, 1.557)
MCMC	β_1	0.084	0.291	−0.501	0.089	0.640	(−0.488, 0.644)
MCMC	β_2	0.975	0.292	0.397	0.976	1.541	(0.395, 1.541)

The results indicate that the MCMC and the HOTA algorithm give virtually indistinguishable results. MCMC is run for a large number of simulations and the usual convergence checks and post processing tasks are applied (e.g., thinning, burn-in, etc.), whereas HOTA is very simple to implement in this example since it is available at little additional computational cost over simple first-order approximations. Moreover, HOTA gives independent samples at a negligible computational cost and it can be used for quick prior sensitivity analyses [62], since it is possible to easily assess the effect of different priors on marginal posterior distributions, given the same Monte Carlo error. This is not generally true for MCMC or importance sampling methods, which in general have to be tuned for the specific model and prior.

4.3 Correlated Binary Data

The pairwise likelihood is particularly useful for modeling correlated binary outcomes, as discussed in [43]. This kind of data arise, e.g., in the context of repeated measurements on the same subject, where a maximum likelihood analysis involves multivariate integrals whose dimension equals the cluster sizes.

Let us focus on a multilevel probit model with constant cluster sizes. In particular, let S_i be a latent q-variate normal with mean $\gamma_i = X_i\beta/\sigma$, with β unknown regression coefficient, σ known scale parameter and X_i design matrix for unit i, and covariance matrix Σ, with $\Sigma_{hh} = \sigma^2$, $\Sigma_{hk} = \sigma^2\rho$, $h \neq k$, $i = 1, \ldots, n$. Then, the observed y_{ih} is equal to 1 if $S_{ih} > 0$, and 0 otherwise, for $h = 1, \ldots, q$.

The full likelihood is cumbersome since it entails calculation of multiple integrals of the multivariate normal distribution. On the other hand, the pairwise log likelihood is (see, e.g., [41,43])

$$p\ell(\beta, \rho) = \sum_{i=1}^{n}\sum_{h=1}^{q-1}\sum_{k=h+1}^{q} \log P(Y_{ih} = y_{ih}, Y_{ik} = y_{ik}; \beta, \rho) \,, \tag{12}$$

where $P(Y_{ih} = 1, Y_{ik} = 1; \beta, \rho) = \Phi_2(\gamma_{ih}, \gamma_{ik}; \rho)$ denotes the standard bivariate normal distribution function with correlation coefficient ρ, and $\gamma_{ih} = x_{ih}\beta/\sigma$ is the component h of γ^i $(i = 1, \ldots, n, h, k = 1, \ldots, q)$. Pairwise likelihood inference is much simpler than using the full likelihood since it involves only bivariate normal integrals.

In principle, the pairwise likelihood can be used directly in the Bayes' theorem as it is a genuine likelihood, giving [73]

$$\pi_{p\ell}(\beta, \rho|y) \propto \pi(\beta, \rho)\ \exp(p\ell(\beta, \rho)) \,.$$

However, [58] suggest to combine a calibrated version of the pairwise likelihood with the prior, obtaining the calibrated posterior

$$\pi^c_{p\ell}(\beta, \rho|y) \propto \pi(\beta, \rho)\ \exp(c\ p\ell(\beta, \rho)) \,, \tag{13}$$

with c suitable constant (see formula (2.3) in [58]). The calibration is necessary in order to alleviate the inefficiency of composite likelihood methods. Moreover, the use of $\pi^c_{p\ell}(\beta, \rho|y)$ recovers, approximately, the asymptotic properties of the pairwise posterior. Examples of $\pi_{p\ell}(\beta, \rho|y)$ and of $\pi^c_{p\ell}(\beta, \rho|y)$ are discussed also in [63,65].

Let us consider an example in [65], which discuss the use of the pairwise likelihood function in Approximate Bayesian Computation (ABC) methods. The data have been generated with $\beta_0 = \rho = 0.5$ and $\beta_1 = \sigma = 1$, and with $n = 50$ and $q = 7$, where β_0 is the intercept and β_1 the coefficient of a covariate, which has been generated from a $U(-1, 1)$. For the parameter $\theta = (\beta_0, \beta_1, \kappa)$, with $\kappa = logit((\rho(q-1)+1)/q)$, a normal prior $N(0, 45)^3$ is assumed.

The marginal pairwise posteriors for ρ, β_0 and β_1, derived from the calibrated and non-calibrated pairwise posteriors, are illustrated in Fig. 3. For the purposes of comparison we report also an MCMC approximation of the posterior based on the full likelihood. Clearly, the non-calibrated pairwise posterior is quite different from the target (MCMC), whereas the calibrated pairwise posterior behaves much better.

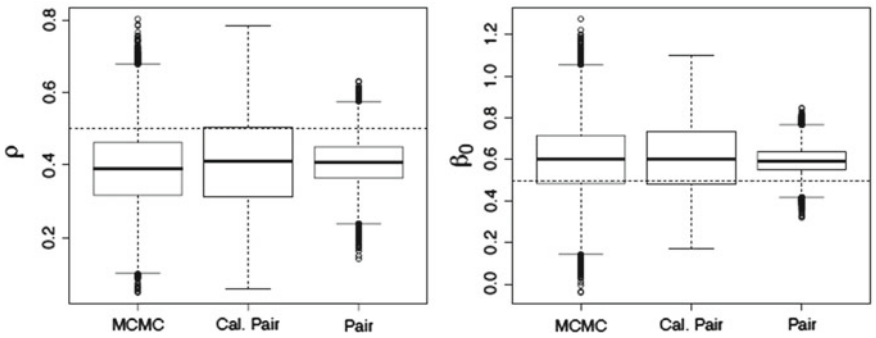


Fig. 3 Correlated binary data: Calibrated pairwise posterior (Cal. Pair) compared with the pairwise (Pair) and the exact (MCMC) posteriors. The *horizontal lines* represent the true parameter values

5 Final Remarks

Posterior distributions based on suitable pseudo-likelihoods have been proved useful for Bayesian inferences on a parameter of interest in several contexts (see also [9]). A first notable situation arises when elimination of a nuisance parameter is of interest. In this case the use of a pseudo-likelihood allows to avoid the elicitation of the prior of the nuisance parameter and the computation of a multidimensional integral in the integrated likelihood. A second striking situation is when the ordinary likelihood, and thus the corresponding posterior distribution, is difficult or even impractical to compute. In this respect, the use of a pseudo-posterior distribution based on the partial and the composite likelihoods may be particularly useful to deal with complex models.

Finally, we note that the interplay between Bayesian and likelihood procedures is still lively and opens to new research topics. A first instance refers to the use of composite likelihood score functions as summary statistics in Approximate Bayesian Computation (ABC) in order to obtain accurate approximations to the posterior distribution in complex models [65]. Moreover, also scoring rules, that generalize the proper and the composite likelihoods, can be used for developing posterior distributions using ABC methods (see the preliminary results in [66]). Finally, in [18] it is shown how higher-order approximations and matching priors are useful to derive an accurate approximation of the measure of evidence for the full Bayesian significance test introduced by [59] for precise hypotheses.

Acknowledgments This work was supported by a grant from the University of Padua (Progetti di Ricerca di Ateneo 2013) and by the grant *Progetto di Ricerca di Base, Legge Regionale Sardegna. 7/2007-2012*.

References

1. Adimari, G.: On the empirical likelihood ratio for smooth functions of M-functionals. Scand. J. Stat. **24**, 47–59 (1997)
2. Adimari, G., Ventura, L.: Quasi-profile log likelihoods for unbiased estimating functions. Ann. Inst. Stat. Math. **54**, 235–244 (2002a)
3. Adimari, G., Ventura, L.: Quasi-likelihood from M-estimators: a numerical comparison with empirical likelihood. Stat. Methods Appl. **11**, 175–185 (2002b)
4. Agostinelli, C., Greco, L.: A weighted strategy to handle likelihood uncertainty in Bayesian inference. Comput. Stat. **28**, 319–339 (2013)
5. Azzalini, A.: Maximum likelihood of order m for stationary stochastic processes. Biometrika **70**, 381–367 (1983)
6. Barndorff-Nielsen, O.E.: Quasi profile and directed likelihoods from estimating functions. Ann. Insti. Stat. Math. **47**, 461–464 (1995)
7. Barndorff-Nielsen, O.E., Chamberlin, S.R.: Stable and invariant adjusted directed likelihoods. Biometrika **81**, 485–499 (1994)
8. Bellio, R., Greco, L., Ventura, L.: Adjusted quasi-profile likelihoods from estimating functions. J. Stat. Plan. Inference **138**, 3059–3068 (2008)
9. Berger, J.O., Bayarri, S.: The interplay between Bayesian and frequentist inference. Stat. Sci. **19**, 58–80 (2004)
10. Berger, J.O., Liseo, B., Wolpert, R.: Integrated likelihood methods for eliminating nuisance parameters. Stat. Sci. **14**, 1–28 (1999)
11. Bertolino, F., Racugno, W.: Analysis of the linear correlation coefficient using pseudo-likelihoods. J. Italian Stat. Soc. **1**, 33–50 (1992)
12. Bertolino, F., Racugno, W.: Robust Bayesian analysis of variance and the x2-test by using marginal likelihoods. The Statistician **43**, 191–201 (1994)
13. Besag, J.E.: Spatial interaction and the statistical analysis of lattice systems (with discussion). J. R. Stat. Soc. B **34**, 192–236 (1974)
14. Bjornstad, J.F.: Predictive likelihood: a review (with discussion). Stat. Sci. **5**, 242–265 (1990)
15. Brazzale, A.R., Davison, A.C., Reid, N.: Applied asymptotics. Case Studies in Small- Sample Statistics. Cambridge University Press, Cambridge (2007)
16. Butler, R.W.: Approximate predictive pivots and densities. Biometrika **76**, 489–501 (1989)
17. Cabras, S., Castellanos, M.E., Racugno, W., Ventura, L.: A matching prior for the shape parameter of the skew-normal distribution. Scand. J. Stat. **39**, 236–247 (2012)
18. Cabras, S., Racugno, W., Ventura, L.: Higher-order asymptotic computation of Bayesian significance tests for precise null hypotheses in the presence of nuisance parameters. J. Stat. Comput. Simul. (to appear) (2014)
19. Chang, H., Mukerjee, R.: Probability matching property of adjusted likelihoods. Stat. Probab. Lett. **76**, 838–842 (2006)
20. Chang, H., Kim, B.H., Mukerjee, R.: Bayesian and frequentist confidence intervals via adjusted likelihoods under prior specification on the interest parameter. Statistics **43**, 203–211 (2009)
21. Chen, M., Ibrahim, J., Shao, Q.: Posterior propriety and computation for the Cox regression model with applications to missing covariates. Biometrika **93**, 791–807 (2006)
22. Cox, D.R.: Regression models and life tables (with discussion). J. R. Stat. Soc. B **34**, 187–200 (1972)
23. Cox, D.T.: Partial likelihood. Biometrika **62**, 269–276 (1975)
24. Cox, D.R., Reid, N.: Orthogonal parameters and approximate conditional inference (with discussion). J. R. Stat. Soc. B. **49**, 1–39 (1987)
25. Cox, D.D., O'Sullivan, F.: Asymptotic analysis of penalized likelihood and related estimators. Ann. Stat. **18**, 1676–1695 (1990)
26. Cox, D.R., Reid, N.: A note on pseudo likelihood constructed from marginal densities. Biometrika **91**, 729–737 (2004)

27. Datta, G.S., Mukerjee, R.: Probability Matching Priors: Higher Order Asymptotics. Springer, Berlin (2004)
28. Davison, A.C., Hinkley, D.V., Worton, B.J.: Bootstrap likelihood. Biometrika **79**, 113–130 (1992)
29. Davison, A.C., Hinkley, D.V., Worton, B.J.: Accurate and efficient construction of bootstrap likelihoods. Stat. Comput. **5**, 257–264 (1995)
30. Efron, B.: Bayes and likelihood calculations from confidence intervals. Biometrika **80**, 3–26 (1993)
31. Epithelial-mesenchymal transition in malignant mesothelioma: Fassina, A., Cappellesso, R., Guzzardo, V., Dalla Via, L., Piccolo, S., Ventura, L., Fassan, M. Mod. Pathol. **25**, 86–99 (2012)
32. Fraser, D.A.S., Reid, N.: Bayes posteriors for scalar interest parameters. Bayesian Stat. **5**, 581–585 (1996)
33. Fraser, D.A.S., Reid, N.: Strong matching of frequentist and Bayesian parametric inference. J. Stat. Plan. Inference **103**, 263–285 (2002)
34. Direct Bayes for interest parameters: Fraser, D.A.S., Reid, N., Wong, A., Yun Yi, G. Valencia **7**, 529–533 (2003)
35. Gong, G., Samaniego, F.J.: Pseudo maximum likelihood estimation: theory and applications. Ann. Stat. **89**, 861–869 (1981)
36. Greco, L., Racugno, W., Ventura, L.: Robust likelihood functions in Bayesian inference. J. Stat. Plan. Inference **138**, 1258–1270 (2008)
37. Green, P.J.: Penalized likelihood for general semi parametric regression models. Int. Stat. Rev. **55**, 245–260 (1987)
38. Hu, F., Zidek, J.V.: The weighted likelihood. Can. J. Stat. **30**, 347–371 (2002)
39. Ibrahim, J., Chen, M., Sinha, D.: Bayesian Survival Analysis. Springer, New York (2002)
40. Kim, Y., Kim, D.: Bayesian partial likelihood approach for tied observations. J. Stat. Plan. Inference **139**, 469–477 (2009)
41. Kuk, A., Nott, D.: A pairwise likelihood approach to analyzing correlated binary data. Stat. Probab. Lett. **47**, 329–335 (2000)
42. Lazar, N.A.: Bayes empirical likelihood. Biometrika **90**, 319–326 (2003)
43. Le Cessie, S., van Houwelingen, J.C.: Logistic regression for correlated binary data. Appl. Stat. **43**, 95–108 (1994)
44. Lee, Y., Nelder, J.A.: Hierarchical generalized linear models (with discussion). J. R. Stat. Soc. B. **58**, 619–678 (1996)
45. Lee, Y., Nelder, J.A., Noh, M.: H-likelihood: problems and solutions. Stat. Comput. **17**, 49–55 (2007)
46. Lin, L.: Quasi Bayesian likelihood. Stat. Methodol. **3**, 444–455 (2006)
47. Markatou, M., Basu, A., Lindsay, B.G.: Weighted likelihood estimating equations with a bootstrap root search. J. Am. Stat. Assoc. **93**, 740–750 (1998)
48. McCullagh, P.: Quasi-likelihood functions. Ann. Stat. **11**, 59–67 (1983)
49. McLeish, D.L., Small, C.G.: A projected likelihood function for semiparametric models. Biometrika **79**, 93–102 (1992)
50. Meng, X.L.: Decoding the H-likelihood. Stat. Sci. **24**, 280–293 (2009)
51. Min, X., Sun, D.: A matching prior based on the modified profile likelihood in a generalized Weibull stress-strength model. Can. J. Stat. **41**, 83–97 (2013)
52. Monahan, J. F., Boos, D. D.: Proper likelihoods for Bayesian analysis. Biometrika. Oxford University Press (OUP), **79**(2), 271–278 (1992). http://dx.doi.org/10.1093/biomet/79.2.271
53. Mykland, P.A.: Dual likelihood. Ann. Stat. **23**, 396–421 (1995)
54. Owen, A.B.: Empirical Likelihood. Chapman & Hall, London (2001)
55. Pace, L., Salvan, A.: Principles of Statistical Inference. World Scientific, Singapore (1997)
56. Pace, L., Salvan, A.: Adjustments of the profile likelihood from a new perspective. J. Stat. Plan. Inference **136**, 3554–3564 (2006)

57. Pace, L., Salvan, A., Ventura, L.: Likelihood based discrimination between separate scale and regression models. J. Stat. Plan. Inference **136**, 3539–3553 (2006)
58. Pauli, F., Racugno, W., Ventura, L.: Bayesian composite marginal likelihoods. Stat. Sin. **21**, 149–164 (2011)
59. Pereira, C., Stern, J.: Evidence and credibility: full Bayesian significance test for precise hypotheses. Entropy **1**, 99–110 (1999)
60. Racugno, W., Salvan, A., Ventura, L.: Bayesian analysis in regression models using pseudo-likelihoods. Commun. Stat. Theory Methods **39**, 3444–3455 (2010)
61. Reid, N., Mukerjee, R., Fraser, D.A.S.: Some aspects of matching priors. Lect. Notes-Monogr. Ser. **42**, 31–43 (2003)
62. Reid, N., Sun, Y.: Assessing sensitivity to priors using higher order approximations. Commun. Stat. Theory Methods **39**, 1373–1386 (2010)
63. Ribatet, M., Cooley, D., Davison, A.: Bayesian inference from composite likelihoods, with an application to spatial extremes. Stat. Sin. **22**, 813–845 (2012)
64. Ruli, E., Sartori, N., Ventura, L.: Marginal posterior simulation via higher-order tail area approximations. Bayesian Anal. **9**, 129–146 (2014)
65. Ruli, E., Sartori, N., Ventura, L.: Approximate Bayesian computation with composite score functions (submitted)
66. Ruli, E., Sartori, N., Ventura, L.: Approximate Bayesian computation with proper scoring rules. Atti della XLVII Riunione Scientifica della SIS. Cagliari, 11-13 giugno 2014, 1–6 (2014)
67. Ruli, E., Ventura, L.: Higher-order Bayesian approximations for pseudo-posterior distributions. Commun. Stat. Simul. Comput. (to appear)
68. Schennach, S.M.: Bayesian exponentially tilted empirical likelihood. Biometrika **92**, 31–46 (2005)
69. Sinha, D., Ibrahim, J., Chen, M.: A Bayesian justification of Cox's partial likelihood. Biometrika **90**, 629–641 (2003)
70. Severini, T.A.: On the relationship between Bayesian and non-Bayesian elimination of nuisance parameters. Stat. Sin. **9**, 713–724 (1999)
71. Severini, T.A.: Likelihood Methods in Statistics. Oxford University Press, New York (2000)
72. Sharma, G., Mathew, T.: Higher order inference for the consensus mean in inter-laboratory studies. Biom. J. **53**, 128–136 (2011)
73. Smith, E.L., Stephenson, A.G.: An extended Gaussian max-stable process model for spatial extremes. J. Stat. Plan. Inference **139**, 1266–1275 (2009)
74. Stein, M.L.: Approximating likelihoods for large spatial data sets. J. R. Stat. Soc. B **66**, 275–296 (2004)
75. Vangel, M.G., Rukhin, A.L.: Maximum likelihood analysis for heteroscedastic one-way random effects ANOVA in interlaboratory studies. Biometrics **55**, 129–136 (1999)
76. Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. Stat. Sin. **21**, 5–42 (2011)
77. Ventura, L., Cabras, S., Racugno, W.: Prior distributions from pseudo-likelihoods in the presence of nuisance parameters. J. Am. Stat. Assoc. **104**, 768–774 (2009)
78. Ventura, L., Cabras, S., Racugno, W.: Default prior distributions from quasi- and quasi-profile likelihoods. J. Stat. Plan. Inference **140**, 2937–2942 (2010)
79. Ventura, L., Racugno, W.: Recent advances on Bayesian inference for $P(X < Y)$. Bayesian Anal. **6**, 411–428 (2011)
80. Ventura, L., Reid, N.: Approximate Bayesian computation with modified log likelihood ratios. Metron **7**, 231–245 (2014)
81. Ventura, L., Sartori, N., Racugno, W.: Objective Bayesian higher-order asymptotics in models with nuisance parameters. Comput. Stat. Data Anal. **60**, 90–96 (2013)
82. Wang, J.: Quadratic artificial likelihood functions using estimating functions. Scand. J. Stat. **33**, 379–390 (2006)