ISSN: 1696-2281
SORT 34 (1) January-June (2010)

SORT

Statistics and Operations Research Transactions

Sponsoring institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Institut d'Estadística de Catalunya

Supporting institution

Spanish Region of the International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

FSC
www.fsc.org
100%
Paper procedent de boscos ben gestionats
FSC™ C081765

SORT

Volume 34

Number 1

January-June 2010

ISSN: 1696-2281

Articles

Selected articles from *Congreso Español de Biometría 2009*

Editor's report

During 2009 the number of manuscripts submitted to SORT - Statistics and Operations Research Transactions has increased substantially. The editorial board confirms that authors find SORT more attractive when it is more efficiently indexed and cited. Table 1 provides the historical summary of editorial processes, submitted to the journal since 2007. This table shows trends in submissions, acceptance rates, and external submission rates (received from outside of Catalonia). The number of submissions increased to an unprecedented total of 55 in 2009, which is more than four times the number received in 2008. The acceptance rate now may decrease to less than 20% and therefore vary a lot compared to past rates above 50% once the manuscripts received in 2009 are fully reviewed.

The increase of submissions implies that the average number of manuscripts received is now one per week, which is a record number in the history of the journal. Moreover, given the current number of articles being submitted and their very high quality, editors are now obliged to be very selective on the manuscripts that are potentially publishable by the journal. Articles are evaluated on the basis that they have an excellent quality of presentation and those results showing an applied orientation are given priority.

Table 1: *Summary of SORT processes*

Year	Submissions	Acceptance rate (%)	External submissions (%)
2007	18	72.22	83.33
2008	13	53.85	69.23
2009	55	9.09*	94.55

*Decision on 24 articles is still pending

The impact assessment, produced by the Journal Citation Reports of the ISI (Institute for Scientific Information) is one of the most frequently used measures of the quality of a journal. The JCR impact factor takes into account citations from articles published in the two preceding years. The 2009 score is based on citations of articles published in 2007 and 2008. The score is the ratio of the number of 2009 citations drawn from articles published in 2007 and 2008 divided by the number of articles published in 2007 and 2008. In 2009 the impact factor that we have calculated unofficially for SORT is at least 0,33. We expect that the release of 2009 scores from ISI in mid June will confirm these good news.

The publisher and the board of sponsors have decided a couple of years ago that SORT will be a fully open-access journal in order to offer timely delivery and electronic availability free of cost. Printed paper copies continue to be distributed to subscribers, and subscription rates only reflect printing costs. We confirm that having the articles available on line has contributed to a dissemination of the research results published in SORT.

The editorial team is grateful to the publisher and sponsors for their generosity and to all authors, associate editors and referees who have been involved in SORT, and to whom we are indebted. We need to acknowledge the work of Elisabet Aznar, who has done an outstanding job in handling the demands of SORT and has coped with the increasing activity of our journal.

Montserrat Guillén
Chief Editor

In memoriam: Stephen W. Lagakos (1946-2009)

Stephen W. Lagakos was a Professor of Biostatistics of the Harvard School of Public Health since 1986. He earned his Ph. D. at the George Washington University (Washington, D.C.) in 1972. After finishing his doctorate he joined the Department of Statistics of the State University of New York at Buffalo, NY. He joined Harvard in 1978, held appointments at Harvard and at the Dana-Farber Cancer Institute from then onwards and served as Chair of the Department of Biostatistics for eight years. He was also the founder and director of the Center for Biostatistics in AIDS Research (CBAR) at the Harvard School of Public Health and a co-founding member of Frontier Science & Technology Research Foundation.

Steve received many awards during his life, such as the Spiegelman Gold Medal Award from the American Public Health Association in 1983 and an honorary doctorate from the National and Kapodistrian University of Athens in 2006, among others. He was also an elected Fellow of the IMS, ASA, ISI and of the American Association for the Advancement of Science.

His research as a biostatistician was outstanding, with a large number of seminal papers in the leading journals of our profession and with very broad scientific interests. He was undoubtedly the leading biostatistician in the world working on clinical trials in AIDS: his interest began with the onset of the epidemic. During all these years he collaborated with clinicians, virologists, immunologists and other scientists to tackle the most relevant scientific questions. Along with his students and colleagues he developed many new statistical methods. He served on committees of the National Institute of Health and the Food and Drug Administration, was Associate Editor of the Journal of the American Statistical Association, a long time member of the Editorial Board of the New England Journal of Medicine and was a member of the Editorial Advisory Committee since the beginning of our journal SORT. His collaboration with SORT started in 2002 when we put together the First Barcelona Workshop on Survival Analysis. This successful event coincided with the transition from the journal **Qüestiió** to **SORT**. Steve's encouragement for this new Catalan journal was remarkable, and he helped us a great deal with getting the submissions of most of the papers for the first issues of SORT.

The first time I talked to him I was finishing my Ph.D at Columbia University, NY. A couple of years later I went to the Biostatistics Department for the first time and our collaboration in AIDS research started. Since then he invited me many times. He was always generous with his time, and he would even feel guilty of not devoting enough of it. Steve was very warm and I was always moved by his tenderness with children. He always made you feel that he was enjoying your and your family's company. We were often invited to his cottage in Rindge, NH, and although we never hiked to Mt. Monadnock we did swim, canoe and enjoy tubing in Lake Monomonac.

Steve was a great, genuinely sincere, funny and modest person who taught me much more than biostatistics. I remember his brilliant sense of humor and common-sense priorities, as well as his ability to make you feel the most important person in the world. He was always interested in other people and was a wonderful listener. His wisdom and advice were invaluable. He always offered encouragement and support during setbacks and losses and was happy to celebrate with you the major milestones of your life: children, papers, promotions. It was a true privilege to collaborate with him and he will be missed greatly. Steve will always live in our hearts.

Guadalupe Gómez, Executive Editor
Barcelona, May 2010

SORT 34 (1) January-June 2010, 3-20

Small-sample inference about variance and its transformations[*]

N. T. Longford

SNTL Statistics Research and Consulting

Abstract

We discuss minimum mean squared error and Bayesian estimation of the variance and its common transformations in the setting of normality and homoscedasticity with small samples, for which asymptotics do not apply. We show that permitting some bias can be rewarded by greatly reduced mean squared error. We apply borderline and equilibrium priors. The purpose of these priors is to reduce the onus on the expert or client to specify a single prior distribution that would capture the information available prior to data inspection. Instead, the (parametric) class of all priors considered is partitioned to subsets that result in the preference for different actions. With the family of conjugate inverse gamma priors, this Bayesian approach can be formulated in the frequentist paradigm, describing the prior as being equivalent to additional observations.

MSC: 62F10 and 62F15

Keywords: Borderline prior, equilibrium prior, expected loss, gamma distribution, mean squared error, plausible prior

1. Introduction

We consider the problem in which a small random sample from a normal distribution, $\mathcal{N}(\mu,\sigma^2)$, is observed and we would like to estimate the variance σ^2 or its transformation, such as σ, $1/\sigma^2$ or $1/\sigma$, or to know whether σ^2 exceeds (or falls short of) a specified threshold σ^2_{R}. We study two approaches: minimum mean squared error (MSE) estimation, to which we refer as *efficient* estimation, and application of (Bayesian) priors. We use only the conjugate family of priors, both for computational simplicity and

[*] N. T. Longford, SNTL and Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25–27, 08005 Barcelona, Spain; email: NTL@sntl.co.uk.

Received: July 2009
Accepted: February 2010

because their representation in terms of additional observations can greatly aid the process of eliciting prior information from an expert. We find a frequentist interpretation of the (Bayesian) posterior distribution which makes the Bayesian approach accessible to frequentist analysis. In the motivating problem, two alternative actions, A and B, are contemplated; A is appropriate when $\sigma^2 < \sigma^2_{\mathrm{R}}$ and B when $\sigma^2 > \sigma^2_{\mathrm{R}}$. There is some prior information, but the analyst's client is either not available or elicitation of a single prior from him or her is unlikely to be constructive.

We are concerned only with analysis of small samples. In large samples, asymptotics apply and maximum likelihood (ML) estimation is satisfactory. The prior information has a diminishing impact and a nonlinear transformation of a parameter is estimated by the same transformation of the ML estimator of the parameter. In small samples, the prior has a non-trivial impact, and efficiency is not maintained by nonlinear transformations. Therefore, efficient (frequentist) estimation of σ^2, σ, $1/\sigma^2$ and $1/\sigma$ are, in principle, distinct problems, and the prior for a Bayesian analysis has to be selected with integrity and care.

The next section deals with efficient estimation. Section 3.1. introduces borderline priors and Section 3.2. equilibrium priors and the related solutions. Equilibrium priors incorporate the losses due to making an incorrect decision (choosing A when $\sigma^2 > \sigma^2_{\mathrm{R}}$ or B when $\sigma^2 < \sigma^2_{\mathrm{R}}$). The perspective of Section 2. is entirely frequentist, while Section 3. might appear at first as entirely Bayesian, exploiting prior information. However, the Bayesian analysis has a frequentist interpretation, with the prior regarded as additional observations. The concluding section summarises the proposed methods.

2. Efficient small-sample estimation

Suppose $y_1,\ldots,y_n$ is a random sample from a normal distribution $\mathcal{N}(\mu,\sigma^2)$. The variance σ^2 is commonly estimated by the corrected mean squares,

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\hat{\mu})^2 \,, \tag{1}$$

where $\hat{\mu} = (y_1+\cdots+y_n)/n$ is the sample mean. The estimator $\hat{\sigma}^2$ has a scaled χ^2 distribution with $n-1$ degrees of freedom:

$$(n-1)\,\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1} \,.$$

The χ^2_k distribution has the density function

$$f(x) = \frac{1}{\Gamma\left(\frac{k}{2}\right)}\left(\frac{1}{2}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1}\exp\left(-\frac{x}{2}\right) .$$

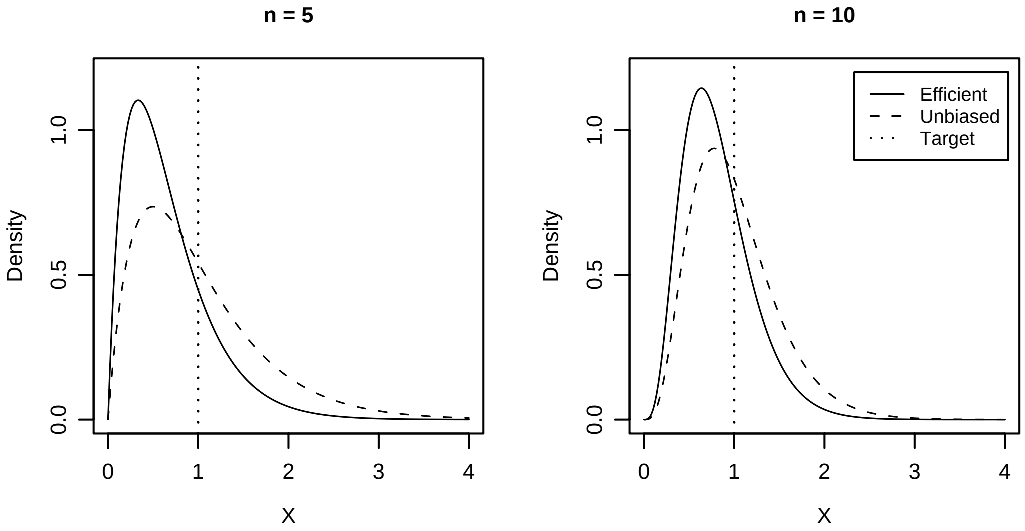


Figure 1: *The densities of the estimators $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ for $n = 5$ and $n = 10$. Both panels are based on the setting $\sigma^2 = 1$.*

The correction for the degree of freedom lost, by using the divisor $n-1$ in (1), is generally regarded as important because the estimator $\hat{\sigma}^2$ is unbiased. However, if we do not insist on unbiasedness we obtain a more efficient estimator as $\tilde{\sigma}^2 = c^*\hat{\sigma}^2$, with $c^* = (n-1)/(n+1)$, that is, with divisor $n+1$ in (1); see Markowitz (1968) and Stuart (1969). The MSE reduction from $2\sigma^4/(n-1)$ to $2\sigma^4/(n+1)$, by $100(1-c^*)\% = 200/(n+1)\%$, converges to zero as $n \to +\infty$, but for small n it is far from trivial. The densities of $\hat{\sigma}^2$ and $\tilde{\sigma}^2$, based on samples of sizes $n = 5$ and 10, are drawn in Figure 1 for target $\sigma^2 = 1$. From the diagram it is difficult to judge that $\hat{\sigma}^2$ (dashes) is less efficient than $\tilde{\sigma}^2$ (solid line) because their densities are distinctly asymmetric and have different shapes. However, the distribution of $\hat{\sigma}^2$ has a thicker right-hand tail than $\tilde{\sigma}^2$, which corresponds to greater probability of large positive estimation errors $\hat{\sigma}^2 - \sigma^2$.

Estimates of variances are used in a variety of roles, and are often involved in nonlinear functions. For example, variance ratios ν^2/σ^2 are estimated when comparing two variances using their (independent) estimators and the standardised value in meta-analysis (Sutton *et al.*, 2000; Longford, 2010) is defined as μ/σ, where μ is the (average) treatment effect and σ the standard deviation of the study-specific treatment effects. The efficiency of $\tilde{\sigma}^2$ is eroded by a nonlinear transformation, so $\tilde{\sigma}^2$ should not be substituted for σ^2 in a nonlinear expression, unless the sampling variation of $\tilde{\sigma}^2$ is very small. For example, neither $1/\hat{\sigma}^2$ nor $1/\tilde{\sigma}^2$ is an efficient or unbiased estimator of the precision $1/\sigma^2$. The respective expectation and variance of $1/\hat{\sigma}^2$ are $(n-1)/(n-3)/\sigma^2$ when $n > 3$ and $2(n-1)^2/\{(n-3)^2(n-5)\sigma^4\}$ when $n > 5$. These expressions are obtained by relating the relevant integrand to another χ^2 distribution or by differentiating the moment generating function; see Stuart (1969) and Stuart and Ord (1994, Chapter 16).

Substituting $\hat{\sigma}^2$ or $\tilde{\sigma}^2$ for σ^2 when it is (a factor) in a denominator is ill-advised when $n < 6$ because the resulting statistic has infinite variance.

We consider first estimators $c/\hat{\sigma}^2$ of $1/\sigma^2$. For $n > 5$, their MSEs are

$$\begin{aligned} & \frac{1}{\sigma^4}\left[\frac{2c^2(n-1)^2}{(n-3)^2(n-5)} + \left\{\frac{c(n-1)}{n-3} - 1\right\}^2\right] \\ = & \frac{1}{\sigma^4}\left\{c^2\frac{(n-1)^2}{(n-3)(n-5)} - 2c\frac{n-1}{n-3} + 1\right\}, \end{aligned}$$

so their minimum is attained for $c^* = (n-5)/(n-1)$. The minimum attained is $2/\{(n-3)\sigma^4\}$, smaller than the MSE of the naive estimator $1/\hat{\sigma}^2$, equal to $2(n+3)/\{(n-3)(n-5)\sigma^4\}$, or the MSE of the unbiased estimator $(n-3)/\{(n-1)\hat{\sigma}^2\}$, equal to $2/\{(n-5)\sigma^4\}$, so long as $n > 5$.

Although $c^*/\hat{\sigma}^2$ is much more efficient than $1/\hat{\sigma}^2$ for $n = 6, \ldots, 10$, it does not address the problem of infinite variance for $n \leq 5$. This problem is resolved by the estimator $1/(d + \hat{\sigma}^2)$ for a positive constant d, but the optimal value of d cannot be derived analytically. (A closed form expression for the MSE of this estimator involves incomplete gamma functions.) We explore this estimator by simulations in the next section.

The variance is often used in a linear function of σ or $1/\sigma$. Efficient estimators of these quantities in the respective classes of estimators $c\hat{\sigma}$ and $c/\hat{\sigma}$ are derived similarly to the efficient estimators of $\hat{\sigma}^2$ and $1/\hat{\sigma}^2$. Let

$$U_n = \frac{\sqrt{2}}{\sqrt{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}.$$

Then $\mathrm{E}(\hat{\sigma}) = \sigma U_n$ and $\mathrm{var}(\hat{\sigma}) = \sigma^2\left(1 - U_n^2\right)$, so the MSE of $c\hat{\sigma}$ is

$$\sigma^2\left\{(1 - cU_n)^2 + c^2\left(1 - U_n^2\right)\right\} = \sigma^2\left(1 - 2cU_n + c^2\right).$$

This function of c attains its minimum for $c^* = U_n$, and the minimum attained is $\sigma^2(1 - U_n^2)$.

For estimating $1/\sigma$ we introduce the constants

$$V_n = \frac{\sqrt{n-1}}{\sqrt{2}}\frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} = \frac{\sqrt{n-1}}{\sqrt{n-2}}\frac{1}{U_{n-1}}.$$

Then

$$\begin{aligned} \mathrm{E}\left(\frac{1}{\hat{\sigma}}\right) &= \frac{V_n}{\sigma} \\ \mathrm{var}\left(\frac{1}{\hat{\sigma}}\right) &= \frac{1}{\sigma^2}\left(\frac{n-1}{n-3} - V_n^2\right). \end{aligned}$$

Hence the MSE of $c/\hat{\sigma}$ is

$$\frac{1}{\sigma^2}\left\{c^2\,\frac{n-1}{n-3} - 2cV_n + 1\right\},$$

and so the estimator of $1/\sigma$ efficient in the class of estimators $c/\hat{\sigma}$ is $\hat{\sigma}^{-1}V_n(n-3)/(n-1)$, so long as $n > 3$. The corresponding MSE is $\{1-V_n^2(n-3)/(n-1)\}/\sigma^2$. Estimators of the form $c/(d+\hat{\sigma})$, with a positive offset d, may be more efficient; they have finite variances for any sample size n.

2.1. Estimating a reciprocal with an offset

We explore next estimating the reciprocal $1/\sigma$ by $1/(d+\hat{\sigma})$. We do this by simulations because we have no convenient expression for the moments of $1/(d+\hat{\sigma})$. Figure 2 displays the empirical biases and root-MSEs of the estimators $1/(d+\hat{\sigma})$ for $d \in (0, 0.5)$ and $\sigma^2 = 0.1, 0.25, 0.5$ and 1.0, based on a sample of size $n = 4$. The values of d for which the estimator is unbiased and for which it attains minimum MSE are marked by vertical ticks at the bottom of the respective panels.

Unbiasedness and minimum MSE are attained for different values of d. The minimum MSE is attained for $d^* = 0.13, 0.21, 0.29$ and 0.41 when $\sigma^2 = 0.1, 0.25, 0.5$ and 1.0, respectively. Although d^* varies substantially with σ^2, the root-MSEs become more and more flat as σ^2 increases. Therefore, the choice of d is less critical for large σ^2, and

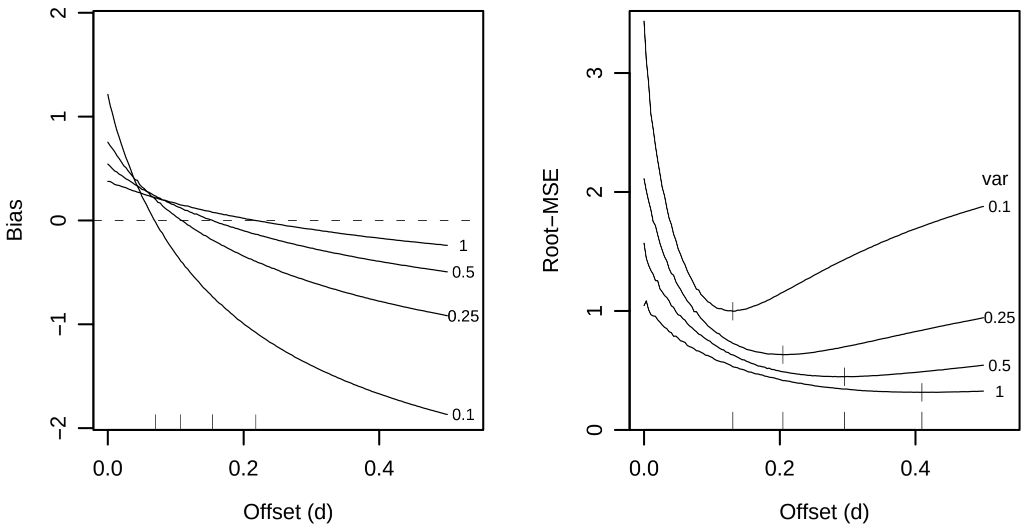


Figure 2: The bias and root-MSE of the estimator $1/(d+\hat{\sigma})$ of $1/\sigma$ as functions of the offset d, with $n = 4$ (3 degrees of freedom). The variances σ^2 are indicated at the right-hand margins. The ticks at the bottom of each panel indicate the value of d for which the estimator is unbiased (left-hand panel) and for which it attains minimum MSE (right-hand panel). Based on 100 000 replications.

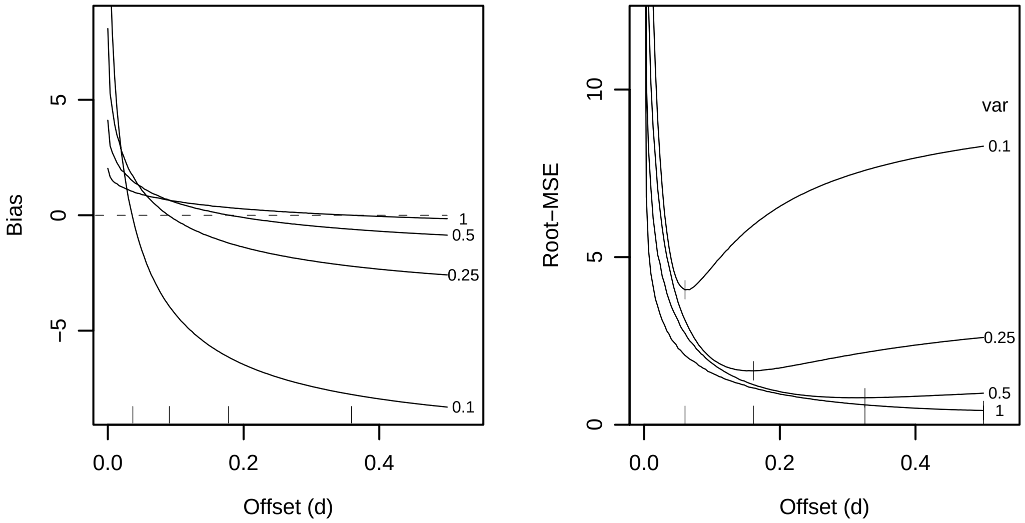


Figure 3: *The bias and root-MSE of the estimator* $1/(d+\hat{\sigma}^2)$ *of* $1/\sigma^2$ *as functions of the offset* d*, with* $n = 4$ *(3 degrees of freedom). Based on 100 000 replications. The same layout is used as in Figure 2.*

should be informed principally by the smallest plausible value of σ^2. This is a better strategy than using the value $\hat{d}^* = d^*(\hat{\sigma}^2)$ that would be optimal if our estimate were exact. If we can rule out small values of σ^2, a value $d > d^*$ is a safe choice because the root-MSE increases very slowly for $d > d^*$.

Figure 3 presents the biases and root-MSEs of the estimator $1/(d+\hat{\sigma}^2)$ of $1/\sigma^2$. It highlights how excessive bias and MSE are avoided by choosing a positive offset d. We arrive at the same general conclusion that if small values of σ^2 can be ruled out, then it is safe to choose a value d that is sufficiently large, because the root-MSEs are flat functions of d for d greater than the optimum offset.

Figures 4 and 5 display the biases and root-MSEs of the respective estimators $1/(d+\hat{\sigma})$ and $1/(d+\hat{\sigma}^2)$ of $1/\sigma$ and $1/\sigma^2$ for sample sizes $n = 6, 11$ and 21. They confirm that the root-MSE is a flat function of d for large σ^2. The precise choice of d is less important for greater variances σ^2, but the estimator $1/(d+\hat{\sigma})$ is very inefficient when too large a value of d is selected, especially when the variance σ^2 is small. Table 1 summarises the results for $\sigma^2 = 1$. The results for different values of σ^2 are obtained by replacing d with d/σ and applying the appropriate rescaling to the bias and MSE. The naive estimator of $1/\sigma$ is perceptibly inefficient even for $n = 21$, and the unbiased estimator is even more inefficient. The difference between the root MSEs of the two estimators that are efficient in the respective classes $c/\hat{\sigma}$ and $1/(d+\hat{\sigma})$ is about 8% for $n = 21$, and much more for smaller n.

One drawback of the estimator $1/(d^*+\hat{\sigma})$ is that the ideal offset d^* depends on σ^2. Therefore, the estimators with an offset can be compared more equitably with $c^*/\hat{\sigma}$ by finding the range of values d for which $1/(d+\hat{\sigma})$ is more efficient than $c^*/\hat{\sigma}$.

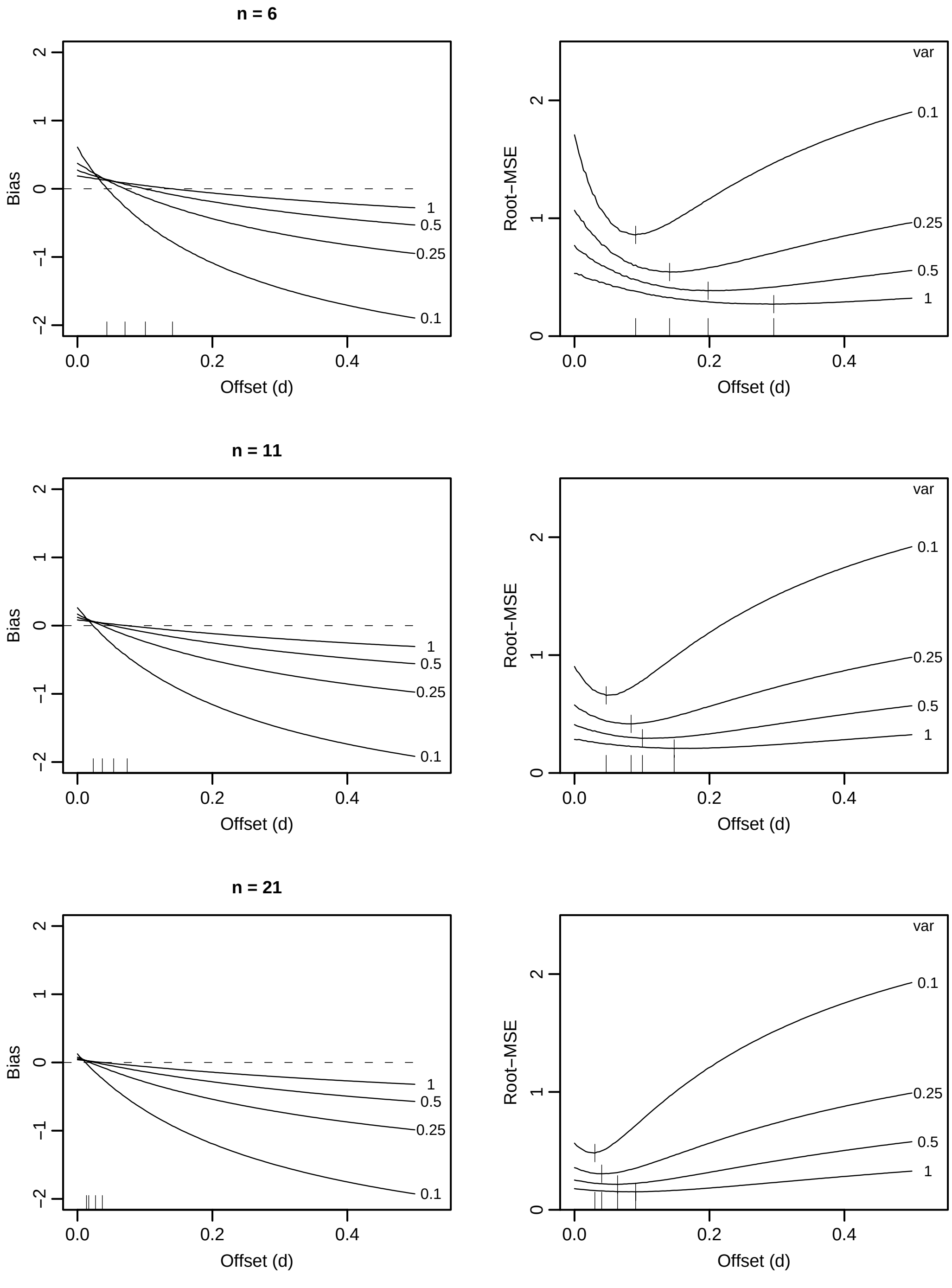


Figure 4: *The bias and root-MSE of the estimator* $1/(d+\hat{\sigma})$ *of* $1/\sigma$ *as functions of the offset* d*, with* $n = 6, 11$ *and 21 (*$n-1$ *degrees of freedom). Based on 50 000 replications.*

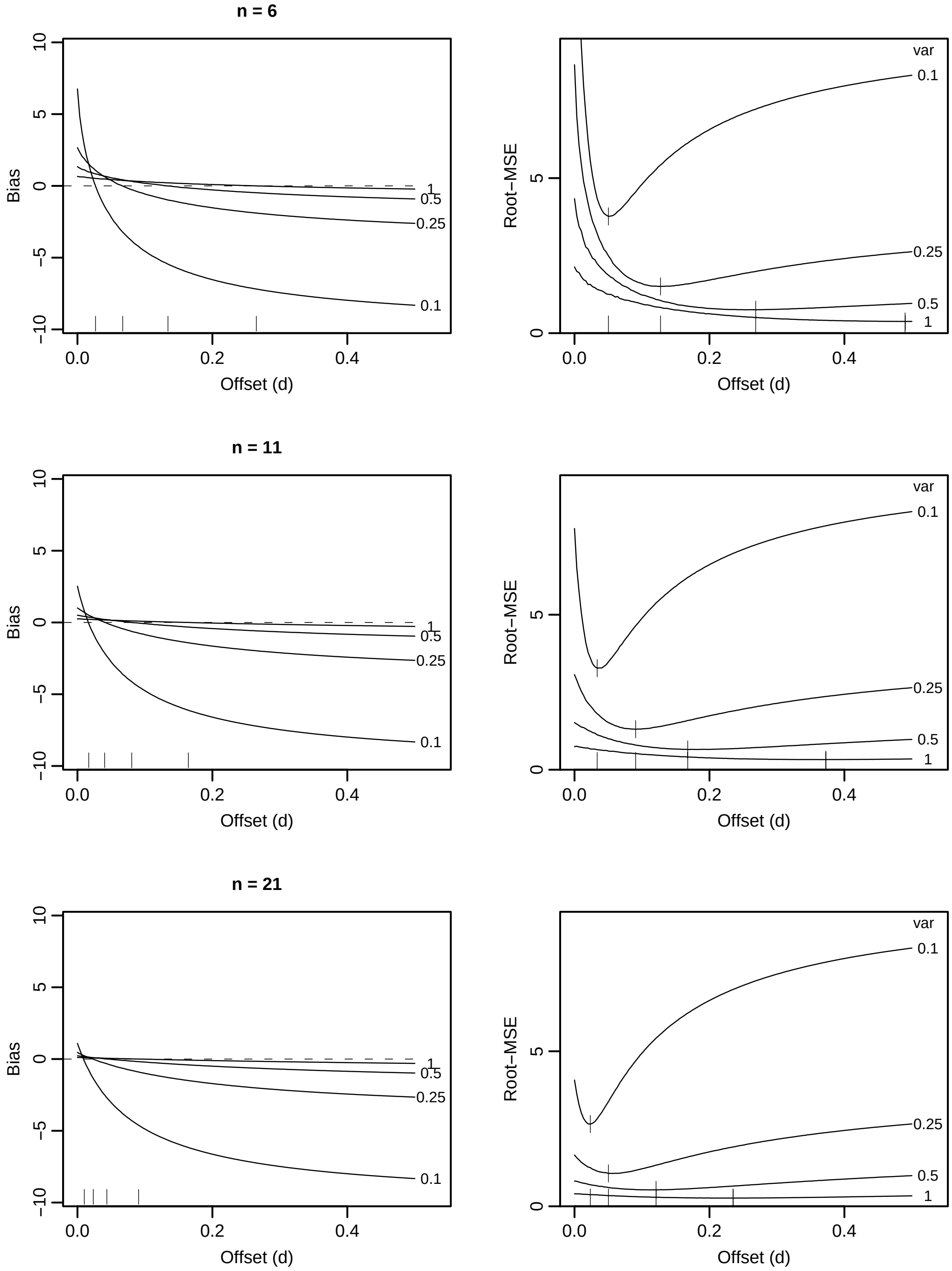


Figure 5: *The bias and root-MSE of the estimator* $1/(d+\hat{\sigma}^2)$ *of* $1/\sigma^2$ *as functions of the offset* d*, with* $n = 6, 11$ *and 21. Based on 50 000 replications.*

Table 1: *Properties of the alternative estimators of the reciprocal standard deviation* $1/\sigma$ *for sample sizes* $n = 6, 11$ *and 21;* $\sigma^2 = 1$.

		Bias			*root-MSE*		
		Sample size			*Sample size*		
Estimator		6	11	21	6	11	21
Naive	$\frac{1}{\hat{\sigma}}$	0.189	0.084	0.040	0.537	0.287	0.179
Unbiased	$\frac{1}{V_n\hat{\sigma}}$	0.000	0.000	0.000	0.633	0.392	0.262
Efficient in $\frac{c}{\hat{\sigma}}$	$\frac{V_n}{\hat{\sigma}}\,\frac{n-3}{n-1}$	-0.151	-0.060	-0.027	0.389	0.246	0.165
Efficient in $\frac{1}{d+\hat{\sigma}}$	$\frac{1}{d^*+\hat{\sigma}}$	-0.124	-0.072	-0.052	0.280	0.208	0.153

When $\sigma^2 = 1.0$, these ranges are $0.084-0.690$, $0.045-0.312$ and $0.025-0.151$ for the respective sample sizes $n = 6, 11$ and 21. When $n = 21$, the optimal offset for $\sigma^2 = 0.1$ is $d^* = 0.030$. Therefore, when σ^2 is in fact equal to 1.0, but we base the value of d^* erroneously on $\sigma^2 = 0.1$, we still obtain an estimator that is more efficient than $c^*/\hat{\sigma}$.

Estimators of the precision $1/\sigma^2$ can be assessed similarly. The offset estimator $1/(d^* + \hat{\sigma}^2)$ is more efficient than $c^*/\hat{\sigma}^2$ even when $n = 21$ (root-MSEs 0.270 *versus* 0.333), but the largest error that we can afford in estimating or guessing the value of d^* is much smaller than for estimating $1/\sigma$. For example, the estimators $1/(d + \hat{\sigma}^2)$ are more efficient than $c^*/\hat{\sigma}^2$ for $0.064 < d < 0.490$. The ideal offset when $\sigma^2 = 0.25$ is $d^\dagger = 0.057$, outside this range, so $1/(d^\dagger + \hat{\sigma}^2)$ is less efficient than $c^*/\hat{\sigma}^2$. In contrast, for $\sigma^2 = 0.30$ we have $d^\dagger = 0.072$, so the offset estimator is more efficient than $c^*/\hat{\sigma}^2$. The gains in efficiency by using offset d in $1/(d + \hat{\sigma}^2)$ to estimate $1/\sigma^2$ are in general not as great as by using $1/(d + \hat{\sigma})$ for estimating $1/\sigma$.

We explored estimators $1/(d + \hat{\sigma})^2$, but found them uniformly less efficient than $1/(d + \hat{\sigma}^2)$. Estimators in the class $1/(d + c\hat{\sigma}^2)$ would be more efficient if the constants c and d were set optimally. However, having to set (or estimate) two constants is likely to be too difficult a task in most settings.

3. Decision about σ^2 with prior information

To estimate σ^2 better than by $\hat{\sigma}^2$, we draw on the prior information in a Bayesian approach. We want to cater for the setting in which no party could be called upon to declare a single prior distribution for σ^2. An expert (client) may not be available at all, the process of elicitation may reach an impasse, or the expert might feel uncomfortable with the declaration of any single prior because some similar priors might equally well be declared, and yet they would lead to appreciably different posterior distributions. See Garthwaite, Kadane and O'Hagan (2005) for a review of methods of elicitation and related issues, such as the uncertainty about the prior. Given the strong impact of a prior on the posterior distribution, and the substantial uncertainty about the prior, drawing any conclusions from the details of any particular posterior distribution is poorly justified. We therefore focus on the tails of the posterior in the context of the problem with a discrete choice. Suppose we have two options, actions A and B; A is preferred when $\sigma^2 < \sigma^2_{\mathrm{R}}$ and B is preferred otherwise. The *reference* variance σ^2_{R} is given. If there is no obvious value of σ^2_{R}, the method described below can be applied to a small number of distinct values of σ^2_{R}.

We consider only the inverse gamma distributions as possible priors for σ^2; their densities are

$$f(s) = \frac{1}{\Gamma(r)} \theta^r s^{-r-1} \exp\left(-\frac{\theta}{s}\right) , \tag{2}$$

where $\theta > 0$ and $r > 0$ are parameters, called the shape and inverse scale, respectively. We regard this class of distributions as sufficiently rich for representing the prior information. Convenience is an important factor in this choice; inverse gamma is the conjugate distribution for (scaled) χ^2, the distribution of the estimator $\hat{\sigma}^2$. The expectation of the inverse gamma is $\theta/(r-1)$, so long as $r > 1$, and its variance is $\theta^2/\{(r-1)^2(r-2)\}$, so long as $r > 2$.

We prefer the parametrisation in terms of the precision $\tau = 1/\sigma^2$, the double-shape $q = 2r$ and the scale $\lambda = 2\theta/q$, because it facilitates an easier interpretation and helps the client make the relevant choices regarding the prior distribution. The prior density for τ that corresponds to (2) is the gamma

$$f(\tau) = \frac{1}{\Gamma\left(\frac{q}{2}\right)} \left(\frac{q\lambda}{2}\right)^{\frac{1}{2}q} \tau^{\frac{1}{2}q-1} \exp\left(-\frac{q\lambda\tau}{2}\right) .$$

The posterior density of τ is

$$f(t \,|\, \hat{\sigma}^2 = y) = C(k,q,\lambda)\, t^{\frac{1}{2}(k+q)-1} \exp\left\{-\frac{t}{2}(ky+q\lambda)\right\} ,$$

where C is the normalising constant. The corresponding distribution is scaled χ^2 with $k+q$ degrees of freedom and the scaling $ky+q\lambda$. In the standard Bayesian approach, all inferential statements about σ^2 are based on this distribution. For example, its expectation $(ky+q\lambda)/(k+q)$ may be quoted as an estimate, and its variance as a measure of uncertainty about σ^2, akin to the (frequentist) sampling variance.

In the frequentist perspective, the impact of the prior on the posterior is equivalent to adding q degrees of freedom (random draws from $\mathcal{N}(\mu, \sigma^2)$ or elementary observations) with a contribution of λ per degree of freedom to the corrected sum of squares, increasing it from $k\hat{\sigma}^2$ to $k\hat{\sigma}^2 + q\lambda$. (Of course, we have to overlook that q may be fractional.) We can regard $\hat{\sigma}^2$ and λ as two independent (elementary) estimators of σ^2. Then the posterior expectation is a composite estimator of σ^2; it combines the two elementary estimators with weights proportional to the associated degrees of freedom.

3.1. Borderline priors

Without being able or willing to commit ourselves to a single prior when the prior would have a strong impact on the posterior distribution, it is not feasible or meaningful to study the entire posterior. Instead, we focus on the tails of the posterior, addressing the concern that the variance σ^2 may be greater (or smaller) than an *a priori* set reference value σ^2_{R}. We can motivate this by adopting the following decision rule. If σ^2_{R} lies in the 100α% right-hand tail of the posterior distribution for σ^2, that is,

$$\mathrm{P}\left(\sigma^2 > \sigma^2_{\mathrm{R}} \,|\, \hat{\sigma}^2\right) < \alpha \,,$$

for a given probability α, we take action A; otherwise we take action B. This is similar to a Bayesian version of hypothesis testing, although we treat the two actions symmetrically and consider both very small and very large values of α (e.g., 0.05 and 0.95).

We want to cater for settings in which the process of elicitation has not been concluded with a single prior (or has not taken place at all), but a set of plausible priors has been agreed (or was declared by the analyst). Such a set may be a rectangle given by the ranges $\lambda \in (\lambda_{\mathrm{L}}, \lambda_{\mathrm{H}})$ and $q \in (q_{\mathrm{L}}, q_{\mathrm{H}})$, or, more generally, a convex set in the parameter space for (λ, q). Since there is no single (prior) distribution that faithfully reflects the prior information, we invert the standard Bayesian solution and seek priors that would yield the so-called *borderline* posteriors. These are posteriors for which the $100(1-\alpha)$ percentile is equal to σ^2_{R}. The corresponding priors are also called borderline.

For a given value of the (prior) parameter q, the borderline value of λ, for which (q, λ) defines a borderline prior, is given by the equation

$$\sigma^2_{\mathrm{R}} \frac{(k+q)^2}{k\hat{\sigma}^2 + q\lambda} = F^{-1}_{k+q}(1-\alpha) \,,$$

in which F_h is the distribution function (and F_h^{-1} the quantile function) of the χ^2 distribution with h degrees of freedom. The solution,

$$\lambda_{\mathrm{B}}(q) = \frac{1}{q} \left\{ \frac{(k+q)^2 \sigma^2_{\mathrm{R}}}{F^{-1}_{k+q}(1-\alpha)} - k\hat{\sigma}^2 \right\} \,,$$

is unique, although λ_{B} may be negative for some values of q. For given α and k, $\lambda_{\mathrm{B}}(q)$ is positive for small $q > 0$ when $\hat{\sigma}^2 < k\sigma^2_{\mathrm{R}}/F_k^{-1}(1-\alpha)$.

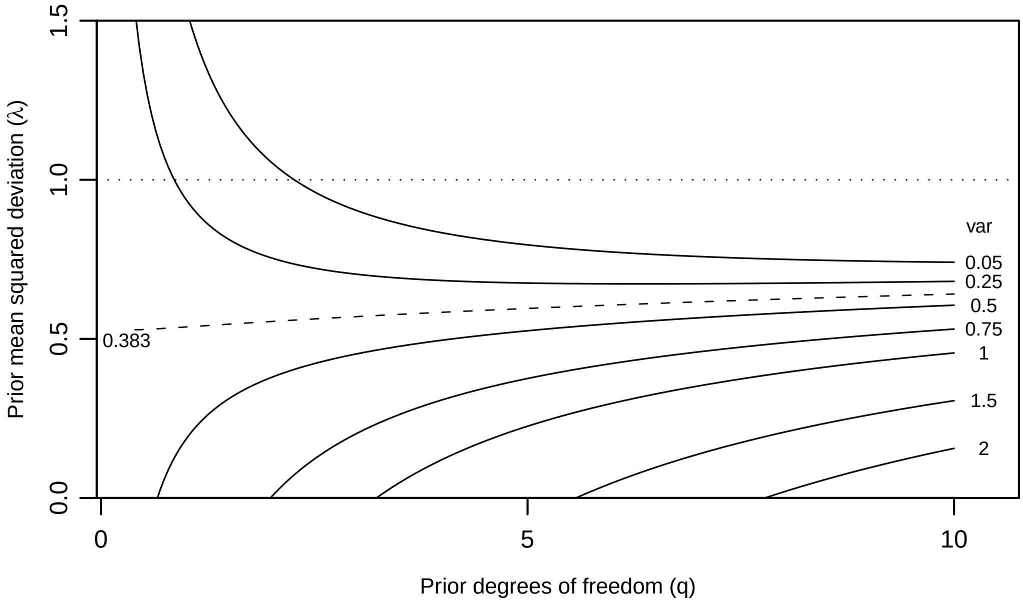


Figure 6: *Borderline priors for the setting with* $k = 3$ *degrees of freedom, the reference variance* $\sigma_{\mathrm{R}}^2 = 1$ *and* $\alpha = 0.05$*. The threshold borderline function, for* $\hat{\sigma}^2 = 0.383$ *is drawn by dashes. The values of* σ^2 *are indicated at the right-hand margin.*

A set of borderline functions $\lambda_{\mathrm{B}}(q)$ is drawn in Figure 6 for $k = 3$, $\sigma_{\mathrm{R}}^2 = 1$, $\alpha = 0.05$ and values of $\hat{\sigma}^2$ indicated at the right-hand margin. The functions are positive for all $q > 0$ when $\hat{\sigma}^2 \leq 0.383$; this threshold value of $\hat{\sigma}^2$ is found by a unidimensional search. All the functions converge to 1.0 as $q \rightarrow +\infty$, but the convergence, of the order $O(1/\sqrt{q})$, is very slow. When $\hat{\sigma}^2 < 0.383$, $\lambda_{\mathrm{B}}(q)$ attains very large values for small q, so that $q\lambda_{\mathrm{B}}(q)$ would make a nontrivial contribution to the posterior expectation $k\hat{\sigma}^2 + q\lambda_{\mathrm{B}}(q)$. When $\hat{\sigma}^2 < 0.383$, very small q is associated with large $\lambda_{\mathrm{B}}(q)$ because the prior contains very little information in relation to the data-based estimator $\hat{\sigma}^2$.

The borderline functions for the complemetary setting, with $\alpha = 0.95$, $k = 3$ and $\sigma_{\mathrm{R}}^2 = 1$, are displayed in Figure 7. For $\hat{\sigma}^2 \in (7.27, 8.53)$, $\lambda_{\mathrm{B}}(0)$ is positive and yet $\lambda_{\mathrm{B}}(q) < 0$ for some positive values of q. For instance, when $\hat{\sigma}^2 = 8.0$, $\lambda_{\mathrm{B}}(q) < 0$ for $q \in (0.36, 5.60)$. By way of an example, suppose $\hat{\sigma}^2 = 6$ with $k = 3$, the prior parameter q is in the range $(3, 5)$ and the prior value of λ does not exceed 0.8 (the shaded box in Figure 7). Then the entire set of plausible prior parameter vectors (q, λ) lies under the borderline function $\lambda_{\mathrm{B}}(q)$, and therefore action A is preferred for every plausible prior; we do not have to hone in on the prior.

If $\hat{\sigma}^2 = 8$, any prior with $q \in (3, 5)$ is located above the borderline function, so action B is preferred. Note that it is not sufficient for both the prior λ and the estimate $\hat{\sigma}^2$ to be smaller than the reference σ_{R}^2 to conclude with preference for small σ^2, because both sources of information are associated with a lot of uncertainty.

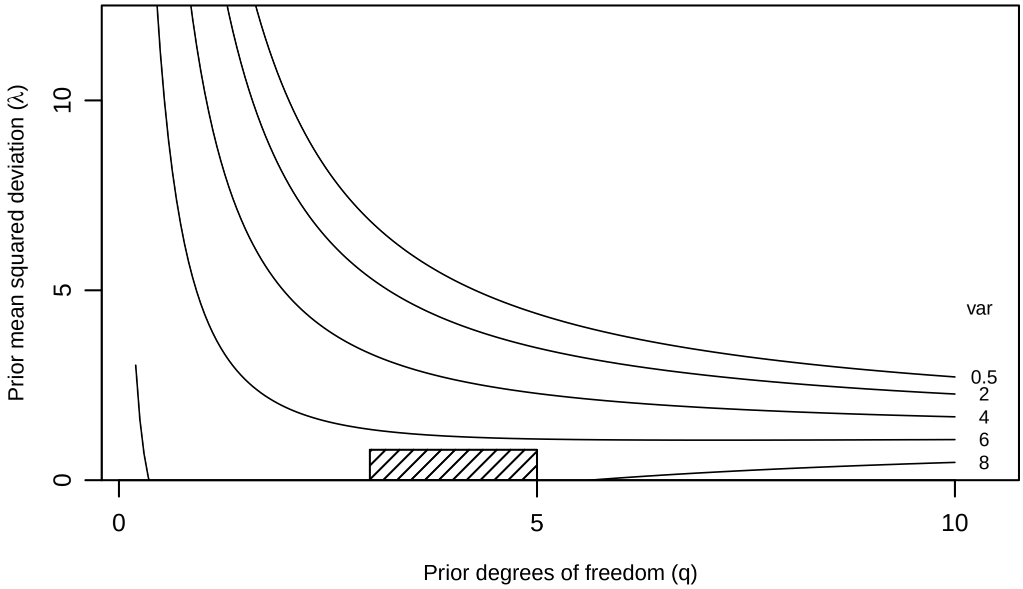


Figure 7: *Borderline priors for the setting with* $k = 3$ *degrees of freedom, the reference variance* $\sigma_{\mathrm{R}}^2 = 1$ *and* $\alpha = 0.95$*. The shaded box represents a set of plausible priors.*

The borderline function divides the space of the prior parameters (q, λ) into the subsets that correspond to the priors which lead to the two decisions. A prior represented by a point under the curve corresponds to preference for values of σ^2 smaller than the reference, and (q, λ) above the curve to preference for values of σ^2 greater than σ_{R}^2. After being presented the borderline curve, an expert (client) has to decide whether any of the borderline priors are plausible. If none are, and the plausible prior parameter vectors (q, λ) are all above (or all below) the curve, we have an unequivocal decision. The advantage of this approach is that we do not have to force the elicitation process to yield a single prior. It suffices to specify a set of plausible priors. Such a set would be non-convex only in some esoteric settings, and it is hard to envisage even a setting in which it would not be a rectangle in (q, λ) or in a different parametrisation. If the borderline curve intersects this plausible set, we cannot choose between the two actions, because for some plausible priors action A, and for others action B, is preferred. There is, therefore, an incentive to reduce the set of plausible priors as much as possible, but not necessarily to a single point, as is required in the standard Bayesian setup.

A single prior has to originate from an expert. This is a serious stumbling block in any secondary analysis when the expert is not available for the necessary dialogue. Also, the expert may not be willing to commit him- or herself to a single prior. The analyst should proceed with the elicitation only as far as it is constructive. While the declaration of a single prior by an analyst on behalf of the client may be rather presumptious, the declaration of a plausible set of priors maintains the integrity of the analysis if this set reflects the analyst's view of what a (real or hypothetical) client's prior may be. In essence, a solution is sought for every prior that the analyst believes the (absent) expert might choose. We do not want to integrate the posteriors over the plausible priors to obtain a single posterior distribution (Gelman *et al.* 2003), because that corresponds to using a (single) prior when some other priors are also plausible.

The dialogue with the expert is simplified by using a parametrisation for the priors that is easy to interpret. Thus, first we settle on the range of plausible prior degrees of freedom q (the strength or extent of prior information), and then on λ (the range of prior magnitudes of σ^2). This leads to a rectangle of plausible priors that may be reviewed further. The reference variance σ^2_{R} is set to reflect the client's priorities; when there is no clear candidate value, the problem may be solved for several references. The tail probabilities are usually set by convention, motivated by the practice of hypothesis testing.

3.2. Equilibrium priors

A drawback of the analysis with the borderline priors is that the consequences of the errors of the two kinds, choosing one action when the other would be appropriate, are ignored. To adapt the analysis, we have to specify the losses associated with such errors. Suppose the gain when we correctly conclude that $\sigma^2 > \sigma^2_{\mathrm{R}}$ (take the right action B) is greater than correctly concluding that $\sigma^2 < \sigma^2_{\mathrm{R}}$ by $|\,\sigma^2 - \sigma^2_{\mathrm{R}}\,|$, and the loss when we incorrectly conclude that $\sigma^2 > \sigma^2_{\mathrm{R}}$ (action B instead of A) is greater than incorrectly concluding that $\sigma^2 < \sigma^2_{\mathrm{R}}$ by $\rho\,|\,\sigma^2_{\mathrm{R}} - \sigma^2\,|$. The positive constant ρ is called the *penalty ratio*. Denote the posterior mean $\hat{\sigma}^2_{\mathrm{post}} = (k\hat{\sigma}^2 + q\lambda)/(k+q)$. The posterior density of σ^2 is $f_{k+q}\{(k+q)z/\hat{\sigma}^2_{\mathrm{post}}\}(k+q)/\hat{\sigma}^2_{\mathrm{post}}$, where f_h is the density of the χ^2 distribution with h degrees of freedom.

Our objective is to find the sign of the expected gain

$$\begin{aligned} & \int_0^{\sigma^2_{\mathrm{R}}} f_{k+q}\left\{\frac{(k+q)\,z}{\hat{\sigma}^2_{\mathrm{post}}}\right\}\frac{k+q}{\hat{\sigma}^2_{\mathrm{post}}}\left(\sigma^2_{\mathrm{R}} - z\right)\mathrm{d}z \\ & -\rho\int_{\sigma^2_{\mathrm{R}}}^{+\infty} f_{k+q}\left\{\frac{(k+q)\,z}{\hat{\sigma}^2_{\mathrm{post}}}\right\}\frac{k+q}{\hat{\sigma}^2_{\mathrm{post}}}\left(z - \sigma^2_{\mathrm{R}}\right)\mathrm{d}z \\ = \;& (1-\rho)\sigma^2_{\mathrm{R}}F_{k+q}\left\{\frac{(k+q)\,\sigma^2_{\mathrm{R}}}{\hat{\sigma}^2_{\mathrm{post}}}\right\} - (1-\rho)\hat{\sigma}^2_{\mathrm{post}}F_{k+q+1}\left\{\frac{(k+q)\,\sigma^2_{\mathrm{R}}}{\hat{\sigma}^2_{\mathrm{post}}}\right\} \\ & +\rho\left(\sigma^2_{\mathrm{R}} - \hat{\sigma}^2_{\mathrm{post}}\right), \end{aligned} \qquad (3)$$

derived using the identity $u\,f_h(u) = h\,f_{h+1}(u)$ for any positive h and u.

An expression similar to (3) can be derived for the loss functions that are piecewise linear in τ. That is, suppose the claim that $\sigma^2 > \sigma^2_{\mathrm{R}}$ $(= 1/\tau_{\mathrm{R}})$, when it is correct, is associated with the gain $\tau_{\rm R} - \tau$, and when it is incorrect, with the loss $\rho(\tau - \tau_{\rm R})$. Then the expected gain is

$$(\rho-1)\tau_{\rm R} F_{k+q}\left(\frac{k+q}{\hat{\sigma}^2_{\rm post}\tau_{\rm R}}\right) - \frac{k+q}{k+q-1}\,\frac{\rho-1}{\hat{\sigma}^2_{\rm post}} F_{k+q-1}\left(\frac{k+q}{\hat{\sigma}^2_{\rm post}\tau_{\rm R}}\right) + \rho\left(\tau - \frac{k+q}{k+q-1}\,\frac{1}{\hat{\sigma}^2_{\rm post}}\right), \qquad (4)$$

so long as $k+q > 1$.

A prior or posterior is called *equilibrium* if the corresponding expected gain is equal to zero. In parallel with the borderline priors, we can represent the equilibrium priors as a function $\lambda^{(0)}(q)$, and discuss whether any of these priors are plausible. If all the plausible priors lie beneath this function, then action A, appropriate when $\sigma^2 < \sigma^2_{\rm R}$, is associated with a positive expected gain; if all the plausible priors are above the function, then action B ($\sigma^2 > \sigma^2_{\rm R}$) is associated with positive expected gain for every plausible prior.

For a given q we find the corresponding equilibrium value of $\lambda^{(0)}(q)$ by the Newton method. Since λ is involved in (3) and (4) only via $\hat{\sigma}^2_{\rm post}$, we can find the 'equilibrium' value of $\hat{\sigma}^2_{\rm post}$, denoted by $\hat{\sigma}^2_{\rm equi}$, and then evaluate $\lambda^{(0)}(q)$ as $\{(k+q)\hat{\sigma}^2_{\rm equi} - k\hat{\sigma}^2\}/q = \hat{\sigma}^2_{\rm equi} + k(\hat{\sigma}^2_{\rm equi} - \hat{\sigma}^2)/q$.

Figure 8 displays the equilibrium function $\lambda^{(0)}$ for the setting with $k = 3$, $\hat{\sigma}^2 = 0.25$, $\rho = 20$ (solid line) and $\rho = 5$ (dashes), and the expected gain given by (3). The shaded

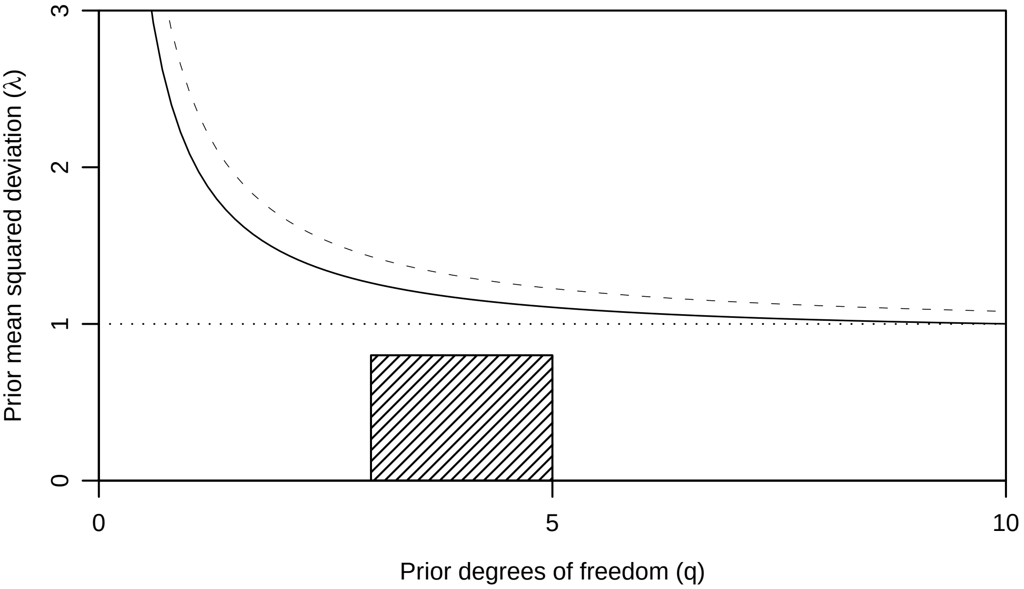


Figure 8: *Equilibrium priors for the setting with $k = 3$ degrees of freedom, the reference variance $\sigma^2_{\rm R} = 1$ and penalty ratios $\rho = 20$ (solid line) and $\rho = 5$ (dashes). The shaded box represents a set of plausible priors.*

box represents a set of plausible priors ($3 < q < 5$ and $0 < \lambda < 0.8$). Since it lies entirely beneath the equilibrium function, it yields an unequivocal conclusion, to take action A, because the expected gain is positive irrespective of which (plausible) prior is a faithful reflection of the prior information. The equilibrium functions in Figure 8 are decreasing in the range $q \in (0,10)$, and they converge to the reference probability $\sigma_{\mathrm{R}}^2 = 1$ as $q \rightarrow +\infty$. However, they are not monotone in $(0,+\infty)$; their values dip under $\sigma_{\mathrm{R}}^2 = 1$. For example, with $\rho = 10$, $\lambda^{(0)}(q)$ attains its minimum of 0.95 at $q \doteq 40$, and with $\rho = 50$ it attains its minimum of 0.78 at $q \doteq 170$.

An aplication of borderline and equilibrium priors in a different small-sample setting is presented in Longford (2009).

4. Conclusion

We explored several alternatives to the established (unbiased) estimator $\hat{\sigma}^2$ of the variance σ^2 in the standard setting of a random sample of small size from $\mathcal{N}(\mu,\sigma^2)$. We demonstrated that estimators of the form $c\hat{\sigma}^2$, and their transformations $g(c\hat{\sigma}^2)$, are superior to $g(\hat{\sigma}^2)$ for functions g equal to the identity and square root, and their reciprocals. The optimal constants c^* are specific to the transformations, but do not depend on σ^2. For the reciprocals, an offset can be applied, as in $1/(d+\hat{\sigma}^2)$ for $d > 0$. The optimal value of d depends on σ^2, but a modicum of error in the value of σ^2 used for the offset $d = d(\sigma^2)$ is tolerated without a substantial loss of efficiency or loss of the superiority over the optimal estimator $c^*/\hat{\sigma}$ or $c^*/\hat{\sigma}^2$.

We introduced the (Bayes) borderline and equilibrium priors for the variance σ^2. Although they require additional specification, a reference variance (σ_{R}^2) and a tail probability (α) or a loss function (penalty ratio), they choose among the two actions optimally with respect to these specifications. Instead of the standard setting in which a single prior is required, it suffices to specify a (convex) set of (plausible) priors. The analysis avoids an impasse and can maintain its integrity when the process of elicitation fails to conclude with a single prior, or when it does not take place at all. However, specifying a smaller set of plausible priors is advantageous because it is less likely to straddle the borderline or equilibrium curve, when the solution (the decision) is not unequivocal. An outstanding challenge is to combine the advantages of the offset and prior information.

Although a suitable (near-optimal) offset d is found by simulations and the borderline or equilibrium curves are found by iterations, only a modest amount of computing is involved (a few minutes of CPU time for all the simulations). The software developed in R is available from the author on request.

Acknowledgements

Research for and preparation of this article were partly supported by the Spanish Ministry of Science and Technology through Grant SEJ2006–13537. Careful reading of the manuscript and helpful suggestions from Jordi Serra i Solanich and two anonymous referees are acknowledged.

References

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100, 680-700.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, 2nd Ed. Chapman & Hall, CRC, New York.

Longford, N. T. (2009). Analysis of all-zero binomial outcomes with borderline and equilibrium priors. *Journal of Applied Statistics*, 36, 1259-1265.

Longford, N. T. (2010). Estimation of the effect size in meta-analysis with few studies. *Statistics in Medicine*, 29, 421-430.

Markowitz, E. (1968). Minimum mean-square-error of estimation of the standard deviation of the normal distribution. *The American Statistician*, 22, 26.

Stuart A. (1969). Reduced mean-square-error estimation of σ^p in normal samples. *The American Statistician*, 23, 27-28.

Stuart, A., and Ord, K. (1994). *Kendall's Advanced Theory of Statistics*, 6th Ed. Volume I. Distribution Theory. Edward Arnold, London.

Sutton, A. J., Jones, D. R., Abrams, K. R., Sheldon, T. A., and Song, F. (2000). *Methods for Meta-analysis in Medical Research*. Wiley, London, UK.

SORT 34 (1) January-June 2010, 21-44

Variance reduction technique for calculating value at risk in fixed income portfolios

Pilar Abad* and Sonia Benito**

Abstract

Financial institutions and regulators increasingly use Value at Risk (VaR) as a standard measure for market risk. Thus, a growing amount of innovative VaR methodologies is being developed by researchers in order to improve the performance of traditional techniques. A variance-covariance approach for fixed income portfolios requires an estimate of the variance-covariance matrix of the interest rates that determine its value. We propose an innovative methodology to simplify the calculation of this matrix. Specifically, we assume the underlying interest rates parameterization found in the model proposed by Nelson and Siegel (1987) to estimate the yield curve. As this paper shows, our VaR calculating methodology provides a more accurate measure of risk compared to other parametric methods.

MSC: 62P05, 91B30

Keywords: Value at Risk, Fixed Income Portfolios, Market risk, Nelson and Siegel model

1. Introduction

One of the most important tasks financial institutions face is evaluating their market risk exposure. This risk is a consequence of changes in the market prices of the assets in their portfolios. A possible way to measure this risk is to evaluate likely losses taking place by means of market price changes, which is what Value at Risk (VaR) methodology does. This methodology has been extensively used in recent times and it has become a basic market risk management tool for financial institutions and regulators.

The VaR of a portfolio is a statistical measure which tells us the maximum amount that an investor may lose over a given time horizon and a probability. Although VaR is

* *Address correspondence to:* Pilar Abad. Departamento de Fundamentos del Análisis Económico. Fac. de Ciencias Jurídicas y Sociales. Universidad Rey Juan Carlos. Campus de Vicálvaro. Paseo Artilleros s/n. 28032-Vicálvaro-Madrid. E-mail: pilar.abad@urjc.es

** Universidad Nacional de Educación a Distancia.

Received: October 2008
Accepted: February 2010

a simple concept, its calculation is not trivial. Formally, VaR(α%) is the percentile α of the probability distribution of changes in the value of a portfolio, i.e.: the value for which α% of the values lie to the left on the distribution. Consequentially, in order to calculate VaR, we first must estimate the probability distribution of the changes in the value of the portfolio. Several methods have been developed to estimate VaR of a portfolio. Among them, parametric methods or variance-covariance approaches, historical and Monte Carlo Simulations were initially proposed[1]. The literature on VaR has focused on two main directions: proposals for methodological innovations which aim to overcome limitations in some of the VaR methods and performance comparisons of VaR methods.

Along the first strand of literature, shortcomings in VaR methods have stimulated development of new methodologies. For example, in the case of the variance-covariance approach: distributions different from the Normal one have been considered [see, e.g., Mittnik *et al.* (2002), Kamdem (2005), Aas and Haff (2006) or Miller and Liu (2006)]; non-parametric distributions have been introduced [see, e.g., Cai (2002), Cakici and Foster (2003), Fan and Gu (2003), or Albanese *et al.* (2004)], or the extreme value theory has been applied to calculate the percentile of the tail distribution [see, e.g., McNeil and Frey (2000), or Brooks *et al.* (2005)]. Furthermore, in a parametric models framework the application of switching volatility models has been proposed [see, e.g., Billio and Pelizzon (2000) or Li and Lin (2004)] and variance reduction techniques which simplify calculations of the variance-covariance matrix needed to compute VaR under the parametric method, [see, e.g., Christiansen (1999), Alexander (2001) or Cabedo and Moya (2003)][2].

The results found in the existing literature regarding relative performances from different VaR models are somewhat inconclusive. No one model is better than others. Recent works include a wider range of methods (historical, Monte Carlo simulation, parametric methods including non-parametric distribution, and the extreme value theory), such as Bao *et al.* (2006), Consigli (2002) and Daníelsson (2002). They show that parametric models provide satisfactory results in stable periods but they are less satisfactory in periods of high volatility. Some further evidence in favour of parametric methods is provided in: Sarma *et al.* (2003) by comparing historical simulation with parametric methods; Daníelsson and Vries (2000) by including the extreme value theory in their analysis and Chong (2004) who uses parametric methods to estimate VaR and compares the Normal distribution against a Student-t distribution to find that VaR performs better under a Normal distribution.

Although consensus on the most accurate model to estimate VaR has not been reached, parametric methods are the most popular in financial practice, as indicated by many authors such as Chong (2004) or Sarma *et al.* (2003). Therefore, this study

[1] Linsmeier and Pearson (2000) discuss the advantages and disadvantages of the three methods for computing VaR.

[2] Other studies also propose variance reduction techniques for estimating VaR, which are used in the Monte Carlo Simulation method, e.g. Glasserman *et al.* (2000).

departs from a parametric or variance-covariance method and proposes a variance reduction technique for estimating VaR. Unlike Christiansen (1999), Alexander (2001) and Cabedo and Moya (2003), our proposal uses the parameterization of interest rates that underlies the model of Nelson and Siegel (1987) to estimate the yield curve.

The parametric approach, based on the assumption that changes in a portfolio's value follow a known distribution, only needs a priori calculation of the conditional variance from changes in the value of the portfolio. However, computing this variance is not a trivial exercise as a variance-covariance matrix for the portfolio assets needs to be estimated. Two types of problems are then involved: (1) a dimensionality problem and (2) a viability problem. The former is related to the large dimension of the matrix which complicates estimation. This is a more sensitive problem for fixed income portfolios where their value depends on a large number of interest rates with different maturities. The later problem stems from the complex task of estimating conditional covariances when sophisticated models such as multivariate GARCH models are used. The estimation of such models is very costly in terms of computation. These type of problems are usually overcome through use of multivariate analysis [Christiansen (1999), Alexander (2001) or Cabedo and Moya (2003)], which are based on the assumption that there are common factors in the volatility of the interest rates and that these same factors explain the changes in the temporal structure of interest rates (TSIR). Under these two assumptions, it turns out possible from a theoretical point of view to obtain, through a multivariate technique and at a low calculation cost, the variance-covariance matrix from a vector of interest rates.

This paper proposes an alternative method of estimating the variance-covariance matrix of interest rates at a low computational cost. No specific assumptions need to be stated to apply our technique. We depart from Nelson and Siegel (1987) model, initially developed to estimate the TSIR. This model gives an expression for interest rates as a function of four parameters. Therefore, we can obtain the interest rates variance-covariance matrix by calculating variances for only four variables – the principal components of the changes in the four parameters. Financial institutions and banks routinely compute the parameters we need from Nelson and Siegel's model for purposes other than VaR related. Consequently, estimated parameters are thus readily available as inputs to be used in a VaR estimation and do not represent an additional computational burden. This fact is an obvious advantage from our approach.

This paper is organized as follows. In section 2 we present our methodological proposal to estimate the variance-covariance matrix for a large vector of interest rates and at a low computational cost. The next three sections evaluate the proposed method for a Spanish market data sample. In section 3 we describe the data we use briefly before applying the method proposed in order to obtain the variance-covariance matrix of a vector of interest rates. In section 4 we evaluate the proposed methodology to calculate VaR for fixed income portfolios so that we can compare the results with those obtained from standard methods of calculation. Finally, section 5 presents the main conclusions from the paper.

2. A parametric model for estimating risk

In this section we present a methodology to calculate the variance-covariance matrix for a large vector of interest rates at a low computational cost. In order to do so we start with the model proposed by Nelson and Siegel (1987), originally designed to estimate the yield curve.

The Nelson and Siegel formulation specifies a parsimonious representation of the forward rate function given by:

$$\varphi_m^t = \beta_0 + \beta_1 e^{\left(-\frac{m}{\tau}\right)} + \beta_2 \frac{m}{\tau} e^{\left(-\frac{m}{\tau}\right)} \tag{1}$$

This expression allows us to accommodate various functional features such as level, slope sign or curve shape in relation to four parameters ($\beta_0,\, \beta_1, \beta_2, \tau$).

Bearing in mind the fact that the spot interest rate at maturity *m* can be expressed as the sum of the instantaneous forward interest rates from 0 up to *m*, that is, by integrating the expression that defines the instantaneous forward rate:

$$r_t(m) = \int_0^m \varphi_u^t d_u \tag{2}$$

we obtain the following expression for the spot interest rate at maturity *m*:

$$r_t(m) = \beta_0 - \beta_1 \frac{\tau}{m} e^{\left(-\frac{m}{\tau}\right)} + \beta_1 \frac{\tau}{m} + \beta_2 \frac{\tau}{m} - \beta_2 e^{\left(-\frac{m}{\tau}\right)} - \beta_2 \frac{\tau}{m} e^{\left(\frac{-m}{\tau}\right)} \tag{3}$$

Equation (3) shows that spot interest rates are a function of only four parameters. In accordance with this function, changes in these parameters are the variables that determine changes in the interest rates. By using a linear approximation we can estimate the change in the zero-coupon interest rate at maturity *m* from the following expression:

$$dr_t(m) \approx \left[\frac{\partial r_t(m)}{\partial \beta_0}, \frac{\partial r_t(m)}{\partial \beta_1}, \frac{\partial r_t(m)}{\partial \beta_2}, \frac{\partial r_t(m)}{\partial \tau} \right] \begin{bmatrix} d\beta_0, t \\ d\beta_1, t \\ d\beta_2, t \\ d\tau_t \end{bmatrix} \tag{4}$$

In a multivariate context, the changes in the vector of interest rates that make up the TSIR can be expressed by generalizing equation (4) in the following way:

$$dr_t = G_t d\beta_t + \varepsilon_t \tag{5}$$

where

$$dr_t = [dr_t(1), dr_t(2), \ldots, dr_t(k)], \qquad d\beta_t^{'} = [d\beta_{0,t}, d\beta_{1,t}, d\beta_{2,t}, d\tau_t],$$

$$G_t = \begin{bmatrix} \dfrac{\partial r_t(1)}{\partial \beta_0} & \dfrac{\partial r_t(1)}{\partial \beta_1} & \dfrac{\partial r_t(1)}{\partial \beta_2} & \dfrac{\partial r_t(1)}{\partial \tau} \\ \dfrac{\partial r_t(2)}{\partial \beta_0} & \dfrac{\partial r_t(2)}{\partial \beta_1} & \dfrac{\partial r_t(2)}{\partial \beta_2} & \dfrac{\partial r_t(2)}{\partial \tau} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial r_t(k)}{\partial \beta_0} & \dfrac{\partial r_t(k)}{\partial \beta_1} & \dfrac{\partial r_t(k)}{\partial \beta_2} & \dfrac{\partial r_t(k)}{\partial \tau} \end{bmatrix}$$

and ε_t is the errors vector.

From expression (5) we can calculate the variance-covariance matrix of a vector of changes in the k interest rates using the following expression:

$$var(dr_t) = G_t \Psi_t G_t^{'} + var(\varepsilon_t) \tag{6}$$

where:

$$\Psi_t = \begin{bmatrix} var(\beta_{0,t}) & cov(\beta_{0,t}\,\beta_{1,t}) & cov(\beta_{0,t}\,\beta_{2,t}) & cov(\beta_{0,t}\,\tau_t) \\ & var(\beta_{1,t}) & cov(\beta_{1,t}\,\beta_{2,t}) & cov(\beta_{1,t}\,\tau_t) \\ & & var(\beta_{2,t}) & cov(\beta_{2,t}\,\tau_t) \\ & & & var(\tau_t) \end{bmatrix}$$

At this point, it is worth noting that we have greatly simplified the dimension of the variance-covariance matrix we need to estimate. Instead of having to estimate $k(k+1)/2$ variances and covariances for a vector of k interest rates, we now only need to estimate 10 second order moments. Nevertheless, the problem associated with the difficulty of the covariances estimations persists.

However, by applying principal components to the vector of the changes in the parameters ($d\beta_t$), we can simplify the calculation of the variance-covariance matrix even further. Accordingly, the vector of changes in the parameters of Nelson and Siegel (1987) model can be expressed as:

$$d\beta_t = AF_t \tag{7}$$

$$F_t = \begin{bmatrix} f_{1,t} & f_{2,t} & f_{3,t} & f_{4,t} \end{bmatrix}$$

and

$$A = \begin{bmatrix} a^1_{\beta_0} & a^2_{\beta_0} & a^3_{\beta_0} & a^4_{\beta_0} \\ a^1_{\beta_1} & a^2_{\beta_1} & a^3_{\beta_1} & a^4_{\beta_1} \\ a^1_{\beta_2} & a^2_{\beta_2} & a^3_{\beta_2} & a^4_{\beta_2} \\ a^1_{\tau} & a^2_{\tau} & a^3_{\tau} & a^4_{\tau} \end{bmatrix}$$

where F_t is the principal components vector associated with the vector $d\beta_t$ and A is the constants matrix from the eigenvectors associated with each of the four eigenvalues for the variance-covariance matrix of changes in the parameters from Nelson and Siegel model ($d\beta_t$).

Substituting equation (7) into equation (5) and given that each principal component is orthogonal to the rest, we can express the interest rates variance-covariance matrix as follows:

$$var(dr_t) = G_t^* \Omega_t G_t^{*'} + var(\varepsilon_t) \tag{8}$$

where:

$$\Omega_t = \begin{bmatrix} var(f_{1,t}) & 0 & 0 & 0 \\ 0 & var(f_{2,t}) & 0 & 0 \\ 0 & 0 & var(f_{3,t}) & 0 \\ 0 & 0 & 0 & var(f_{4,t}) \end{bmatrix}$$

and $G_t^* \approx G_t \times A$

Ignoring $var(\varepsilon_t)$, let us approximate:

$$var(dr_t) \approx G_t^* \Omega_t G_t^{*'} \tag{9}$$

Therefore, equation (9) provides us an alternative method to estimate the variance-covariance matrix of changes in a k interest rates vector by using the four principal components estimation for changes in the parameters of Nelson and Siegel (1987) model. In this way, the dimensionality problem associated with the calculation of the covariance has finally been solved.

Note that $var(dr_t)$ will be positive semi-definite, but it may not be strictly positive definite unless $\varepsilon_t = 0$. Although Ω_t is positive definite because it is a diagonal matrix with positive elements, nothing guarantees that $G_t^* \, \Omega_t G_t^{*'}$ will be positive definite when $\varepsilon_t \neq 0$. If the covariance matrix is based on (9), we should ensure strictly positive definiteness through checking eigenvalues. However, it is reasonable to expect that approximation (9) will give a strict positive definite variance-covariance matrix if representation (5) is done with a high degree of accuracy.

In the following sections we evaluate this method, to calculate both the variance matrix of a vector of interest rates and VaR for fixed income portfolios.

3. Estimating the variance-covariance matrix

3.1. The data

With the purpose of examining the method proposed in this paper, we estimate a daily term structure of interest rates using the actual mean for daily prices of Treasury transactions. The original data set consists of daily observations from actual transactions in all bonds traded on the Spanish government debt market. The database for bonds traded on the secondary market of Treasury debt covers the period from January, 1st 2002 to December, 31th 2004. We use this daily database to estimate the daily term structure of interest rates. We fit Nelson and Siegel's (1987) exponential model for the estimation of the yield curve and minimise price errors weighted by duration. We work with daily data for interest rates at 1, 2,. . . , 15 year maturities.

3.2. The results

In this section we examine this new approach to variance-covariance matrix estimation. The first section begins by comparing the changes in estimated and observed interest rates. Changes in interest rates are modelled by equation (5) so that we can then compare them with observed ones.

We then proceed to estimate the variance-covariance matrix of a vector of 10 interest rates, using the methodology proposed in the previous section. We compare these estimations (Indirect Estimation) with those obtained through some common univariate procedures (Direct Estimation).

In both cases, direct and indirect estimation, we need a method for estimating variances and covariance. For the indirect estimation case, the estimation method gives us the variances of the four principal components of changes in the parameters in Nelson and Siegel model. Indeed, this enables us to obtain the interest rates variance-covariance matrix from equation (9).

We use two alternative measures of volatility to estimate the variance-covariance matrixes of interest rates changes and principal components variance: exponentially weighted moving average (EWMA) and Generalized Autoregressive Conditional Heteroskedasticity models (GARCH)[3].

(1) Under the first alternative, the variance-covariance matrix is estimated with RiskMetrics methodology as developed by J. P. Morgan (1995). RiskMetrics uses the so called exponentially weighted moving average (EWMA) method. Accordingly, the estimator for the variance is:

[3] GARCH models are standard in finance (see Ferenstein and Gasowski, 2004).

$$var(dx_t) = (1-\lambda)\sum_{j=0}^{N-1}\lambda^j(dx_{t-j}-\overline{dx})^2 \tag{10}$$

and the estimator for the covariance is:

$$cov(dx_t dy_t) = (1-\lambda)\sum_{j=0}^{N-1}\lambda^j(dx_{t-j}-\overline{dx})(dy_{t-j}-\overline{dy}) \tag{11}$$

J.P. Morgan uses the EWMA method to estimate VaR in their portfolios. For $\lambda = 0.94$ with $N = 20$, on a widely diversified international portfolio RiskMetrics produces the best back-testing results. Subsequently, we use both of these values in the paper.

Therefore, we obtain direct estimations of the interest rates variance-covariance matrix (D_EWMA) from equations (10) and (11) where x_t and y_t are interest rates at different maturities. For the case of indirect estimation of the variance-covariance matrix (I_EWMA), we use equation (10) to calculate the principal components variances (where x_t are now these principal components). Equation (9) gives us then the relevant matrix.

(2) The EWMA methodology currently used for RiskMetrics$^{\text{TM}}$ data is quite acceptable for calculating VaR measures. Alternatively, some authors suggest using variance-covariance matrices obtained from multivariate GARCH. Nevertheless, the large variance-covariance matrices used in VaR calculations could never be estimated directly by implementing a full multivariate GARCH model due to insurmountable, computational complexity. For this reason we only compute variances of interest rates changes with univariate GARCH models and avoid computation of the covariance[4].

Given that indirect estimation (I_GARCH) does not require the estimation of covariance, we estimate the principal components conditional variance from changes in Nelson and Siegel model's parameters by using univariate GARCH models.

In sub-section two, we compare alternative estimations for the variance-covariance matrix as described above. Comparisons are then summarised in Table 1.

Table 1: *Type of variance-covariance matrix estimation.*

		Type of variance models	
		EWMA	**GARCH**
Type of estimation	**Direct Estimation**	D-EWMA	D-GARCH*
	Indirect Estimation	I-EWMA	I-GARCH

* We have not estimated multivariate GARCH model because of the computational complexity are insurmountable, so that only present the result of the variances which have been estimated using univariate GARCH models. All GARCH models are Exponential GARCH (EGARCH) model (see Nelson, 1991).

[4] We use the most suitable model for each series. All of them are Exponential General Autoregressive Conditional Heteroskedastic (EGARCH) model (Nelson, 1991).

Note that estimating the variance-covariance matrix with the methodology proposed in this study (indirect estimation) involves a minimum calculation cost, since it is only necessary to estimate the variance of four variables (the principal components of daily changes in the parameters of the Nelson and Siegel model).

3.2.1. Comparing interest rates changes

Firstly, we have evaluated the capacity of the model that we propose here to estimate daily changes in an interest rates vector. We need to compare observed interest rates with their estimations from equation (5). In Figure 1 we show a scatter diagram relating observed changes with estimated changes in 1-year interest rates, the graph shows that they are closely related regardless of the maturity.

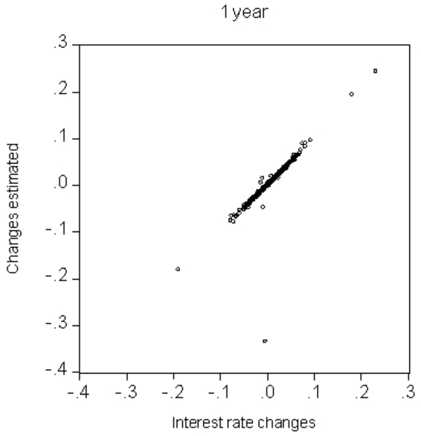


Figure 1: *Comparing the changes of 1-year interest rate observed with the estimated changes (equation (5)).*

In Table 2 we report some descriptive statistics for the interest rate estimation errors. The average error is less than a half basic point for all maturities i.e quite small. In relative terms, this error represents approximately 0.5% from the interest rates average. It is also worth noting that the average error and the standard deviation are very similar in all maturities therefore the model appears to be accurate for all maturities.

Table 2: *Estimation errors in interest rates. Descriptive statistics.*

	1-year	**2-year**	**3-year**	**4-year**	**5-year**	**6-year**	**7-year**	**8-year**	**9-year**	**10-year**
Mean[(a)]	0.2	0.3	0.4	0.5	0.4	0.4	0.4	0.3	0.3	0.3
Standard deviation	1.5	3.2	2.7	2.1	1.8	1.8	1.7	1.6	1.4	1.3
Maximum error	32.9	66.0	45.4	32.8	34.5	33.2	30.4	27.1	23.8	20.9
Minimum error	−4.1	−1.5	−0.1	−0.1	−0.2	−4.0	−7.8	−9.2	−9.4	−9.4

Note: Sample period is from 1/1/2002 to 12/31/2004 (510 observations). The errors are the difference between the observed interest rates and their estimations from equation (5). The errors (and all statistics) are expressed in basic points. [(a)] The average error is calculated in absolute value.

These results imply that when estimating changes in zero-coupon interest rates using equation (5) the error is virtually non-existent. In what follows, we evaluate the differences in the estimation of the variance-covariance matrix under various alternatives.

3.2.2. Comparing the estimations of variance-covariance matrix

In Figure 2 we show the conditional variances for the 1-year interest rate. We apply the exponentially weighted moving average method for direct and indirect: D_EWMA versus I_EWMA. Furthermore, in Figure 3 we show the direct estimation of the conditional variances for the interest rates using the GARCH (D_GARCH) models as well as an indirect estimation (I_GARCH). The variances estimated using the method proposed in this paper are very similar to the direct estimates for most of the maturities.

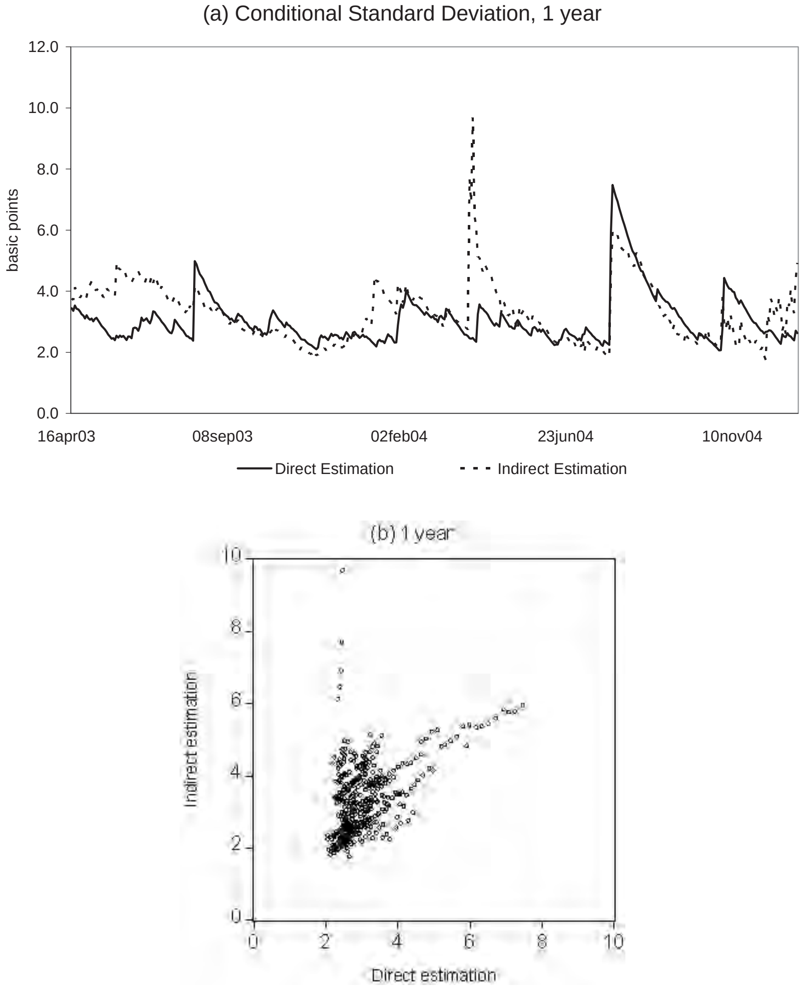
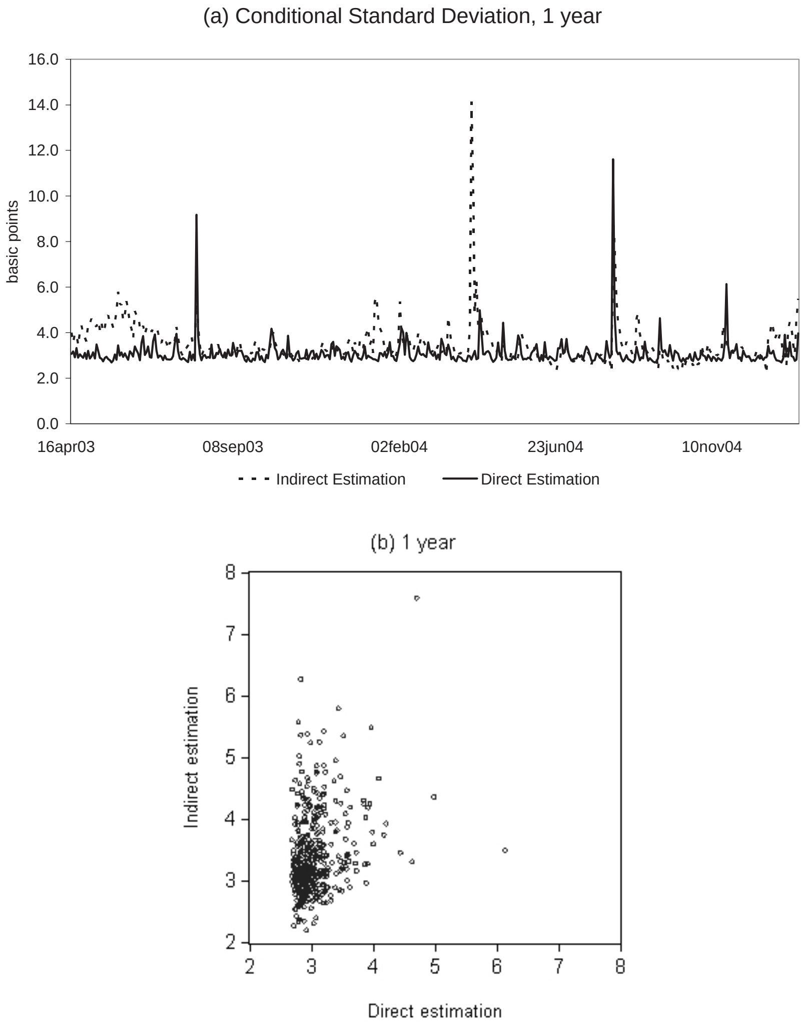


Figure 2: *Comparing the variance of changes of 1-year interest rate: Direct and indirect estimation using exponentially weighted moving average model.*



(a) Conditional Standard Deviation, 1 year

(b) 1 year

Figure 3: *Comparing the variance of changes of 1-year interest rate: Direct and indirect estimation using GARCH model.*

The descriptive statistics for the standard deviations differences estimated with both procedures are reported in Table 3. We compare the direct and indirect estimation methods using an EWMA model in panel (a), and using a GARCH model in panel (b). Panel (a) shows that the absolute value of the average differences for the EWMA specification, oscillates between 0.7 and 1.4 basic points. These average differences represent between 20% and 40% of the average of the estimated series. Panel (b) in Table 3 also shows that the average difference in absolute value for EGARCH specification is quite small. These differences are smaller than those of panel (a) taken as a percentage of the estimated conditional variance series. We can note that for both comparisons the range of differences for each pair of estimates is much wider for the 6-, 7- and 8-year interest rate than for the other maturities.

Table 3: *Differences in the estimation of the standard deviation on interest rates: direct vs. indirect method. Descriptive statistics.*

	1-year	**2-year**	**3-year**	**4-year**	**5-year**	**6-year**	**7-year**	**8-year**	**9-year**	**10-year**
	Panel (a): Comparing D_EWMA vs. I_EWMA									
Mean[(a)]	0.7	0.9	1.1	1.2	1.2	1.4	1.4	1.3	1.1	0.9
Standard deviation	0.9	1.7	2.0	1.8	1.7	1.8	1.9	1.8	1.5	1.2
Maximum error	1.6	4.4	6.6	3.9	1.4	1.3	1.2	1.1	1.0	0.8
Minimum error	−7.2	−19.7	−17.4	−12.9	−10.0	−8.6	−10.0	−10.0	−7.8	−8.0
	Panel (b): Comparing D_GARCH vs. I_GARCH									
Mean[(a)]	0.6	0.6	0.7	0.9	1.1	1.2	1.3	1.2	0.9	0.7
Standard deviation	0.9	1.2	1.2	1.3	1.5	1.6	1.7	1.7	1.6	1.5
Maximum error	3.6	4.8	3.9	2.3	1.4	1.1	0.9	0.7	0.4	0.4
Minimum error	−11.3	−18.0	−14.1	−12.0	−18.5	−22.5	−25.8	−27.0	−25.2	−23.4

Note: Sample period is from 1/1/2002 to 12/31/2004 (510 observations). I_EWMA indirect estimation (equation (9)) and D_EWMA direct estimation. RiskMetrics methodology (EWMA). I_GARCH: indirect estimation (equation (9)) and D_GARCH direct estimation. Conditional autoregressive volatility models (GARCH). [(a)] The average of the differences has been calculated in absolute value. Differences are measured in base points.

We now compare directly estimated covariances with those obtained with the procedure suggested in this paper. As above mentioned, given the extreme complexity of the GARCH multivariate model estimations, the direct estimation of the covariances is only approached with EWMA models.

Figure 4 shows estimated covariances for 3 and 1-year interest rates and for both procedures: D_EWMA versus I_EWMA. As it can be checked, estimated covariances behave similarly, although it should be noted that in most maturities there are greater differences than for the variances. In Table 4 we report some of the descriptive statistics of the estimated covariances. The average difference in absolute value is very small, between 0.0004 and 0.0014. However, this represents about 40% of the average estimated covariance.

Table 4: *Differences in the estimation of covariances of interest rates: direct vs. indirect method. Descriptive statistics.*

Comparing D_EWMA vs. I_EWMA

		1-year		**3-year**		**5-year**
	3-year	**5-year**	**10-year**	**5-year**	**10-year**	**10-year**
Mean[(a)]	0.0004	0.0005	0.0005	0.0013	0.0013	0.0014
Standard deviation	0.001	0.001	0.001	0.004	0.003	0.003
Maximum error	0.002	0.004	0.003	0.008	0.001	0.001
Minimum error	−0.022	−0.014	−0.003	−0.042	−0.018	−0.021

Note: Sample period is from 1/1/2002 to 12/31/2004 (510 observations). I_EWMA indirect estimation (equation (8)) and D_EWMA direct estimation. RiskMetrics methodology (EWMA). [(a)] The average difference is calculated in absolute value.

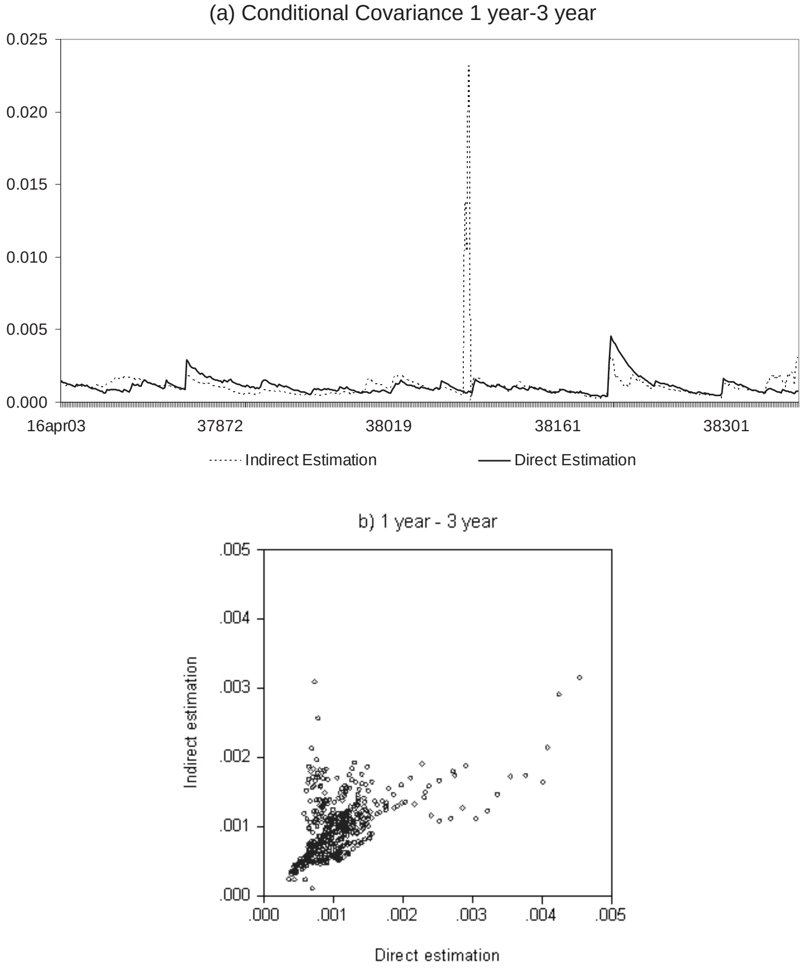


Figure 4: *Comparing the covariance between 3-year and 1-year interest rates: Direct and indirect estimation using exponentially weighted moving average (EWMA) model.*

In order to summarize the section we can conclude that we have shown how the procedure proposed in this paper to estimate the variance-covariance matrix of a large interest rates vector generates quite satisfactory results. In the following section we evaluate whether these small differences are important for risk management. Consequently, we apply the proposed methodology VaR calculation in several fixed income portfolios.

4. Estimating value at risk

In this section we evaluate the utility of our methodological proposal for risk management in fixed income portfolios. Thus, we create a parametric measure of VaR as an indicator of the risk of a given portfolio.

4.1. Value at risk

The VaR of a portfolio is a measure of the maximum loss that the portfolio may suffer over a given time horizon and with a given probability. Formally, the VaR measure is defined as the lower limit of the confidence interval of one tail:

$$\Pr\left[\Delta V_t(\tau) < VaR_t\right] = \alpha \tag{12}$$

where α is the level of confidence and $\Delta V_t(\tau)$is the change in the value of the portfolio over the time horizon τ.

The methods based on the parametric or variance-covariance approaches depart from the assumption that changes in the value of a portfolio follow a Normal distribution. Assuming that the average change is zero, the VaR for one day of portfolio j is obtained as:

$$VaR_{j,t}(\alpha\%) = \sigma_{t,dV_j} k_{\alpha\%} \tag{13}$$

where $k_{\alpha\%}$is the α percentile of the Standard Normal distribution, and the parameter that needs to be estimated is the standard deviation conditional of the value of portfolio j (σ_{t,dV_j}).

In a fixed income asset portfolio, duration can be used to obtain the variance of the value of portfolio j from the interest rates variance as shown in Jorion (2000):

$$\sigma^2_{t,dV_j} = D_{j,t}\Sigma_t D'_{j,t} \tag{14}$$

where Σ_t is the variance-covariance matrix of the interest rates and $D_{j,t}$ is the vector of the duration of portfolio j. This vector represents the sensitivity of the value of the portfolio to changes in the interest rates that determine its value.

In this section, VaR measures are calculated and compared. In the parametric approach, we use the estimations of the variance-covariance matrix as obtained in the previous section (see Table 1). Table 5 illustrates the four measures of VaR that we develop from the four variance-covariance models:

Table 5: *Type of VaR measures.*

	Type of variance-covariance matrix estimation	**Type o VaR measure**
Direct Estimation	D_EWMA	VaR_D_EWMA
	D_GARCH	VaR_D_GARCH[*]
Indirect Estimation	I_EWMA	VaR_I_EWMA
	I_GARCH	VaR_I_GARCH

[*] We did not compute VAR_D_GARCH because of the impossibility to estimate a multivariate GARCH model with 10 variables.

For the first VaR measure, VaR_D_EWMA, VaR is obtained by directly estimating Σ_t with an EWMA model. This is a popular approach to measuring market risk, used by JP Morgan (RiskMetricTM). The second VaR measure, VaR_D_GARCH, is also calculated by directly estimating the variance-covariance matrix, but using GARCH models to estimate the second order moments. Given that the large variance-covariance matrices used in VaR calculations could never be estimated directly using a full multivariate GARCH model, this VaR measure has not been calculated as computational complexity would be insurmountable.

Two final VaR measures are then computed by estimating the variance-covariance matrix of the interest rates following the procedure described in Section 2. We can estimate the variance-covariance matrix of interest rates indirectly, by substituting equation (9) into equation (14) to deduct a new expression for the variance of the changes in the value of the portfolio:

$$\sigma^2_{t,dV_j} = D_{j,t} G^* \Omega_t G^{*'} D'_{j,t} = D^m_{j,t} \Omega_t D^{m'}_{j,t} \tag{15}$$

In indirect estimation, $\Omega_t$is a diagonal matrix containing the conditional variance for the principal components of changes in the four parameters of Nelson and Siegel's model in its main diagonal. Also, $D^m_{j,t}$ is the modified vector of durations of portfolio j (with 1×4 dimension) which represents the sensitivity of the value of the portfolio to changes in the principal components of the four parameters in Nelson and Siegel model. In the VaR_I_EWMA, we use an EWMA model to estimate the variance of the principal components; and a suitable GARCH model to estimate these variances for the calculation of the VaR_I_GARCH measure.

4.2. The portfolios

In order to evaluate the procedure proposed in this paper for VaR calculation, we have considered 4 different portfolios made up of theoretical bonds with maturities at 3-, 5-, 10- and 15-years and constructed from real data from the Spanish debt market. The bond coupon is 3.0% in every portfolio. The period of analysis goes from April, 15th 2002 to December, 31st 2004, which allows us to perform 437 estimations of daily VaR for each portfolio.

In order to estimate the daily VaR we have assumed that the main portfolios features remain constant during the analysis period: the initial value of the portfolio, the maturity date and the coupon rate. In this case, results are comparable for the entire period of analysis since we avoid both the pull to par effect (the value of the bonds tends to par as the maturity date of the bond approaches) and the roll down effect (the volatility of the bond decreases over time).

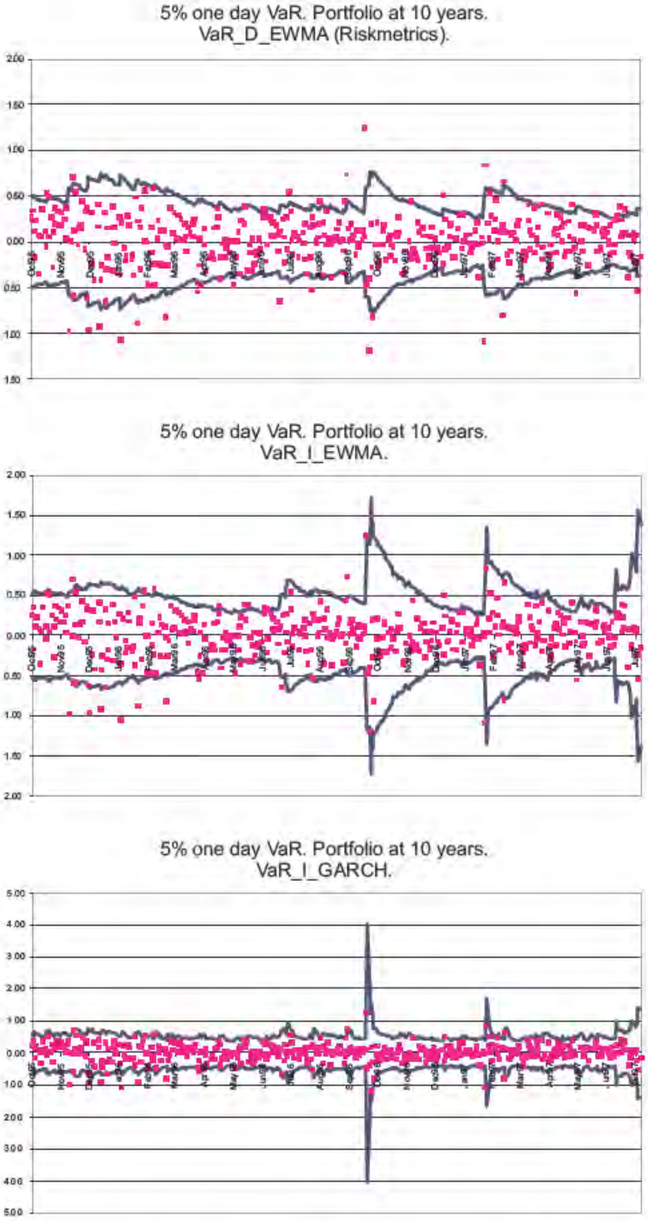


Figure 5: The 5% one day VaR for a 10 year portfolio. Direct estimation using an exponentially weighted moving average model [VaR_D_EWMA(5%)], indirect estimation using an exponentially weighted moving average model [VaR_I_EWMA(5%)] and indirect estimation using a GARCH model [VaR_I_GARCH(5%)].

4.3. Comparing VaR measures

In this section VaR measures are compared. We calculate daily VaR at a 5%, 4%, 3%, 2% and 1% confidence level for all portfolios. First, before formally evaluating the precision of the VaR measures under comparison, we examine actual daily portfolio value changes (as implied by daily fluctuations in the zero-coupon interest rate) and compare them with the 5% VaR. In Figure 5 we show the actual change in a 10-year portfolio together with the VaR at 5% for the three VaR measures we consider: VaR_D_EWMA (Graph 1), VaR_I_EWMA (Graph 2) and VaR_I_GARCH (Graph 3). In Graph 1 and 2, we observe that the portfolio's value falls below VaR more often than in Graph 3. In all cases, the number of times the value of the portfolio falls below VaR is closer to its theoretical level. This is also a clear result for the other portfolios considered, but we will not show it due to space limitations. This preliminary analysis suggests that VaR estimations from both models, both directly and indirectly are very precise; however, a more rigorous evaluation of the precision of the estimations is required[5].

We then compare VaR measures with the actual change in a portfolio value on day t+1, denoted as ΔV_{t+1}. When $\Delta V_{t+1} < VaR$, we have an exception. For testing purposes, we define the exception indicator variable as

$$I_{t+1} = \begin{cases} 1 \text{ if } \Delta V_{t+1} < VaR \\ 0 \text{ if } \Delta V_{t+1} \geq VaR \end{cases} \tag{16}$$

a) Testing the level

The most basic test of a VaR procedure is to see if the stated probability level is actually achieved. The mean of the exception indicator series is the level of achievement for the procedure. If we assume a constant probability for the exception, the number of exceptions follows the binomial distribution. Thus, it is possible to build up confidence intervals for the level of each VaR measure (see Kupiec, 1995).

Table 6 shows the level achieved and the 95% confidence interval for each of the 1-day VaR estimates. An * indicates the cases in which the level is out of the confidence interval, evidence obtained rejects the null hypothesis at the 5% confidence level. The number of exceptions is inside the interval confidence for the three measures and almost all portfolios considered. Therefore, VaR estimates (direct and indirect) seem to be good.

Only for three cases is the number of exceptions out of the confidence interval. Specifically, for VaR_I_GARCH measure for 4% and 5% confidence level in 5- and 10-year portfolios. The number of exceptions in these cases is much lower than the theoretical level, so it would seem this measure overestimates the risk of these portfolios.

[5] The empirical assessment of VaR is not developed through analysing the mean squared error (Longford, 2008). Instead, we use the standard test of VaR.

Table 6: *Testing the Level.*

	Number of exceptions				
VaR measures	**3-year**	**5-year**	**10-year**	**15-year**	**Confidence intervals at the 95% level**
Var_D_EWMA (1%)	6	7	7	5	(1-10)
Var_D_EWMA (2%)	9	9	14	10	(5-17)
Var_D_EWMA (3%)	15	12	17	14	(8-23)
Var_D_EWMA (4%)	17	14	17	16	(12-30)
Var_D_EWMA (5%)	20	20	23	20	(17-36)
Var_I_EWMA (1%)	7	7	10	8	(1-10)
Var_I_EWMA (2%)	13	11	11	13	(5-17)
Var_I_EWMA (3%)	16	12	15	17	(8-23)
Var_I_EWMA (4%)	19	15	16	19	(12-30)
Var_I_EWMA (5%)	19	19	19	24	(17-36)
Var_I_GARCH (1%)	6	6	6	6	(1-10)
Var_I_GARCH (2%)	11	9	8	9	(5-17)
Var_I_GARCH (3%)	12	10	11	11	(8-23)
Var_I_GARCH (4%)	16	10*	13	13	(12-30)
Var_I_GARCH (5%)	19	13*	14*	19	(17-36)

Note: Sample period 4/15/2002 to 12/31/2004 (437 observations). Confidence intervals derived from the number of exceptions follows the binomial distribution (437, *x*%) for *x* = 1, 2, 3, 4 and 5. An * indicates the cases in which the number of exceptions is out of the confidence interval, so that, we obtain evidence to reject the null hypothesis at the 5% level type I error rate.

b) Testing consistency of level

We want the VaR level found to be the stated level on average, but we also want to find the stated level at all points in time. One approach to test the consistency of the level is the Ljung-Box portmanteau test (Ljung and Box, 1978) on the exception indicator variable of zeros and ones. When using Ljung-Box tests, there is a choice of the number of lags in which to look for autocorrelation. If the test uses only a few lags but autocorrelation occurs over a long time frame, the test will miss some of the autocorrelation. Conversely, should a large number of lags be used in the test when the autocorrelation is only in a few lags, then the test will not be as sensitive as if the number of lags in the test matched the autocorrelation.

Different lags have been used for each estimate in order to have a good picture of autocorrelation. Table 7 shows the Ljung-Box statistics at lags of 4 and 8. We only detect the existence of autocorrelation in the 10-year portfolio with the measures VaR_I_EWMA(3%) and (4%). In general, the results of the Ljung-Box comparison indicate that autocorrelation is not present. When we consider other lags not shown for space reasons, the result is very similar. We can also conclude from this test the VaR estimates are good.

Table 7: *Testing Consistency of Level.*

	3-year		**5-year**		**10-year**		**15-year**	
Lags	**4**	**8**	**4**	**8**	**4**	**8**	**4**	**8**
VaR_D_EWMA (1%)	0.35	0.71	0.38	0.64	0.38	0.64	0.24	0.49
	(0.987)	(1.000)	(0.984)	(1.000)	(0.984)	(1.000)	(0.993)	(1.000)
VaR_D_EWMA (2%)	0.79	5.18	0.67	1.17	2.14	3.65	0.98	2.00
	(0.940)	(0.739)	(0.955)	(0.997)	(0.711)	(0.887)	(0.913)	(0.981)
VaR_D_EWMA (3%)	1.97	4.13	1.26	3.73	2.25	4.17	2.19	4.24
	(0.740)	(0.845)	(0.869)	(0.881)	(0.690)	(0.842)	(0.700)	(0.835)
VaR_D_EWMA (4%)	2.14	3.86	1.76	3.67	2.25	4.17	5.29	7.98
	(0.709)	(0.870)	(0.780)	(0.886)	(0.690)	(0.842)	(0.259)	(0.435)
VaR_D_EWMA (5%)	1.88	3.61	1.90	3.63	5.60	7.92	3.85	7.93
	(0.758)	(0.890)	(0.754)	(0.889)	(0.231)	(0.441)	(0.427)	(0.440)
VaR_I_EWMA (1%)	0.47	0.97	0.47	0.97	0.98	4.49	5.68	6.33
	(0.976)	(0.998)	(0.976)	(0.998)	(0.913)	(0.811)	(0.225)	(0.611)
VaR_I_EWMA (2%)	2.30	4.65	2.90	5.82	2.89	5.81	2.29	4.04
	(0.680)	(0.795)	(0.576)	(0.667)	(0.577)	(0.668)	(0.683)	(0.853)
VaR_I_EWMA (3%)	5.31	7.64	2.52	5.08	14.52*	16.77*	4.58	7.64
	(0.257)	(0.469)	(0.641)	(0.749)	(0.006)	(0.033)	(0.333)	(0.470)
VaR_I_EWMA (4%)	8.19	10.18	6.32	8.45	12.08*	14.41	3.70	7.55
	(0.085)	(0.253)	(0.177)	(0.391)	(0.017)	(0.072)	(0.449)	(0.479)
VaR_I_EWMA (5%)	8.19	10.18	8.19	8.32	9.07	12.00	8.42	12.30
	(0.085)	(0.253)	(0.085)	(0.403)	(0.059)	(0.151)	(0.077)	(0.138)
VaR_I_GARCH (1%)	0.35	0.71	0.35	0.71	0.35	0.71	0.35	0.71
	(0.987)	(1.000)	(0.987)	(1.000)	(0.987)	(1.000)	(0.987)	(1.000)
VaR_I_GARCH (2%)	1.19	4.12	0.79	5.18	0.62	6.33	0.79	1.62
	(0.879)	(0.846)	(0.940)	(0.739)	(0.961)	(0.610)	(0.940)	(0.991)
VaR_I_GARCH (3%)	1.43	3.98	0.98	4.49	1.19	4.12	1.19	2.44
	(0.840)	(0.859)	(0.913)	(0.811)	(0.879)	(0.846)	(0.879)	(0.965)
VaR_I_GARCH (4%)	5.64	10.97	0.98	4.49	2.29	4.63	2.29	4.63
	(0.228)	(0.203)	(0.913)	(0.811)	(0.683)	(0.796)	(0.683)	(0.796)
VaR_I_GARCH (5%)	4.61	8.37	2.30	5.23	2.18	4.43	4.60	6.60
	(0.329)	(0.398)	(0.680)	(0.733)	(0.702)	(0.816)	(0.330)	(0.581)

Note: Sample period 4/15/2002 to 12/31/2004. The Ljung-Box Q-statistics on the exception indicator variable and their *p*-values. The Q-statistic at lag 4(8) for the null hypothesis that there is no autocorrelation up to order 4(8). An * indicates that there is evidence to reject the null hypothesis at the 5% level type I error date.

c) *The back-testing criterion*

The back-testing criterion is used to evaluate the performance of VaR measures. The most popular back-testing measure for accuracy of the quantile estimator is the percentage of returns that falls below the quantile estimate which is denoted as $\hat{\alpha}$. For an accurate estimator of an α quantile, $\hat{\alpha}$ will be very close to α%. In order to determine the significance of α departure of from $\hat{\alpha}$%, the following test statistic is used:

$$Z = (T\hat{\alpha} - T\alpha\%) / \sqrt{T\alpha\%(1-\alpha\%)} \longrightarrow N(0,1) \tag{17}$$

where T is the sample size.

Table 8: *The Back-testing Criterion.*

	% of exceptions			
	3-year	**5-year**	**10-year**	**15-year**
VaR_D_EWMA (1%)	1.37% [0.784]	1.60% [1.264]	1.60% [1.264]	1.14% [0.303]
VaR_D_EWMA (2%)	2.06% [0.089]	3.10% [1.644]	3.88%* [2.801]	4.26%* [3.380]
VaR_D_EWMA (3%)	3.43% [0.530]	2.75% [−0.311]	3.89% [1.091]	3.20% [0.250]
VaR_D_EWMA (4%)	3.89% [−0.117]	3.20% [−0.850]	3.89% [−0.117]	3.66% [−0.361]
VaR_D_EWMA (5%)	4.58% [−0.406]	4.58% [−0.406]	5.26% [0.252]	4.58% [−0.406]
VaR_I_EWMA (1%)	1.60% [1.264]	1.60% [1.264]	2.29%* [2.707]	1.83% [1.745]
VaR_I_EWMA (2%)	2.97% [1.456]	2.52% [0.772]	2.52% [0.772]	2.97% [1.456]
VaR_I_EWMA (3%)	3.66% [0.810]	2.75% [−0.311]	3.43% [0.530]	3.89% [1.091]
VaR_I_EWMA (4%)	4.35% [0.371]	3.43% [−0.605]	3.66% [−0.361]	4.35% [0.371]
VaR_I_EWMA (5%)	4.35% [−0.626]	4.35% [−0.626]	4.35% [−0.626]	5.49% [0.472]
VaR_I_GARCH (1%)	1.37% [0.784]	1.37% [0.784]	1.37% [0.784]	1.37% [0.784]
VaR_I_GARCH (2%)	2.52% [0.772]	2.06% [0.089]	1.83% [−0.253]	2.06% [0.089]
VaR_I_GARCH (3%)	2.75% [−0.311]	2.29% [−0.872]	2.52% [−0.592]	2.52% [−0.592]
VaR_I_GARCH (4%)	3.66% [−0.361]	2.29% [−1.826]	2.97% [−1.094]	2.97% [−1.094]
VaR_I_GARCH (5%)	4.35% [−0.626]	2.97% [−1.942]	3.20% [−1.723]	4.35% [−0.626]

Note: Sample period 4/15/2002 to 12/31/2004. Percentage of exceptions. In square brackets Back-testing Criterion: The Z statistic for determining the significance of departure for $\hat{\alpha} = x/T$ from $\alpha\%$. An * indicates that there is evidence to reject the null hypothesis at the 5% level type I error rate.

Table 8 presents the percentage of exception and, in square brackets, the *Z* statistic for VaR measures. For measures computed with EWMA (independently of the quantile considered) we reject the null hypothesis that the percentage of exceptions coincides with the corresponding quantile in three cases. More precisely, in two occasions with the VaR_D_EWMA measure and once with the VaR_I_EWMA measure. On the other hand, the null hypothesis is never rejected for the VaR_I_GARCH measure.

In summary, we can say that the VaR measures we obtain using the simplification proposed in this paper are, at least as good as those computed with RiskMetrics method (VaR_D_EWMA). Nevertheless, the advantage of the proposed method is a much lower computational cost to calculate VaR.

5. Conclusion

When we use the most commonly implemented parametric approach, we need to estimate the variance-covariance matrix of the portfolio assets. The variance-covariance matrix of prices of a bonds vector from a portfolio depends on the variance-covariance matrix of the interest rates that determine its value. The estimation of the interest rates matrix entails two types of practical problems: dimensionality (the number of variances and covariances to be estimated may be very large), and feasibility (the estimation of interest rates covariances using multivariate methods becomes unfeasible as the dimension increases).

The aim of this paper is to propose a method for calculating the variance-covariance matrix of a large set of interest rates with a low computational cost. The suggested methodology exploits the parameterization of the underlying interest rates proposed by Nelson and Siegel (1987) for estimating the term structure of interest rate (TSIR). Our method turns out to be useful for estimating Value at Risk (VaR), since it considerably simplifies the calculation of this measure.

We start with an explanatory model of interest rates: the Nelson and Siegel (1987) model originally developed to estimate the TSIR. This model provides a relationship to account for changes in interest rates as a function of changes in four parameters, using a linear approximation. Although this approximation reduces the dimension of the variance-covariance matrix, it still requires covariance to be estimated. In order to solve this problem, we propose applying principal components of the changes in the four parameters of the Nelson and Siegel model. Given orthogonality among principal components, the resulting diagonal variance-covariance matrix has a smaller dimension, i.e., all covariances are zero.

The procedure we propose in this paper has been tested using data from the Spanish debt market. The results obtained from applying our methodology are very satisfactory. On the one hand, the variances estimated with our procedure and those from a direct estimation are quite similar, regardless of the method used to estimate the volatility (exponentially weighted moving average or RiskMetrics methodology vs. Generalized

Autoregressive Conditional Heteroskedasticity models). As for VaR calculation, the estimations we obtain with this procedure are quite precise, independently of the method used to estimate the volatility.

An additional advantage of the proposed method is that it is not necessary to decompose the assets into cash-flow and subsequently assign cash to a series of vertexes (RiskMetrics cash flow mapping method). This stems from the fact that our method allows us to estimate the variances and covariances of a vector of interest rates at the same cost and independently from the dimension of the problem. It is unnecessary to reduce the TSIR to a small number of vertexes.

Finally, we should mention that the methodology proposed in this paper presupposes a small implementation cost for financial institutions, since the majority of them already use the Nelson and Siegel (1987) method to estimate TSIR or yield curve. Therefore, these institutions already have the information required to implement our procedure.

Acknowledgements

We wish to thank A. Novales and the anonymous referees of this journal for their useful comments. Of course, they should not bear any responsibility for the remaining errors. The authors acknowledge and appreciate the funding from the Spanish Ministry of Education and FEDER through projects ECO2008-02752/ECON and BEC2003-03965. An earlier version of this article was published as FUNCAS Working Paper# 300.

References

Aas, K. and Haff, I. H. (2006). The generalized hyperbolic Skew student’s t-distribution. *Journal of Financial Econometrics*, 4, 275-309.

Albanese, C., K. Jackson and Wiberg, P. (2004). A new Fourier transform algorithm for value-at-risk. *Quantitative Finance*, 4, 328-338.

Alexander, C. O. (2001). “*Orthogonal GARCH*,” *Mastering Risk* (C. O. Alexander, Ed.) Volume 2. Financial Times-Prentice Hall, 21-38.

Alexander, C. O. and Leigh, C. T. (1997). On the covariance matrices used in value at risk models. *Journal of Derivatives*, 4, 50-62.

Bao, Y., Lee, T.-H. and Saltoglu, B. (2006). Evaluating predictive performance of value at risk models in emerging markets: a reality check. *Journal of Forecasting*, 25, 101-128.

Billio, M. and Pelizzon, L. (2000). Value-at-risk: a multivariate switching regime approach. *Journal of Empirical Finance*, 7, 531-554.

Brooks, C., Clare, A. D., Dalle Molle, J. W. and Persand, G. (2005). A comparison of extreme value theory approaches for determining value at risk. *Journal of Empirical Finance,* 12, 339-352.

Cabedo, J. D. and Moya, I. (2003). Value at risk calculation through ARCH factor methodology: proposal and comparative analysis. *European Journal of Operational Research*, 150, 516-528.

Cai, Z. (2002). Regression quantiles for time series. *Econometric Theory*, 18, 169-192.

Cakici, N. and Foster, K. R. (2003). Value at risk for interest rate-dependent securities. *Journal of Fixed Income*, 12, 81-96.

Chong, J. (2004). Value at risk from econometric models and implied from currency options. *Journal of Forecasting*, 23, 603-620.

Christiansen, C. (1999). Value at risk using the factor-ARCH model. *Journal of Risk*, 1, 65-87.

Consigli, G. (2002). Tail estimation and mean-VaR portfolio selection in markets subject to financial instability. *Journal Banking and Finance*, 26, 1355-1382.

Daníelsson, J. (2002). The emperor has no clothes: limits to risk modelling. *Journal Banking of Finance*, 26, 1273-1296.

Daníelsson, J. and Vries, C. G. (2000). *Value-at-Risk and Extreme Returns*, Mimeo, Tinbergen Institute Rotterdam.

Fan, J. and Gu, J. (2003). Semiparametric estimation of value at risk. *Econometrics Journal*, 6, 261-290.

Ferenstein, E. and Gasowski, M. (2004). Modelling stock returns with AR-GARCH processes. *Statistics and Operations Research Transactions Journal (SORT)*, 28, 55-68.

Glasserman, P., Heidelberger, P. and Shahabuddin, P. (2000). Variance reduction techniques for estimating value-at-risk. *Management Science*, 46, 1349-1364.

Jorion, P. (2000). *Value at Risk: The New Benchmark for Managing Financial Risk*, published by McGraw-Hill.

Kamdem, J. S. (2005). Value-at-risk and expected shortfall for linear portfolios with elliptically distributed risk factors. *International Journal of Theoretical & Applied Finance*, 8, 537-551.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 2, 73-84.

Li, M-YL and Lin, H-WW. (2004). Estimating value-at-risk via Markov switching ARCH models-an empirical study on stock index returns. *Applied Economics Letters*, 11, 679-691.

Linsmeier, T. J. and Pearson, N. D. (2000). Value at risk. *Financial Analysts Journal*, 56, 47-67.

Longford, N. T. (2008). An alternative analysis of variance. *Statistics and Operations Research Transactions Journal (SORT)*, 32, 77-92.

McNeil, A. J., Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, 7, 271-300.

Miller, D. J. and Liu, W. (2006). Improved estimation of portfolio value-at-risk under copula models with mixed marginals. *Journal of Futures Markets*, 26, 997-1018.

Mittnik, S., Paolella, M. S. and Rachev, S. T. (2002). Stationarity of stable power-GARCH processes. *Journal of Econometrics*, 106, 97-107.

Morgan, J. P. (1995). *RiskMetrics Technical Document*, 3d ed. New York.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica*, 59, 347-370.

Nelson, C. R. and Siegel, A. F. (1987). Parsimonious modelling of yield curves. *Journal of Business*, 60, 473-489.

Sarma, M., Thomas, S. and Shah, A. (2003). Selection of value at risk models. *Journal of Forecasting*, 22, 337-358.

SORT 34 (1) January-June 2010, 45-64

A family of ratio estimators for population mean in extreme ranked set sampling using two auxiliary variables

Abdul Haq and Javid Shabbir*

Quaid-i-Azam University

Abstract

In this paper we have adopted the Khoshnevisan ***et al.*** (2007) family of estimators to extreme ranked set sampling (***ERSS***) using information on single and two auxiliary variables. Expressions for mean square error (***MSE***) of proposed estimators are derived to first order of approximation. Monte Carlo simulations and real data sets have been used to illustrate the method. The results indicate that the estimators under ***ERSS*** are more efficient as compared to estimators based on simple random sampling (***SRS***), when the underlying populations are symmetric.

MSC: 62DO5

Keywords: Ratio estimator, ranked set sampling, extreme ranked set sampling.

1. Introduction

Ranked set sampling (*RSS*) was introduced by McIntyre (1952) and suggested using *RSS* as a costly efficient alternative as compared to *SRS*. Takahasi and Wakimoto (1968) developed the mathematical theory and proved that the sample mean of a ranked set sample is an unbiased estimator of the population mean and possesses smaller variance than the sample mean of a simple random sample with the same sample size. Samawi and Muttlak (1996) suggested the use of *RSS* to estimate the population ratio and showed that it gives more efficient estimates as compared to *SRS*. Samawi *et al.* (1996) introduced *ERSS* to estimate the population mean and showed that the sample mean under *ERSS*

* Department of Statistics, Quaid-i-Azam University, Islamabad, 45320, Pakistan.
Corresponding Emails: aaabdulhaq@yahoo.com and jsqau@yahoo.com.
Received: August 2009
Accepted: April 2010

is an unbiased and is more efficient than the sample mean based on *SRS*. Samawi (2002) introduced the ratio estimation in estimating the population ratio using *ERSS* and showed that the ratio estimator under *ERSS* is an approximately unbiased estimator of the population ratio. Also in the case of symmetric populations ratio estimators under *ERSS* are more efficient than ratio estimators under *SRS*. Samawi and Saeid (2004) investigated the use of the separate and the combined ratio estimators in *ERSS*. Samawi *et al.* (2004) studied the use of regression estimator in *ERSS* and showed that for symmetric distributions, the regression estimator under *ERSS* is more efficient as compared to *SRS* and *RSS*.

In this paper, *SRS* and *ERSS* methods are used for estimating the population mean of the study variable Y by using information on the auxiliary variables X and Z.

The organization of this paper is as follows. Section 2 includes sampling methods like *SRS* and *ERSS*. In Section 3, main notations and results are given. Sections 4 and 5 comprise of a family of ratio estimators using single and two auxiliary variables. Section 6 describes of simulation and empirical studies and Section 7 finally provides the conclusion.

2. Sampling methods

2.1. Simple random sampling

In *SRS*, m units out of N units of a population are drawn in such a way that every possible combination of items that could make up a given sample size has an equal chance of being selected. In usual practice, a simple random sample is drawn unit by unit.

2.2. Ranked set sampling

RSS procedure involves selection of m sets, each of m units from the population. It is assumed that units within each set can be ranked visually at no cost or at little cost. From the first set of m units, the lowest ranked unit is selected; the remaining units of the sample are discarded. From the second set of m units, the second lowest ranked unit is selected and the remaining units are discarded. The procedure is continued until from the mth set, the mth ranked unit is selected. This completes one cycle of a ranked set sample of size m. The whole process can be repeated r times to get a ranked set sample of size $n = mr$.

2.3. Extreme ranked set sampling

Samawi *et al.* (1996) introduced a new variety of ranked set sampling, named as *ERSS* to estimate the population mean and have shown that *ERSS* gives more efficient estimates as compared to *SRS*.

In *ERSS*, m independent samples, each of m units are drawn from infinite population to estimate the unknown parameter. Here we assume that lowest and largest units of these samples can be detected visually with no cost or with little cost as explained by Samawi (2002). From the first set of m units, the lowest ranked unit is measured, similarly from the second set of m units, the largest ranked unit is measured. Again in the third set of m units the lowest ranked unit is measured and so on. The procedure continues until from $(m-1)$ units, $(m-1)$ units are measured. From the last mth sample, the selection of the unit depends whether m is even or not. It can be measured in two ways:

(i) If m is even then the largest ranked unit is to be selected; we denote such a sample with notation $ERSS_a$.

(ii) If m is odd then for the measurement of the mth unit, we take the average of the lowest and largest units of the mth sample; such a sample will be donated by $ERSS_b$ or we take the median of the mth sample; such a sample is denoted by $ERSS_c$.

The choice of a sample $ERSS_b$ will be more difficult as compared to the choices of $ERSS_a$ and $ERSS_c$ (see Samawi *et al.* 1996). The above procedure can be repeated r times to select an *ERSS* of size mr units.

3. Notations under *SRS* and *ERSS*

Let $(X_1,\, Y_1),(X_2,\, Y_2),\ldots,(X_m,\, Y_m)$ be a random sample from a bivariate normal distribution with probability density function $f(X,Y)$, having parameters $\mu_X,\ \mu_Y,\ \sigma_X, \sigma_Y$ and ρ. We assume that the ranking is performed on the auxiliary variable X for estimating the population mean (μ_Y). Let $(X_{11},\, Y_{11}),(X_{12},\, Y_{12}),\ldots,(X_{1m},\, Y_{1m})$, $(X_{21},\, Y_{21})$, $(X_{22},\, Y_{22}),\ldots,(X_{2m},\, Y_{2m})$,..., $(X_{m1},\, Y_{m1}),(X_{m2},\, Y_{m2}),\ldots,(X_{mm},\, Y_{mm})$ be m independent bivariate random vectors each of size m, $\left(X_{i(1)},\, Y_{i[1]}\right),\left(X_{i(2)},\, Y_{i[2]}\right),\ldots,\left(X_{i(m)},\, Y_{i[m]}\right)$ be the *RSS* for $i=1,2,\ldots m$. In *ERSS*, if m is even then $\left(X_{1(1)j},\, Y_{1[1]j}\right),\left(X_{2(m)j},\, Y_{2[m]j}\right)$, $\ldots,\left(X_{m-1(1)j},\, Y_{m-1[1]j}\right),\left(X_{m(m)j},\, Y_{m[m]j}\right)$, denoted by $ERSS_a$, and if m is odd then $\left(X_{1(1)j},\, Y_{1[1]j}\right),\left(X_{2(m)j},\, Y_{2[m]j}\right),\ldots,\left(X_{m-1(m)j},\, Y_{m-1[m]j}\right),\left(X_{m\left(\frac{m+1}{2}\right)j},\, Y_{m\left[\frac{m+1}{2}\right]j}\right)$, denoted by $ERSS_c$, for the jth cycle, where j=1,2,..., r.

Considering ranking on the auxiliary variable X, we use the following notations and results.

Let $E(X_i)=\mu_X,\ E(Y_i)=\mu_Y,\ Var(X_i)=\sigma_X^2,\ Var(Y_i)=\sigma_Y^2,\ E\left(X_{i(m)}\right)=\mu_{X(m)}$, $E\left(Y_{i[m]}\right)=\mu_{Y[m]}$, $E\left(X_{i(1)}\right)=\mu_{X(1)}$, $E\left(Y_{i[1]}\right)=\mu_{Y[1]}$, $Var\left(X_{i(1)}\right)=\sigma_{X(1)}^2$, $Var\left(Y_{i[1]}\right)=\sigma_{Y[1]}^2$,

$$Var\left(X_{i(m)}\right)=\sigma^2_{X(m)}, \qquad Var\left(Y_{i[m]}\right)=\sigma^2_{Y[m]},$$
$$E\left(X_{i\left(\frac{m+1}{2}\right)}\right)=\mu_{X\left(\frac{m+1}{2}\right)}, \qquad E\left(Y_{i\left[\frac{m+1}{2}\right]}\right)=\mu_{Y\left[\frac{m+1}{2}\right]},$$
$$Var\left(X_{i\left(\frac{m+1}{2}\right)}\right)=\sigma^2_{X\left(\frac{m+1}{2}\right)}, \quad Var\left(Y_{i\left[\frac{m+1}{2}\right]}\right)=\sigma^2_{Y\left[\frac{m+1}{2}\right]}$$

and

$$Cov\left(X_{i(h)},Y_{i[k]}\right)=\sigma_{X(h)Y[k]}.$$

In *SRS* the sample means of variables X and Y are

$$\bar{X}=\frac{1}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m}X_{ij}$$

and

$$\bar{Y}=\frac{1}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m}Y_{ij}$$

In $ERSS_a$, the sample means of X and Y are

$$\bar{X}_{(a)}=\frac{1}{2}\left(\bar{X}_{(1)}+\bar{X}_{(m)}\right),$$

where

$$\bar{X}_{(1)}=\frac{2}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m/2}X_{2i-1(1)j}, \qquad \bar{X}_{(m)}=\frac{2}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m/2}X_{2i(m)j}$$

and

$$\bar{Y}_{[a]}=\frac{1}{2}\left(\bar{Y}_{[1]}+\bar{Y}_{[m]}\right),$$

where

$$\bar{Y}_{[1]}=\frac{2}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m/2}Y_{2i-1[1]j}, \qquad \bar{Y}_{[m]}=\frac{2}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m/2}Y_{2i[m]j}.$$

In $ERSS_c$, we define

$$\bar{X}_{(c)}=\frac{\sum_{j=1}^{r}\left(X_{1(1)j}+X_{2(m)j}+\cdots+X_{m-1(m)j}+X_{m\left(\frac{m+1}{2}\right)j}\right)}{mr}=\frac{\left(\frac{m-1}{2}\right)\left(\bar{X}'_{(1)}+\bar{X}'_{(m)}\right)+\bar{X}'_{\left(\frac{m+1}{2}\right)}}{m},$$

where

$$\bar{X}'_{(1)} = \frac{2}{r\,(m-1)}\sum_{j=1}^{r}\sum_{i=1}^{(m-1)/2} X_{2i-1(1)j}, \quad \bar{X}'_{(m)} = \frac{2}{r\,(m-1)}\sum_{j=1}^{r}\sum_{i=1}^{(m-1)/2} X_{2i(m)j},$$

$$\bar{X}'_{\left(\frac{m+1}{2}\right)} = \frac{1}{r}\sum_{j=1}^{r} X_{m\left(\frac{m+1}{2}\right)j}.$$

Also for Y, we have

$$\bar{Y}_{[c]} = \frac{\sum_{j=1}^{r}\left(Y_{1[1]j}+Y_{2[m]j}+\cdots+Y_{m-1[m]j}+Y_{m\left[\frac{m+1}{2}\right]j}\right)}{mr} = \frac{\left(\frac{m-1}{2}\right)\left(\bar{Y}'_{[1]}+\bar{Y}'_{[m]}\right)+\bar{Y}'_{\left[\frac{m+1}{2}\right]}}{m},$$

where

$$\bar{Y}'_{[1]} = \frac{2}{r\,(m-1)}\sum_{j=1}^{r}\sum_{i=1}^{(m-1)/2} Y_{2i-1[1]j}, \quad \bar{Y}'_{[m]} = \frac{2}{r\,(m-1)}\sum_{j=1}^{r}\sum_{i=1}^{(m-1)/2} Y_{2i[m]j},$$

$$\bar{Y}'_{\left[\frac{m+1}{2}\right]} = \frac{1}{r}\sum_{j=1}^{r} Y_{m\left[\frac{m+1}{2}\right]j}.$$

Similarly, in case of the two auxiliary variables X and Z, when ranking is done on Z, we use the following notations.

$$\begin{array}{lll}
E\left(Y_{i[m]}\right)=\mu_{Y[m]}, & E\left(X_{i[m]}\right)=\mu_{X[m]}, & E\left(Z_{i(m)}\right)=\mu_{Z(m)}, \\
E\left(Y_{i[1]}\right)=\mu_{Y[1]}, & E\left(X_{i[1]}\right)=\mu_{X[1]}, & E\left(Z_{i(1)}\right)=\mu_{Z(1)}, \\
Var\left(Y_{i[1]}\right)=\sigma^2_{Y[1]}, & Var\left(X_{i[1]}\right)=\sigma^2_{X[1]}, & Var\left(Z_{i(1)}\right)=\sigma^2_{Z(1)}, \\
Var\left(Y_{i[m]}\right)=\sigma^2_{Y[m]}, & Var\left(X_{i[m]}\right)=\sigma^2_{X[m]}, & Var\left(Z_{i(m)}\right)=\sigma^2_{Z(m)}, \\
E\left(Y_{i\left[\frac{m+1}{2}\right]}\right)=\mu_{Y\left[\frac{m+1}{2}\right]}, & E\left(X_{i\left[\frac{m+1}{2}\right]}\right)=\mu_{X\left[\frac{m+1}{2}\right]}, & E\left(Z_{i\left(\frac{m+1}{2}\right)}\right)=\mu_{Z\left(\frac{m+1}{2}\right)}, \\
Var\left(Y_{i\left[\frac{m+1}{2}\right]}\right)=\sigma^2_{Y\left[\frac{m+1}{2}\right]}, & Var\left(X_{i\left[\frac{m+1}{2}\right]}\right)=\sigma^2_{X\left[\frac{m+1}{2}\right]}, & Var\left(Z_{i\left(\frac{m+1}{2}\right)}\right)=\sigma^2_{Z\left(\frac{m+1}{2}\right)},
\end{array}$$

$Cov\left(X_{i[h]},Y_{i[k]}\right)=\sigma_{X[h]Y[k]}$, $Cov\left(X_{i[h]},Z_{i(k)}\right)=\sigma_{X[h]Z(k)}$ and $Cov\left(Y_{i[h]},Z_{i(k)}\right)=\sigma_{Y[h]Z(k)}$.

In *SRS* the sample means of variables X, Y and Z are

$$\bar{X}=\frac{1}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m} X_{ij}, \qquad \bar{Y}=\frac{1}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m} Y_{ij} \qquad \text{and} \qquad \bar{Z}=\frac{1}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m} Z_{ij}.$$

In $ERSS_a$, the sample means of X, Y and Z are

$$\overline{X}_{[a]}=\frac{1}{2}\left(\overline{X}_{[1]}+\overline{X}_{[m]}\right),$$

where

$$\overline{X}_{[1]}=\frac{2}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m/2}X_{2i-1[1]j},\qquad \overline{X}_{[m]}=\frac{2}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m/2}X_{2i[m]j},\qquad \overline{Y}_{[a]}=\frac{1}{2}\left(\overline{Y}_{[1]}+\overline{Y}_{[m]}\right),$$

where

$$\overline{Y}_{[1]}=\frac{2}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m/2}Y_{2i-1[1]j},\qquad \overline{Y}_{[m]}=\frac{2}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m/2}Y_{2i[m]j}\qquad \text{and}\qquad \overline{Z}_{(a)}=\frac{1}{2}\left(\overline{Z}_{(1)}+\overline{Z}_{(m)}\right),$$

where

$$\overline{Z}_{(1)}=\frac{2}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m/2}Z_{2i-1(1)j},\qquad \overline{Z}_{(m)}=\frac{2}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m/2}Z_{2i(m)j}.$$

In $ERSS_c$, the sample means for X, Y and Z are

$$\overline{X}_{[c]}=\frac{\left(\frac{m-1}{2}\right)\left(\overline{X}'_{[1]}+\overline{X}'_{[m]}\right)+\overline{X}'_{\left[\frac{m+1}{2}\right]}}{m},$$

where

$$\overline{X}'_{[1]}=\frac{2}{r\left(m-1\right)}\sum_{j=1}^{r}\sum_{i=1}^{(m-1)/2}X_{2i-1[1]j},\qquad \overline{X}'_{[m]}=\frac{2}{r\left(m-1\right)}\sum_{j=1}^{r}\sum_{i=1}^{(m-1)/2}X_{2i[m]j},$$

$$\overline{X}'_{\left[\frac{m+1}{2}\right]}=\frac{1}{r}\sum_{j=1}^{r}X_{m\left[\frac{m+1}{2}\right]j},\qquad \overline{Y}_{[c]}=\frac{\left(\frac{m-1}{2}\right)\left(\overline{Y}'_{[1]}+\overline{Y}'_{[m]}\right)+\overline{Y}'_{\left[\frac{m+1}{2}\right]}}{m},$$

where

$$\overline{Y}'_{[1]}=\frac{2}{r\left(m-1\right)}\sum_{j=1}^{r}\sum_{i=1}^{(m-1)/2}Y_{2i-1[1]j},\qquad \overline{Y}'_{[m]}=\frac{2}{r\left(m-1\right)}\sum_{j=1}^{r}\sum_{i=1}^{(m-1)/2}Y_{2i[m]j},$$

$$\overline{Y}'_{\left[\frac{m+1}{2}\right]}=\frac{1}{r}\sum_{j=1}^{r}Y_{m\left[\frac{m+1}{2}\right]j}\qquad \text{and}\qquad \overline{Z}_{(c)}=\frac{\left(\frac{m-1}{2}\right)\left(\overline{Z}'_{(1)}+\overline{Z}'_{(m)}\right)+\overline{Z}'_{\left(\frac{m+1}{2}\right)}}{m},$$

where

$$\bar{Z}'_{(1)} = \frac{2}{r\,(m-1)} \sum_{j=1}^{r} \sum_{i=1}^{(m-1)/2} Z_{2i-1(1)j}, \qquad \bar{Z}'_{(m)} = \frac{2}{r\,(m-1)} \sum_{j=1}^{r} \sum_{i=1}^{(m-1)/2} Z_{2i(m)j},$$

$$\bar{Z}'_{\left(\frac{m+1}{2}\right)} = \frac{1}{r} \sum_{j=1}^{r} Z_{m\left(\frac{m+1}{2}\right)j}.$$

4. Proposed estimators using the single auxiliary variable

4.1. A family of ratio estimators using $ERSS_a$

Following Khoshnevisan *et al.* (2007), we propose a family of ratio estimators in $ERSS_a$ using the single auxiliary variable, when ranking is performed on the auxiliary variable X and is given by

$$\hat{\bar{Y}}_{ERSS_a} = \bar{Y}_{[a]} \left[\frac{a\mu_X + b}{\alpha\left(a\,\bar{X}_{(a)} + b\right) + (1-\alpha)\,(a\mu_X + b)} \right]^g, \tag{1}$$

where α and g are suitable constants, also a and b are either real numbers or functions of known parameters for the auxiliary variable X, like coefficient of variation (C_X) or coefficient of kurtosis (β_{2X})or standard deviation (S_X) or coefficient of correlation (ρ_{YX}).

Using bivariate Taylor series expansion, we have

$$\left(\hat{\bar{Y}}_{ERSS_a} - \mu_Y\right) \cong \frac{1}{2}\left[\bar{Y}_{[1]} - E\left(\bar{Y}_{[1]}\right)\right] + \frac{1}{2}\left[\bar{Y}_{[m]} - E\left(\bar{Y}_{[m]}\right)\right] - \frac{\mu_Y\,(a\alpha g)\left[\bar{X}_{(1)} - E\left(\bar{X}_{(1)}\right)\right]}{2\,(a\mu_X + b)}$$
$$- \frac{\mu_Y\,(a\alpha g)\left[\bar{X}_{(m)} - E\left(\bar{X}_{(m)}\right)\right]}{2\,(a\mu_X + b)}. \tag{2}$$

Solving (2) and using assumption of symmetry of distribution, the approximate *MSE* of $\hat{\bar{Y}}_{ERSS_a}$ is given by

$$MSE\left(\hat{\bar{Y}}_{ERSS_a}\right) \cong \frac{1}{mr}\left(\sigma^2_{Y[1]} + w^2\sigma^2_{X(1)} - 2w\sigma_{X(1)Y[1]}\right), \tag{3}$$

where $w = \dfrac{\mu_Y\,(a\alpha g)}{(a\mu_X + b)}$.

Minimizing (3) with respect to *w*, we get the optimum value of *w* i.e.

$$w_{(opt)} = \frac{\sigma_{X(1)Y[1]}}{\sigma^2_{X(1)}}.$$

The minimum *MSE* of $\hat{\bar{Y}}_{ERSS_a}$is given by

$$MSE_{\min}\left(\hat{\bar{Y}}_{ERSS_a}\right) \cong \frac{\sigma^2_{Y[1]}\left(1-\rho^2_{X(1)Y[1]}\right)}{mr}, \tag{4}$$

where $\rho^2_{X(1)Y[1]} = \dfrac{\sigma^2_{X(1)Y[1]}}{\sigma^2_{X(1)}\sigma^2_{Y[1]}}$.

Note that the minimum *MSE* in (4) is equal to the *MSE* of the traditional regression estimator based on single auxiliary variable under $ERSS_a$.

4.2. A Family of ratio estimators using $ERSS_c$

We propose the same family of ratio estimators in $ERSS_c$ as

$$\hat{\bar{Y}}_{ERSS_c} = \bar{Y}_{[c]}\left[\frac{a\mu_X+b}{\alpha\left(a\,\bar{X}_{(c)}+b\right)+(1-\alpha)\left(a\mu_X+b\right)}\right]^g, \tag{5}$$

where α, g, $a \neq 0$ and b are defined earlier.

Using bivariate Taylor series expansion, we have

$$\begin{aligned}\left(\hat{\bar{Y}}_{ERSS_c}-\mu_Y\right) \cong{}& \frac{(m-1)}{2m}\left[\bar{Y}'_{[1]}-E\left(\bar{Y}'_{[1]}\right)\right]+\frac{(m-1)}{2m}\left[\bar{Y}'_{[m]}-E\left(\bar{Y}'_{[m]}\right)\right]\\ &+\frac{\left[\bar{Y}'_{\left[\frac{m+1}{2}\right]}-E\left(\bar{Y}'_{\left[\frac{m+1}{2}\right]}\right)\right]}{m}-\frac{\mu_Y\left(a\alpha g\right)(m-1)\left[\bar{X}'_{(1)}-E\left(\bar{X}'_{(1)}\right)\right]}{2m\left(a\mu_X+b\right)}\\ &-\frac{\mu_Y\left(a\alpha g\right)(m-1)\left[\bar{X}'_{(m)}-E\left(\bar{X}'_{(m)}\right)\right]}{2m\left(a\mu_X+b\right)}-\frac{\mu_Y\left(a\alpha g\right)\left[\bar{X}'_{\left(\frac{m+1}{2}\right)}-E\left(\bar{X}'_{\left(\frac{m+1}{2}\right)}\right)\right]}{m\left(a\mu_X+b\right)}.\end{aligned} \tag{6}$$

Using the assumption of symmetry of distribution, the approximate *MSE* of $\hat{\bar{Y}}_{ERSS_c}$, is given by

$$MSE\left(\hat{\bar{Y}}_{ERSS_c}\right) \cong \frac{1}{mr}\left(\frac{(m-1)\,\sigma^2_{Y[1]}+\sigma^2_{Y\left[\frac{m+1}{2}\right]}}{m}+w^2\frac{(m-1)\,\sigma^2_{X(1)}+\sigma^2_{X\left(\frac{m+1}{2}\right)}}{m}\right.$$

$$\left.-2w\frac{(m-1)\,\sigma_{X(1)Y[1]}+\sigma_{X\left(\frac{m+1}{2}\right)Y\left[\frac{m+1}{2}\right]}}{m}\right), \tag{7}$$

where w is defined earlier.

Also (7) can be written as

$$MSE\left(\hat{\bar{Y}}_{ERSS_c}\right) \cong \frac{1}{mr}\left[\sigma^{2*}_{Y[1]}+w^2\sigma^{2*}_{X(1)}-2w\sigma^{*}_{X(1)Y[1]}\right], \tag{8}$$

where

$$\sigma^{2*}_{Y[1]}=\frac{(m-1)\,\sigma^2_{Y[1]}+\sigma^2_{Y\left[\frac{m+1}{2}\right]}}{m}, \qquad \sigma^{2*}_{X(1)}=\frac{(m-1)\,\sigma^2_{X(1)}+\sigma^2_{X\left(\frac{m+1}{2}\right)}}{m}$$

and

$$\sigma^{*}_{X(1)Y[1]}=\frac{(m-1)\,\sigma_{X(1)Y[1]}+\sigma_{X\left(\frac{m+1}{2}\right)Y\left[\frac{m+1}{2}\right]}}{m}.$$

The minimum MSE of $\hat{\bar{Y}}_{ERSS_c}$ at the optimum value of w given by $w_{(opt)}=\dfrac{\sigma^{*}_{X(1)Y[1]}}{\sigma^{2*}_{X(1)}}$ is

$$MSE_{\min}\left(\hat{\bar{Y}}_{ERSS_c}\right) \cong \frac{\sigma^{2*}_{Y[1]}\left(1-\rho^{2*}_{X(1)Y[1]}\right)}{mr}, \tag{9}$$

where $\rho^{2*}_{X(1)Y[1]}=\dfrac{\sigma^{2*}_{X(1)Y[1]}}{\sigma^{2*}_{X(1)}\sigma^{2*}_{Y[1]}}$.

Note that the minimum MSE in (9) is of similar form to the MSE of the regression estimator based on the single auxiliary variable under $ERSS_c$. Also from (1) and (5), several different forms of ratio and product estimators can be generalized by taking different values of α,g, a and b. It is to be noted that for g = +1 and $g = -1$, we can make the ratio and product family of estimators respectively under $ERSS_a$ and $ERSS_c$ using the single auxiliary variable.

5. Proposed estimators using the two auxiliary variables

5.1. A family of ratio estimators in $ERSS_a$

Following Khoshnevisan *et al.* (2007), we propose a family of ratio estimators in $ERSS_a$ using information on the two auxiliary variables, when ranking is performed on the auxiliary variable Z.

$$\hat{\tilde{Y}}'_{ERSS_a} =$$

$$\bar{Y}_{[a]}\left[\frac{a\mu_X+b}{\alpha_1\left(a\,\bar{X}_{[a]}+b\right)+(1-\alpha_1)\left(a\mu_X+b\right)}\right]^{g_1}\left[\frac{c\mu_Z+d}{\alpha_2\left(c\,\bar{Z}_{(a)}+d\right)+(1-\alpha_2)\left(c\mu_Z+d\right)}\right]^{g_2}, \tag{10}$$

where α_1, α_2, g_1 and g_2 are suitable constants, also a, b, c and d are either real numbers or functions of known parameters for the auxiliary variables X and Z respectively.

Using multivariate Taylor series expansion, we have

$$\begin{aligned}\left(\hat{\tilde{Y}}'_{ERSS_a}-\mu_Y\right) \cong \frac{1}{2}\left[\bar{Y}_{[1]}-E\left(\bar{Y}_{[1]}\right)\right]+\frac{1}{2}\left[\bar{Y}_{[m]}-E\left(\bar{Y}_{[m]}\right)\right]-\frac{\mu_Y\left(a\alpha_1 g_1\right)\left[\bar{X}_{[1]}-E\left(\bar{X}_{[1]}\right)\right]}{2\left(a\mu_X+b\right)}\\ -\frac{\mu_Y\left(a\alpha_1 g_1\right)\left[\bar{X}_{[m]}-E\left(\bar{X}_{[m]}\right)\right]}{2\left(a\mu_X+b\right)}-\frac{\mu_Y\left(c\alpha_2 g_2\right)\left[\bar{Z}_{(1)}-E\left(\bar{Z}_{(1)}\right)\right]}{2\left(c\mu_Z+d\right)}\\ -\frac{\mu_Y\left(c\alpha_2 g_2\right)\left[\bar{Z}_{(m)}-E\left(\bar{Z}_{(m)}\right)\right]}{2\left(c\mu_Z+d\right)}.\end{aligned} \tag{11}$$

Squaring both sides, taking expectation of (11) and using assumption of symmetry of distribution, the MSE of $\hat{\tilde{Y}}'_{ERSS_a}$ is given by

$$MSE\left(\hat{\tilde{Y}}'_{ERSS_a}\right) \cong$$

$$\frac{1}{mr}\left(\sigma^2_{Y[1]}+w_1^2\sigma^2_{X[1]}+w_2^2\sigma^2_{Z(1)}-2w_1\sigma_{X[1]Y[1]}-2w_2\sigma_{Y[1]Z(1)}+2w_1w_2\sigma_{X[1]Z(1)}\right). \tag{12}$$

Minimizing $MSE\left(\hat{\tilde{Y}}'_{ERSS_a}\right)$ with respect to w_1 and w_2, the optimum values of w_1 and w_2, are given by

$$w_{1(opt)}=\frac{\sigma^2_{Z(1)}\sigma_{X[1]Y[1]}-\sigma_{X[1]Z(1)}\sigma_{Y[1]Z(1)}}{\sigma^2_{X[1]}\sigma^2_{Z(1)}-\sigma^2_{X[1]Z(1)}}$$

and

$$w_{2(opt)} = \frac{\sigma^2_{X[1]}\sigma_{Y[1]Z(1)} - \sigma_{X[1]Z(1)}\sigma_{X[1]Y[1]}}{\sigma^2_{X[1]}\sigma^2_{Z(1)} - \sigma^2_{X[1]Z(1)}}.$$

Substituting the optimum values of w_1 and w_2 in (12), we get

$$MSE_{\min}\left(\tilde{\bar{Y}}'_{ERSS_a}\right) \cong \frac{\sigma^2_{Y[1]}\left(1 - R^2_{Y[1].X[1]Z(1)}\right)}{mr}, \tag{13}$$

where $R^2_{Y[1].X[1]Z(1)} = \dfrac{\rho^2_{X[1]Y[1]} + \rho^2_{Y[1]Z(1)} - 2\rho_{X[1]Y[1]}\rho_{Y[1]Z(1)}\rho_{X[1]Z(1)}}{1 - \rho^2_{X[1]Z(1)}}$ is the multiple correlation coefficient of $Y[1]$ on $X[1]$ and $Z(1)$ in $ERSS_a$. The minimum MSE of $\tilde{\bar{Y}}'_{ERSS_a}$ is equal to the MSE of the regression estimator when using the two auxiliary variables.

5.2. A family of ratio estimators in $ERSS_c$

We propose a following family of estimators in $ERSS_c$ using the two auxiliary variables X and Z as

$$\tilde{\bar{Y}}'_{ERSS_c} =$$

$$\bar{Y}_{[c]}\left[\frac{a\mu_X + b}{\alpha_1\left(a\,\bar{X}_{[c]} + b\right) + (1-\alpha_1)(a\mu_X + b)}\right]^{g_1}\left[\frac{c\mu_Z + d}{\alpha_2\left(c\,\bar{Z}_{(c)} + d\right) + (1-\alpha_2)(c\mu_Z + d)}\right]^{g_2}, \tag{14}$$

where α_1,α_2,g_1, g_2, a, b, c and d are suitable constants as described earlier.

Using multivariate Taylor series expansion, we have

$$\begin{aligned}\left(\tilde{\bar{Y}}'_{ERSS_c} - \mu_Y\right) \cong{}& \frac{(m-1)}{2m}\left[\bar{Y}'_{[1]} - E\left(\bar{Y}'_{[1]}\right)\right] + \frac{(m-1)}{2m}\left[\bar{Y}'_{[m]} - E\left(\bar{Y}'_{[m]}\right)\right] \\ & - \frac{1}{m}\left[\bar{Y}'_{\left[\frac{m+1}{2}\right]} - E\left(\bar{Y}'_{\left[\frac{m+1}{2}\right]}\right)\right] - \frac{\mu_Y(a\alpha_1 g_1)(m-1)\left[\bar{X}'_{[1]} - E\left(\bar{X}'_{[1]}\right)\right]}{2m(a\mu_X + b)} \\ & - \frac{\mu_Y(a\alpha_1 g_1)(m-1)\left[\bar{X}'_{[m]} - E\left(\bar{X}'_{[m]}\right)\right]}{2m(a\mu_X + b)} \\ & - \frac{\mu_Y(c\alpha_2 g_2)(m-1)\left[\bar{Z}'_{(1)} - E\left(\bar{Z}'_{(1)}\right)\right]}{2m(c\mu_Z + d)}\end{aligned}$$

$$-\frac{\mu_Y\left(c\alpha_2 g_2\right)(m-1)\left[\bar{Z}'_{(m)}-E\left(\bar{Z}'_{(m)}\right)\right]}{2m\left(c\mu_Z+d\right)}$$

$$-\frac{\mu_Y\left(a\alpha_1 g_1\right)\left[\bar{X}'_{\left[\frac{m+1}{2}\right]}-E\left(\bar{X}'_{\left[\frac{m+1}{2}\right]}\right)\right]}{m\left(a\mu_X+b\right)}-\frac{\mu_Y\left(c\alpha_2 g_2\right)\left[\bar{Z}'_{\left(\frac{m+1}{2}\right)}-E\left(\bar{Z}'_{\left(\frac{m+1}{2}\right)}\right)\right]}{m\left(c\mu_Z+d\right)}. \quad (15)$$

Squaring, taking expectation and using assumption of symmetry of distribution, we have

$$MSE\left(\hat{\bar{Y}}'_{ERSS_c}\right)\cong\frac{1}{mr}\left(\frac{(m-1)\,\sigma^2_{Y[1]}+\sigma^2_{Y\left[\frac{m+1}{2}\right]}}{m}+w_1^2\frac{(m-1)\,\sigma^2_{X[1]}+\sigma^2_{X\left[\frac{m+1}{2}\right]}}{m}\right.$$

$$+w_2^2\frac{(m-1)\,\sigma^2_{Z(1)}+\sigma^2_{Z\left(\frac{m+1}{2}\right)}}{m}-2w_1\frac{(m-1)\,\sigma_{X[1]Y[1]}+\sigma_{X\left[\frac{m+1}{2}\right]Y\left[\frac{m+1}{2}\right]}}{m}$$

$$\left.-2w_2\frac{(m-1)\,\sigma_{Y[1]Z(1)}+\sigma_{Y\left[\frac{m+1}{2}\right]Z\left(\frac{m+1}{2}\right)}}{m}+2w_1w_2\frac{(m-1)\,\sigma_{X[1]Z(1)}+\sigma_{X\left[\frac{m+1}{2}\right]Z\left(\frac{m+1}{2}\right)}}{m}\right). \quad (16)$$

The above expression can be written as

$$MSE\left(\hat{\bar{Y}}'_{ERSS_c}\right)\cong$$

$$\frac{1}{mr}\left[\sigma^{2*}_{Y[1]}+w_1^2\sigma^{2*}_{X[1]}+w_2^2\sigma^{2*}_{Z(1)}-2w_1\sigma^*_{X[1]Y[1]}-2w_2\sigma^*_{Y[1]Z(1)}+2w_1w_2\sigma^*_{X[1]Z(1)}\right], \quad (17)$$

where

$$w_1=\frac{\mu_Y\left(a\alpha_1 g_1\right)}{\left(a\mu_X+b\right)}, \qquad w_2=\frac{\mu_Y\left(c\alpha_2 g_2\right)}{\left(c\mu_Z+d\right)}, \qquad \sigma^{2*}_{Y[1]}=\frac{(m-1)\,\sigma^2_{Y[1]}+\sigma^2_{Y\left[\frac{m+1}{2}\right]}}{m},$$

$$\sigma^{2*}_{X[1]}=\frac{(m-1)\,\sigma^2_{X[1]}+\sigma^2_{X\left[\frac{m+1}{2}\right]}}{m}, \qquad \sigma^{2*}_{Z(1)}=\frac{(m-1)\,\sigma^2_{Z(1)}+\sigma^2_{Z\left(\frac{m+1}{2}\right)}}{m},$$

$$\sigma^*_{X[1]Y[1]}=\frac{(m-1)\,\sigma_{X[1]Y[1]}+\sigma_{X\left[\frac{m+1}{2}\right]Y\left[\frac{m+1}{2}\right]}}{m},$$

$$\sigma^*_{Y[1]Z(1)} = \frac{(m-1)\,\sigma_{Y[1]Z(1)} + \sigma_{Y\left[\frac{m+1}{2}\right]Z\left(\frac{m+1}{2}\right)}}{m}$$

and

$$\sigma^*_{X[1]Z(1)} = \frac{(m-1)\,\sigma_{X[1]Z(1)} + \sigma_{X\left[\frac{m+1}{2}\right]Z\left(\frac{m+1}{2}\right)}}{m}.$$

Using (17), the optimum values of w_1 and w_2 are given by

$$w_{1(opt)} = \frac{\sigma^{2*}_{Z(1)}\sigma^*_{X[1]Y[1]} - \sigma^*_{X[1]Z(1)}\sigma^*_{Y[1]Z(1)}}{\sigma^{2*}_{X[1]}\sigma^{2*}_{Z(1)} - \sigma^{2*}_{X[1]Z(1)}}$$

and

$$w_{2(opt)} = \frac{\sigma^{2*}_{X[1]}\sigma^*_{Y[1]Z(1)} - \sigma^*_{X[1]Z(1)}\sigma^*_{X[1]Y[1]}}{\sigma^{2*}_{X[1]}\sigma^{2*}_{Z(1)} - \sigma^{2*}_{X[1]Z(1)}}.$$

Substituting the optimum values of w_1 and w_2 in (17), we get the minimum *MSE* of $\hat{\bar{Y}}^{\prime}_{ERSS_c}$, which is given by

$$MSE_{\min}\left(\hat{\bar{Y}}^{\prime}_{ERSS_c}\right) \cong \frac{\sigma^{2*}_{Y[1]}\left(1 - R^{2*}_{Y[1].X[1]Z(1)}\right)}{mr}, \tag{18}$$

where

$$R^{2*}_{Y[1].X[1]Z(1)} = \frac{\rho^{2*}_{X[1]Y[1]} + \rho^{2*}_{Y[1]Z(1)} - 2\rho^*_{X[1]Y[1]}\rho^*_{Y[1]Z(1)}\rho^*_{X[1]Z(1)}}{\left(1 - \rho^{2*}_{X[1]Z(1)}\right)}$$

is the multiple correlation coefficient of $Y[1]$ on $X[1]$ and $Z(1)$ in $ERSS_c$. The expression given in (18) is equal to the *MSE* of the regression estimator when using the two auxiliary variables under $ERSS_c$.

Note: For different choices of g_1 and g_2 in (10) and (14), we have

$g_1 = g_2 = 1,$	ratio estimator,
$g_1 = g_2 = -1,$	product estimator,
$g_1 = 1$ and $g_2 = -1,$	ratio-product estimator,
$g_1 = -1$ and $g_2 = 1,$	product-ratio estimator.

6. Simulation study

A simulation study has been made to examine the performance of the considered estimators in *SRS* and *ERSS* for estimating the population mean, when ranking is done on the auxiliary variables X and Z separately. Following Samawi (2002), bivariate random observations were generated from bivariate normal distribution having parameters $\mu_X = 6$, $\mu_Y = 3$, $\sigma_X = \sigma_Y = 1$ and $\rho_{XY} = \pm 0.99$, ± 0.95, ± 0.90, ± 0.70 and ± 0.50. Using 4000 simulations, estimates of *MSE*s for ratio estimators were computed as given in Tables 1-5 (see Appendix). We consider $m(r)$ as 4(2), 4(4), 5(2), 6(2) and 6(4) respectively to study the performances of the ratio estimators under *SRS*, $ERSS_a$ and $ERSS_c$.

Further simulation has also been done for the same family of ratio estimators using the two auxiliary variables. For this trivariate random observations were generated from trivariate normal distribution having parameters $\mu_X = 6$, $\mu_Y = 3$, $\mu_Z = 8$, $\sigma_X = \sigma_Y = \sigma_Z = 1$ and for different values of ρ_{XY}. The correlation coefficients between $(Y,\ Z)$ and $(X,\ Z)$ are assumed to be $\rho_{YZ} = 0.70$ and $\rho_{XZ} = 0.60$ respectively as shown in Tables 6-8, with different sample sizes m and different cycles r. Again 4000 simulations have been made to study the performances of a family of the ratio estimators using the two auxiliary variables.

From Tables 1-5 (see Appendix), it is noted that all considered ratio estimators using the one auxiliary variable (X) perform better under *ERSS* as compared to *SRS* for different values of ρ_{XY}. In the case of using the two auxiliary variables X and Z (see Tables 6-8, Appendix), for $r = 1$ and $r = 2$, *ERSS* again gives more precise estimates as compared to *SRS*. Also as we increase r=1 to $r = 2$, the *MSE* values of each estimator decreases under both *SRS* and *ERSS* schemes.

6.1. Empirical study

In this section, we have illustrated the performance of various estimators of population mean under *SRS* and *ERSS* through natural data sets. *ERSS* performs better than *SRS* in case of symmetric populations. In order to generate the symmetric data from positively skewed data, we have taken the logarithm of the study variable (Y) and the auxiliary variables (X and Z).

Table 9 provides the estimated *MSE* values of all considered estimators using the single auxiliary variable (X) based on 4000 samples drawn with replacement. It is immediate to observe that the proposed estimators under *ERSS* perform better than the estimators based on *SRS*. Among all estimators, the estimator $\hat{\overline{Y}}_{1ERSS_a}$ is more efficient for all values of m.

Table 10 gives the estimated *MSE* values of all considered estimators using the two auxiliary variables (X and Z) based on 4000 samples drawn with replacement. The proposed estimators under *ERSS* also perform better than the estimators based on *SRS*. For this data set, the estimator $\hat{\overline{Y}}'_{1ERSS_a}$ has the smaller *MSE* values than other considered estimators $\hat{\overline{Y}}'_{iERSS_a}$ $(i = 2, 3, 4)$.

7. Conclusion

In the present paper, we have studied the problem of estimating the population mean using single and two auxiliary variables in *ERSS*, when we have known information about the population parameters. A given family of estimators includes several ratio type estimators, which have also been adopted by different authors in *SRS*. We examined the effect of transformations on the same family of estimators in *ERSS*. From Tables 1-5, the estimators $\hat{\bar{Y}}_{4ERSS_a}$ and $\hat{\bar{Y}}_{4ERSS_c}$, with $a = \alpha = g = 1$ and $b = S_X$, perform better than all other estimators when $\rho_{XY} < 0$. In Tables 1-5 for $\rho_{XY} > 0$, the estimator $\hat{\bar{Y}}_{3ERSS_a}$, with $a = \beta_{2X}$, $\alpha = g = 1$ and $b = C_X$, generally give more precise estimates as compared to other estimators. In case of two auxiliary variables (see Tables 7 and 8), the ratio estimators $\hat{\bar{Y}}'_{3ERSS_a}$ and $\hat{\bar{Y}}'_{3ERSS_c}$, with choices $\alpha_1 = \alpha_2 = g_1 = g_2 = 1$, $a = \beta_{2X}$, $b = C_X$, $c = \beta_{2Z}$ and $d = C_Z$, are efficient in all other estimators for all values of ρ_{XY} with different sample size m. Finally, it is recommended to use *ERSS* over *SRS* in symmetric populations, in order to get more precise estimates of population mean.

References

Khoshnevisan, M., Singh, R., Chauhan, P., Sawan, N. & Smarandache, F. (2007). A general family of estimators for estimating population mean using known value of some population parameter(s). *Far East Journal of Theoretical Statistics*, 22, 181-191.

McIntyre, G. A. (1952). A method of unbiased selective sampling using ranked sets. *Australian Journal Agriculture Research*, 3, 385-390.

Murthy, M. N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta, India.

Samawi, H. M. (2002). On ratio estimation using extreme ranked set samples. *Abhath Al-Yarmouk, Basic Sciences and Engineering*, 11, 815-827.

Samawi, H. M., Ahmed, M. S. & Abu-Dayyeh, W. (1996). Estimating the population mean using extreme ranked set sample. *Biometrical Journal*, 38(5), 577-586.

Samawi, H. M., Al-Samarraie, A. Y. A. & Al-Saidy, O. M. (2004). On regression estimators using extreme ranked set samples. *Journal of Science of Technology Sultan, Qaboos Universtiy*, 9, 67-86.

Samawi, H. M. & Muttlak, H. A. (1996). Estimation of ratio using rank set sampling. *Biometrical Journal*, 38, 753-764.

Samawi, H. M. & Saeid, L. J. (2004). Stratified extreme ranked set sample with application to ratio estimators. *Journal of Modern Applied Statistical Methods*, 3(1) 117-133.

Takahasi, K. & Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of Institute of Statistical Mathematics*, 20, 1-31.

Appendix

Table 1: *MSE values of different estimators using SRS and ERSS for $m = 4, r = 2$.*

Estimator	ρ_{XY}	0.99	0.9	0.8	0.7	0.5	−0.99	−0.9	−0.8	−0.7	−0.5
$\hat{\hat{Y}}_{1ERSS_a} = \bar{y}_{[a]}\left[\frac{\mu_X+\rho_{XY}}{\bar{x}_{(a)}+\rho_{XY}}\right]$	*SRS*	0.0421445	0.0459502	0.0536968	0.0768756	0.0966342	0.3431534	0.332503	0.3220496	0.2717683	0.2336323
	ERSS	0.0213953	0.0280724	0.0375345	0.0680077	0.0953278	0.1614094	0.1560954	0.1592598	0.1694472	0.1643586
$\hat{\hat{Y}}_{2ERSS_a} = \bar{y}_{[a]}\left[\frac{\mu_X+C_X}{\bar{x}_{(a)}+C_X}\right]$	*SRS*	0.0342081	0.0394587	0.0459384	0.0698305	0.0957964	0.2810607	0.2736063	0.2645737	0.2302537	0.2125556
	ERSS	0.0186795	0.0259662	0.0352951	0.0643397	0.0906895	0.1374189	0.1399317	0.1437876	0.1578388	0.1538846
$\hat{\hat{Y}}_{3ERSS_a} = \bar{y}_{[a]}\left[\frac{\beta_{2X}\mu_X+C_X}{\beta_{2X}\bar{x}_{(a)}+C_X}\right]$	*SRS*	0.0343591	0.0376966	0.046057	0.0689711	0.0985591	0.2839814	0.2827303	0.2785348	0.2435255	0.2236065
	ERSS	0.0178802	0.0247913	0.0346036	0.0654427	0.0937825	0.1419155	0.1458068	0.141693	0.1573677	0.1596848
$\hat{\hat{Y}}_{4ERSS_a} = \bar{y}_{[a]}\left[\frac{\mu_X+S_X}{\bar{x}_{(a)}+S_X}\right]$	*SRS*	0.0431915	0.0477002	0.0535503	0.0718311	0.094948	0.2536793	0.2470929	0.2436818	0.2271933	0.2072925
	ERSS	0.0227204	0.0286125	0.0369096	0.0703841	0.09294	0.1271421	0.1330668	0.1314235	0.1430808	0.1503416

Table 2: *MSE values of different estimators using SRS and ERSS for $m = 4, r = 4$.*

Estimator	ρ_{XY}	0.99	0.9	0.8	0.7	0.5	−0.99	−0.9	−0.8	−0.7	−0.5
$\hat{\hat{Y}}_{1ERSS_a} = \bar{y}_{[a]}\left[\frac{\mu_X+\rho_{XY}}{\bar{x}_{(a)}+\rho_{XY}}\right]$	*SRS*	0.021717	0.0235508	0.0249642	0.0366157	0.0464009	0.1658218	0.1563325	0.152332	0.1354153	0.1171365
	ERSS	0.0107867	0.0147103	0.0190514	0.0347934	0.0460406	0.0783965	0.0774434	0.0807905	0.0812439	0.0815075
$\hat{\hat{Y}}_{2ERSS_a} = \bar{y}_{[a]}\left[\frac{\mu_X+C_X}{\bar{x}_{(a)}+C_X}\right]$	*SRS*	0.0174917	0.0190349	0.0221608	0.0349446	0.045639	0.1358885	0.1438294	0.1313778	0.1232827	0.1116966
	ERSS	0.0092618	0.0131617	0.01722	0.0320217	0.0465724	0.0681436	0.0708875	0.0721552	0.0742076	0.0752605
$\hat{\hat{Y}}_{3ERSS_a} = \bar{y}_{[a]}\left[\frac{\beta_{2X}\mu_X+C_X}{\beta_{2X}\bar{x}_{(a)}+C_X}\right]$	*SRS*	0.0168429	0.0192322	0.0216478	0.0353861	0.0478808	0.1430786	0.1383039	0.1367204	0.116574	0.1148881
	ERSS	0.0089404	0.0123172	0.0166099	0.0348346	0.0480005	0.0686419	0.0688538	0.0731911	0.0790872	0.0751499
$\hat{\hat{Y}}_{4ERSS_a} = \bar{y}_{[a]}\left[\frac{\mu_X+S_X}{\bar{x}_{(a)}+S_X}\right]$	*SRS*	0.0218861	0.0244523	0.0257768	0.0381466	0.0447497	0.123449	0.1257248	0.12019	0.1146243	0.1009705
	ERSS	0.0110367	0.0144961	0.0186961	0.0342148	0.0485601	0.0637671	0.0646209	0.0668673	0.0728451	0.0727519

Table 3: *MSE values of different estimators using SRS and ERSS for $m = 6, r = 2$.*

Estimator	ρ_{XY}	0.99	0.9	0.8	0.7	0.5	−0.99	−0.9	−0.8	−0.7	−0.5
$\hat{\hat{Y}}_{1ERSS_a} = \bar{y}_{[a]}\left[\frac{\mu_X + \rho_{XY}}{\bar{x}_{(a)} + \rho_{XY}}\right]$	*SRS*	0.0276638	0.0310981	0.0339876	0.0492321	0.0623409	0.2198379	0.2056067	0.2040597	0.1750035	0.15697
	ERSS	0.0125093	0.0179924	0.0232678	0.0444828	0.0625907	0.0909455	0.0915863	0.0956375	0.0963579	0.1004085
$\hat{\hat{Y}}_{2ERSS_a} = \bar{y}_{[a]}\left[\frac{\mu_X + C_X}{\bar{x}_{(a)} + C_X}\right]$	*SRS*	0.0228799	0.028244	0.0311799	0.0452084	0.0644545	0.1936849	0.1822071	0.1833652	0.1721255	0.1467526
	ERSS	0.0102854	0.0154112	0.0217249	0.0437198	0.0611709	0.0788966	0.0813828	0.08313	0.0966679	0.098809
$\hat{\hat{Y}}_{3ERSS_a} = \bar{y}_{[a]}\left[\frac{\beta_{2X}\mu_X + C_X}{\beta_{2X}\bar{x}_{(a)} + C_X}\right]$	*SRS*	0.023543	0.0266576	0.0307184	0.0468753	0.0627374	0.1894285	0.187493	0.1834673	0.1657038	0.148337
	ERSS	0.0101325	0.0154559	0.0211064	0.0454989	0.0635411	0.0769606	0.0848768	0.08612	0.0936018	0.0972394
$\hat{\hat{Y}}_{4ERSS_a} = \bar{y}_{[a]}\left[\frac{\mu_X + S_X}{\bar{x}_{(a)} + S_X}\right]$	*SRS*	0.0284588	0.0308181	0.0338651	0.048739	0.0646897	0.1614895	0.1726636	0.1640717	0.1508197	0.1369065
	ERSS	0.0128794	0.0173838	0.0238402	0.0439633	0.0633808	0.0704374	0.0764649	0.0776193	0.0874932	0.0983925

Table 4: *MSE values of different estimators using SRS and ERSS for $m = 6, r = 4$.*

Estimator	ρ_{XY}	0.99	0.9	0.8	0.7	0.5	−0.99	−0.9	−0.8	−0.7	−0.5
$\hat{\hat{Y}}_{1ERSS_a} = \bar{y}_{[a]}\left[\frac{\mu_X + \rho_{XY}}{\bar{x}_{(a)} + \rho_{XY}}\right]$	*SRS*	0.0133039	0.0148865	0.0170387	0.0239208	0.0312614	0.1091832	0.1102434	0.0999666	0.090003	0.0764261
	ERSS	0.0065831	0.0088927	0.0114613	0.0226373	0.0310097	0.0456315	0.0447123	0.0472675	0.0501376	0.0510662
$\hat{\hat{Y}}_{2ERSS_a} = \bar{y}_{[a]}\left[\frac{\mu_X + C_X}{\bar{x}_{(a)} + C_X}\right]$	*SRS*	0.0117771	0.0131748	0.0154717	0.0228518	0.030882	0.0919903	0.0920865	0.0866318	0.0833786	0.0718192
	ERSS	0.0052139	0.007885	0.0110488	0.0216899	0.0321166	0.0385758	0.0400184	0.0407774	0.0457131	0.0481164
$\hat{\hat{Y}}_{3ERSS_a} = \bar{y}_{[a]}\left[\frac{\beta_{2X}\mu_X + C_X}{\beta_{2X}\bar{x}_{(a)} + C_X}\right]$	*SRS*	0.011078	0.013142	0.0149048	0.0243094	0.0308484	0.0915886	0.0917993	0.0891019	0.079593	0.0754453
	ERSS	0.0048665	0.0075851	0.0106771	0.0222789	0.0308866	0.0390286	0.0414351	0.0413122	0.0467898	0.0487322
$\hat{\hat{Y}}_{4ERSS_a} = \bar{y}_{[a]}\left[\frac{\mu_X + S_X}{\bar{x}_{(a)} + S_X}\right]$	*SRS*	0.013897	0.0148891	0.0174452	0.0236295	0.0314632	0.0877026	0.0829768	0.0808907	0.0719681	0.0665056
	ERSS	0.0062653	0.0092167	0.0120893	0.022158	0.0312568	0.0361718	0.0369842	0.0391167	0.0410935	0.0468384

Table 5: MSE values of different estimators using SRS and ERSS for $m = 5, r = 2$.

Estimator	ρ_{XY}	0.99	0.9	0.8	0.7	0.5	−0.99	−0.9	−0.8	−0.7	−0.5
$\hat{\hat{Y}}_{1ERSS_c} = \bar{y}_{[c]}\left[\frac{\mu_X+\rho_{XY}}{\bar{x}_{(c)}+\rho_{XY}}\right]$	*SRS*	0.0337842	0.0364239	0.0406543	0.0606875	0.078297	0.262396	0.2657483	0.255815	0.2125039	0.1911728
	ERSS	0.0156276	0.0217249	0.028424	0.0537139	0.0739129	0.1054132	0.1133406	0.1088354	0.1215593	0.1208256
$\hat{\hat{Y}}_{2ERSS_c} = \bar{y}_{[c]}\left[\frac{\mu_X+C_X}{\bar{x}_{(c)}+C_X}\right]$	*SRS*	0.0272919	0.0313738	0.0354903	0.0559322	0.0752243	0.2202212	0.2207612	0.2131319	0.1939198	0.1723402
	ERSS	0.013084	0.0188902	0.0268646	0.0524263	0.0737738	0.091501	0.0921889	0.0989449	0.1129062	0.1136797
$\hat{\hat{Y}}_{3ERSS_c} = \bar{y}_{[c]}\left[\frac{\beta_{2X}\mu_X+C_X}{\beta_{2X}\bar{x}_{(c)}+C_X}\right]$	*SRS*	0.0277326	0.0304337	0.0352675	0.0577364	0.0741742	0.2386541	0.2225659	0.21779	0.1927738	0.182693
	ERSS	0.0123854	0.0193276	0.0258538	0.0518648	0.0761968	0.0972475	0.0981524	0.1052062	0.1178463	0.1166842
$\hat{\hat{Y}}_{4ERSS_c} = \bar{y}_{[c]}\left[\frac{\mu_X+S_X}{\bar{x}_{(c)}+S_X}\right]$	*SRS*	0.0353522	0.0371511	0.0436183	0.0577978	0.0757769	0.2082045	0.2042378	0.2015726	0.1789549	0.1646477
	ERSS	0.0147976	0.0206443	0.0283466	0.053068	0.0742967	0.0853865	0.0904359	0.0950412	0.1041661	0.1122718

Table 6: MSE values of different estimators using SRS and ERSS for $m = 4, r = 1$.

Estimator	ρ_{XY}	0.99	0.95	0.90	0.80	0.75
$\hat{\hat{Y}}'_{1ERSS_a} = \bar{y}_{[a]}\left[\frac{\mu_X+\rho_{XY}}{\bar{x}_{[a]}+\rho_{XY}}\right]\left[\frac{\mu_Z+\rho_{YZ}}{\bar{z}_{(a)}+\rho_{YZ}}\right]$	*SRS*	0.0407724	0.0501944	0.0592333	0.0824846	0.0930625
	ERSS	0.0389418	0.0395594	0.0377836	0.0373521	0.0376105
$\hat{\hat{Y}}'_{2ERSS_a} = \bar{y}_{[a]}\left[\frac{\mu_X+C_X}{\bar{x}_{[a]}+C_X}\right]\left[\frac{\mu_Z+C_Z}{\bar{z}_{(a)}+C_Z}\right]$	*SRS*	0.0325959	0.0391727	0.0554632	0.0779341	0.0896016
	ERSS	0.0289404	0.029944	0.0299202	0.0294217	0.0303673
$\hat{\hat{Y}}'_{3ERSS_a} = \bar{y}_{[a]}\left[\frac{\beta_{2X}\mu_X+C_X}{\beta_{2X}\bar{x}_{[a]}+C_X}\right]\left[\frac{\beta_{2Z}\mu_Z+C_Z}{\beta_{2Z}\bar{z}_{(a)}+C_Z}\right]$	*SRS*	0.0311288	0.0393582	0.0503898	0.0760209	0.0873974
	ERSS	0.0269807	0.0272861	0.0293202	0.0304836	0.0275289
$\hat{\hat{Y}}'_{4ERSS_a} = \bar{y}_{[a]}\left[\frac{\mu_X+S_X}{\bar{x}_{[a]}+S_X}\right]\left[\frac{\mu_Z+S_Z}{\bar{z}_{(a)}+S_Z}\right]$	*SRS*	0.0446412	0.0502939	0.0623651	0.0846077	0.0935501
	ERSS	0.0386108	0.0411331	0.0394753	0.0393358	0.0392025

Table 7: MSE values of different estimators using SRS and ERSS for $m = 5, r = 1$.

Estimator	ρ_{XY}	0.99	0.95	0.90	0.80	0.75
$\hat{\hat{Y}}'_{1ERSS_c} = \bar{y}_{[c]}\left[\frac{\mu_X+\rho_{XY}}{\bar{x}_{[c]}+\rho_{XY}}\right]\left[\frac{\mu_Z+\rho_{YZ}}{\bar{z}_{(c)}+\rho_{YZ}}\right]$	*SRS*	0.0317161	0.0380387	0.0458682	0.0654852	0.0720379
	ERSS	0.0299674	0.028886	0.0288459	0.0299894	0.0306687
$\hat{\hat{Y}}'_{2ERSS_c} = \bar{y}_{[c]}\left[\frac{\mu_X+C_X}{\bar{x}_{[c]}+C_X}\right]\left[\frac{\mu_Z+C_Z}{\bar{z}_{(c)}+C_Z}\right]$	*SRS*	0.0242259	0.0321638	0.0398709	0.0599825	0.0709018
	ERSS	0.0221972	0.0233933	0.022709	0.0232737	0.0233012
$\hat{\hat{Y}}'_{3ERSS_c} = \bar{y}_{[c]}\left[\frac{\beta_{2X}\mu_X+C_X}{\beta_{2X}\bar{x}_{[c]}+C_X}\right]\left[\frac{\beta_{2Z}\mu_Z+C_Z}{\beta_{2Z}\bar{z}_{(c)}+C_Z}\right]$	*SRS*	0.0238334	0.0311472	0.0413165	0.0602494	0.0688909
	ERSS	0.0214479	0.0212721	0.0222061	0.0205364	0.0219412
$\hat{\hat{Y}}'_{4ERSS_c} = \bar{y}_{[c]}\left[\frac{\mu_X+S_X}{\bar{x}_{[c]}+S_X}\right]\left[\frac{\mu_Z+S_Z}{\bar{z}_{(c)}+S_Z}\right]$	*SRS*	0.0330982	0.0391957	0.049966	0.069275	0.0761463
	ERSS	0.029793	0.0300217	0.0308689	0.0312941	0.0305486

Table 8: MSE values of different estimators using SRS and ERSS for $m = 5, r = 2$.

Estimator	ρ_{XY}	0.99	0.95	0.90	0.80	0.75
$\hat{\hat{Y}}'_{1ERSS_c} = \bar{y}_{[c]}\left[\frac{\mu_X+\rho_{XY}}{\bar{x}_{[c]}+\rho_{XY}}\right]\left[\frac{\mu_Z+\rho_{YZ}}{\bar{z}_{(c)}+\rho_{YZ}}\right]$	*SRS*	0.0159022	0.0182493	0.0223628	0.0317303	0.0351792
	ERSS	0.0149805	0.0146252	0.0147981	0.0150615	0.0149301
$\hat{\hat{Y}}'_{2ERSS_c} = \bar{y}_{[c]}\left[\frac{\mu_X+C_X}{\bar{x}_{[c]}+C_X}\right]\left[\frac{\mu_Z+C_Z}{\bar{z}_{(c)}+C_Z}\right]$	*SRS*	0.0115080	0.0154952	0.0197208	0.0313548	0.0351996
	ERSS	0.0110513	0.0109730	0.0110084	0.0111618	0.0111697
$\hat{\hat{Y}}'_{3ERSS_c} = \bar{y}_{[c]}\left[\frac{\beta_{2X}\mu_X+C_X}{\beta_{2X}\bar{x}_{[c]}+C_X}\right]\left[\frac{\beta_{2Z}\mu_Z+C_Z}{\beta_{2Z}\bar{z}_{(c)}+C_Z}\right]$	*SRS*	0.0109745	0.0143726	0.0195953	0.0311480	0.0356664
	ERSS	0.0107352	0.0106756	0.0103141	0.0105737	0.0103239
$\hat{\hat{Y}}'_{4ERSS_c} = \bar{y}_{[c]}\left[\frac{\mu_X+S_X}{\bar{x}_{[c]}+S_X}\right]\left[\frac{\mu_Z+S_Z}{\bar{z}_{(c)}+S_Z}\right]$	*SRS*	0.0160405	0.0187243	0.0234514	0.0316207	0.0367001
	ERSS	0.0146324	0.0153025	0.0146813	0.0143856	0.0153041

Population-I: Source: Murthy (1967).

$\log(Y)$: output of a factory, $\log(X)$: fixed capital.

$N = 80$, $m = 4, 6, 8$, $r = 1$, $\mu_Y = 8.480904$, $\mu_X = 6.750716$ and $\rho_{XY} = 0.9640175$.

Table 9: Estimated MSE values.

Estimator	*SRS* and *ERSS*	$m = 4$	$m = 6$	$m = 8$
$\hat{\bar{Y}}_{1ERSS_a} = \bar{y}_{[a]}\left[\dfrac{\mu_X + \rho_{XY}}{\bar{x}_{(a)} + \rho_{XY}}\right]$	*SRS*	0.05314492	0.03622249	0.02676690
	ERSS	0.02439540	0.01409169	0.01021740
$\hat{\bar{Y}}_{2ERSS_a} = \bar{y}_{[a]}\left[\dfrac{\mu_X + C_X}{\bar{x}_{(a)} + C_X}\right]$	*SRS*	0.07966297	0.05430063	0.04004554
	ERSS	0.03555703	0.01935204	0.01299920
$\hat{\bar{Y}}_{3ERSS_a} = \bar{y}_{[a]}\left[\dfrac{\beta_{2X}\mu_X + C_X}{\beta_{2X}\bar{x}_{(a)} + C_X}\right]$	*SRS*	0.08237797	0.05614814	0.04140083
	ERSS	0.03670370	0.01989512	0.01329300
$\hat{\bar{Y}}_{4ERSS_a} = \bar{y}_{[a]}\left[\dfrac{\mu_X + S_X}{\bar{x}_{(a)} + S_X}\right]$	*SRS*	0.05842170	0.03982448	0.02941517
	ERSS	0.02660899	0.01513016	0.01075533

Population-II: Source: Murthy (1967).

$\log(Y)$: output of a factory, $\log(X)$: fixed capital and $\log(Z)$: number of workers.

$N = 80, m = 4, 6, 8, r = 1, \mu_Y = 8.480904, \mu_X = 6.750716, \mu_Z = 5.233816,$

$\rho_{XY} = 0.9640175$ and $\rho_{YZ} = 0.916134$.

Table 10: Estimated MSE values.

Estimator	*SRS* and *ERSS*	$m = 4$	$m = 6$	$m = 8$
$\hat{\bar{Y}}'_{1ERSS_a} = \bar{y}_{[a]}\left[\dfrac{\mu_X + \rho_{XY}}{\bar{x}_{[a]} + \rho_{XY}}\right]\left[\dfrac{\mu_Z + \rho_{YZ}}{\bar{z}_{(a)} + \rho_{YZ}}\right]$	*SRS*	0.7599230	0.4833926	0.3632148
	ERSS	0.2603936	0.1286980	0.0919748
$\hat{\bar{Y}}'_{2ERSS_a} = \bar{y}_{[a]}\left[\dfrac{\mu_X + C_X}{\bar{x}_{[a]} + C_X}\right]\left[\dfrac{\mu_Z + C_Z}{\bar{z}_{(a)} + C_Z}\right]$	*SRS*	1.0409530	0.6582852	0.4931993
	ERSS	0.3510492	0.1686980	0.1163822
$\hat{\bar{Y}}'_{3ERSS_a} = \bar{y}_{[a]}\left[\dfrac{\beta_{2X}\mu_X + C_X}{\beta_{2X}\bar{x}_{[a]} + C_X}\right]\left[\dfrac{\beta_{2Z}\mu_Z + C_Z}{\beta_{2Z}\bar{z}_{(a)} + C_Z}\right]$	*SRS*	1.0751500	0.6794242	0.5088861
	ERSS	0.3618654	0.1734536	0.1193290
$\hat{\bar{Y}}'_{4ERSS_a} = \bar{y}_{[a]}\left[\dfrac{\mu_X + S_X}{\bar{x}_{[a]} + S_X}\right]\left[\dfrac{\mu_Z + S_Z}{\bar{z}_{(a)} + S_Z}\right]$	*SRS*	0.7816216	0.4970571	0.3733052
	ERSS	0.2679331	0.1318287	0.0934524

Selected articles from
Congreso Español de Biometría 2009

SORT 34 (1) January-June 2010, 67-78

Modelling spatial patterns of distribution and abundance of mussel seed using Structured Additive Regression models

María P. Pata[1], María Xosé Rodríguez-Álvarez[2,3], Vicente Lustres-Pérez[1]
Eugenio Fernández-Pulpeiro[1], Carmen Cadarso-Suárez[2,3]

Abstract

As mussel farming depends on sources of natural mussel seed, knowledge of factors is required to regulate both the spatial distribution and abundance of this resource. These spatial patterns were modelled using Bayesian STructured Additive Regression (STAR) models for categorical data, based on a mixed-model representation. We used Bayesian penalized splines for modelling the continuous covariate effects and a Markov random field prior for estimating the spatial effects.

MSC: 62F15, 62G08, 62P12, 92D40.

Keywords: Mussel seed, Bayesian structured additive regression (STAR) models, spatial effects, Bayesian P-splines.

1. Introduction

Knowledge of spatial patterns of distribution and abundance of species is essential in order to understand the ecological processes that have generated such processes (Underwood, Chapman and Connell (2000)). In the case of marine resources, knowledge of these patterns is of crucial interest.

Mussel farming is widely developed along most of Galicia's Atlantic coastline, and indeed this region is the largest producer in Europe (200,000 MT/year). As mussel

[1] Departamento de Zoología y Antropología Física, Universidade de Santiago de Compostela (USC), Spain.

[2] Unidad de Bioestadística, Departamento de Estadística e Investigación Operativa, USC, Spain.

[3] Instituto de Investigación Sanitaria de Santiago (IDIS), Santiago de Compostela, Spain.

Received: November 2009
Accepted: May 2010

farming depends on natural mussel seed resources, knowledge of its distribution and abundance is fundamental to prevent depletion of natural populations.

In ecology, Generalized Linear Models (GLM, McCullagh and Nelder (1997)) are the most widely used statistical models to assess relationships between species distribution and environment. In recent years, however, biomedical researchers have shown a great interest in the use of Generalized Additive Models (GAM, Hastie and Tibshirani (1990); Wood (2006)), due the latter's ability to cover the complex non-linear effects had by continuous covariates on the outcome of interest. Recent applications of GAMs in ecology (see, for instance, Austin (2002), Austin (2007), Guisan, Edwards and Hastie (2002)) show that GAM regression models are useful tools for analysing relationships between species' distributions and their environment. Yet, spatial autocorrelation often exists in the data because the sample points are close to one another and subject to the same environmental factors (see Kneib, Müller and Hothorn (2008)). Since spatial correlation is difficult to handle within a GAM framework, a more general regression model is thus called for.

Accordingly, this study modelled the spatial distribution of mussel seed within a Bayesian STructured Additive Regression Model (STAR, Fahrmeir and Lang (2001)) framework. Inference was based on a mixed-model representation (Kneib and Fahrmeir (2006)). The use of STAR models affords several advantages when analysing spatial data, including, among others, the possibility of incorporating: (a) flexible forms of the effects of continuous covariates, by using Bayesian P-splines (Eilers and Marx (1996), Lang and Brezger (2004)); (b) flexible spatial effects; and, (c) random effects to explain the overdispersion caused by unobserved heterogeneity or the presence of autocorrelation in spatial data (Fahrmeir and Lang (2001)). Models that enable smooth effects of continuous covariates and spatial effects with flexible forms to be incorporated are known as geoadditive models (Kammann and Wand (2003)). In this paper, we used a geoadditive multicategorical regression model (Kneib and Fahrmeir (2006)), in which the response variable was assumed to follow a multinomial distribution.

The paper is structured as follows: the mussel seed data are introduced in Section 2; the statistical methodology is described in Section 3; the results from fitting the proposed STAR models to mussel seed data are shown in Section 4; and the paper concludes with a Discussion Section.

2. The mussel seed data

This study was undertaken during spring tides at 62 sites along Galicia's Atlantic seaboard, between $43^\circ 21'$ N, $8^\circ 21'$ W and $42^\circ 44'$ N, $9^\circ 04'$ W, from March to September in 2005 and 2006.

At each site, a transect perpendicular to the coastline was placed in the intertidal zone. A sample quadrant (20×20 cm) was set at 50-centimetre intervals and the percentage cover of mussel seed then measured. Information from a set of covariates was taken in order to explain the distribution pattern of the mussel seed. These covariates were tidal height (in metres), percentage of pools, and positioning related to cardinal points divided into the following five categories: NN; NE; SE; SW; and NW.

For study purposes, the outcome of interest was percentage cover (from 5% upwards, in multiples of 5%). This variable was treated as categorical, and the following four categories were established:

Category 1: low abundance, [0% - 5%]. This was used as the reference category;
Category 2: medium, (5% - 25%];
Category 3: high, (25% - 50%];
Category 4: very high, $>$ 50%.

Within the STAR framework, several approaches can be used to analyse categorical responses, such as the multinomial model for nominal categories or the cumulative logit probit models, among others, for ordered categories. Despite the fact that a better option might have been the cumulative model, we nevertheless chose to use a multinomial model in view of the biological interest that this option could afford.

All computations were performed using the BayesX package (Belitz *et al.* (2009)).

3. Statistical methodology: geoadditive multicategorical regression model

In multicategorical data the response variable Y is observed in categories $r \in (1,\ldots,k)$. Analysis of this type of data calls for an appropriate model to take into account the additional information supplied by these categories (Boeck and Wilson (2004)). In this paper, a multinomial logit model was considered, with the probability of the category r expressed as follows:

$$P(Y=r|u)=\pi^{(r)}=h^{(r)}\left(\eta^{(1)},\ldots,\eta^{(q)}\right)=\frac{\exp\left(\eta^{(r)}\right)}{1+\sum_{s=1}^{q}\exp\left(\eta^{(s)}\right)},\ r=1,\ldots,q=k-1,$$

with k as reference category, and the linear predictor $\eta^{(r)}=u'\alpha^{(r)}$, depending on covariates u and category-specific vector of regression coeficients $\alpha^{(r)}$. It is possible to obtain the general multinomial model

$$\pi=h(\eta),\qquad \eta=V\gamma,$$

by defining the design matrix

$$V=\begin{pmatrix} \upsilon_1' \\ \vdots \\ \upsilon_q' \end{pmatrix}=\begin{pmatrix} u' & & 0 \\ & \ddots & \\ 0 & & u' \end{pmatrix}$$

and the overall vector of regression parameters (Kneib (2006); Kneib and Fahrmeir (2006))

$$\gamma = \left(\alpha^{(1)}, \ldots, \alpha^{(q)}\right).$$

To take into account the spatial information for each unit (administrative areas in our example), the following geoadditive multicategorical model (defined by the geoadditive predictor) is then considered

$$\eta_i^{(r)} = u_i'\alpha^{(r)} + f_1^{(r)}\left(x_{i1}\right) + \ldots + f_l^{(r)}\left(x_{il}\right) + f_{spat}^{(r)}\left(s_i\right),$$

where $f_1^{(r)}, \ldots, f_l^{(r)}$ are unknown smooth functions of the covariates $x_1, \ldots, x_l$, and $f_{spat}^{(r)}$ is the non-linear effect of spatial index $s_i \in \{1, \ldots, S\}$ (administrative area in our example).

This specification of the model allows for flexible incorporation of non-linear effects of continuous covariates and spatial effects. Furthermore, the different types of covariates are considered in a unified framework (Fahrmeir and Lang (2001); Kneib and Fahrmeir (2006)).

Since spatial correlation and/or heterogeneity due to unobserved spatially varying covariates are usually present in spatial data, it seems appropiate for the spatial effect to be broken down into a spatially correlated part (structured part: f_{str}) and a spatially uncorrelated part (unstructured part: f_{unstr}):

$$f_{spat}^{(r)}(s) = f_{str}^{(r)}(s) + f_{unstr}^{(r)}(s).$$

This representation of the spatial effects makes it possible to distinguish between the two kinds of unobserved covariates, namely, those that display a strong spatial structure and those that are present locally (Besag, York and Mollié (1991); Fahrmeir *et al.* (2003)).

To estimate smooth effect functions and model parameters, an empirical Bayesian approach based on mixed model representation is used. Assigning appropriate priors for parameters and functions is crucial. For the fixed effects parameter γ, diffuse priors $p(\gamma) \propto const$ are asssumed.

For specifying smoothness priors for continous covariates, a Bayesian version of the P-splines approach of Eilers and Marx (1996) is used (Lang and Brezger (2004)). This approach assumes that the effect f of a covariate x can be approximated by a polinomial spline of degree l defined on a set of equally spaced knots $x_{min} = \xi_0 < \xi_1 < \ldots < \xi_{r-1} < \xi_r = x_{max}$. This can be written in terms of a linear combination of $M_j = r_j + l_j$ B-spline basis functions

$$f_j(x) = \sum_{m=1}^{M_j} \beta_{jm} B_m(x),$$

where β_j is the vector of the unknown regression coefficients.

The main problem when dealing with these splines lies in the selection of the number of knots and their placement. The idea of P-splines is to select a generous number of knots and define a roughness penalty on adjacent regression coefficients to regularise the problem and avoid overfitting (Eilers and Marx (1996)). In the frequentist approach, first- or second-order differences are usually used. From a Bayesian perspective, these are replaced by their stochastic analogues, namely, first-or second-order random walks. For the purposes of this study, we used second-order random walks for the regression coefficients, defined as

$$\beta_{jm} = 2\beta_{j,m-1} - \beta_{j,m-2} + u_{jm},$$

with Gaussian errors $u_{jm} \sim N\left(0, \tau_j^2\right)$. (Lang and Brezger (2004)). The variance parameter τ_j^2 controls the amount of smoothnes.

Since the spatial locations are clustered in connected geographical regions, a Markov Random Field prior (Besag *et al.* (1991)) is selected for the structured spatial effects. This spatial smoothness prior is defined by

$$\left\{ f_{str}(s) | f_{str}(s'); \, s \neq s', \tau^2 \right\} \sim N\left(\sum_{s \in \delta_s} \frac{f_{str}(s')}{N_s}, \frac{\tau^2}{N_s} \right),$$

where N_s is the number of adjacent sites, $s \in \delta_s$ ndicates that site $s^{'}$ is neighbour of site s, that is, they share a common boundary (Fahrmeir and Lang (2001), Kneib (2006)).

The unstructured spatial effects are assumed to be i.i.d. random effects $f_{unstr}(s) \sim N\left(0, \tau^2\right)$ (Fahrmeir *et al.* (2004); Kneib and Fahrmeir 2006).

Inference is performed with empirical Bayes (EB) posterior analysis based on generalized linear mixed model (GLMM) methodology, once an appropiate reparameterization of the regression terms is given. For empirical Bayes inference, the variances τ_j^2 are considered as unknown constants to be estimated from their marginal likelihood. Based on the GLMM approach regression and variance parameters can be estimated using iteratively weighted least squares (IWLS) and (approximate) restricted maximum likelihood (REML) developed for GLMM's. For detailed description of the estimation procedure see Fahrmeir *et al.* (2004) and Kneib and Fahrmeir (2006).

4. Results

To analyse the spatial distribution of mussel seed with respect to the relevant explanatory variables, a geoadditive multinomial logit model was applied. The parametric effects of sites' positioning in terms of cardinal points as well as the smooth effects of tidal height and percentage of pools were included in the model.

A summary of the estimated effects of site positioning for each category is shown in Table 1. The category with the highest frequency, SW, was chosen as the reference category. As can be seen from Table 1, the results were only significant for Category 2 (mussel seed abundance of 5%-25%), and as NN-, NE- and NW-positioning of sites reduced the presence of this category, SW-positioning was therefore the best for the presence of Category 2.

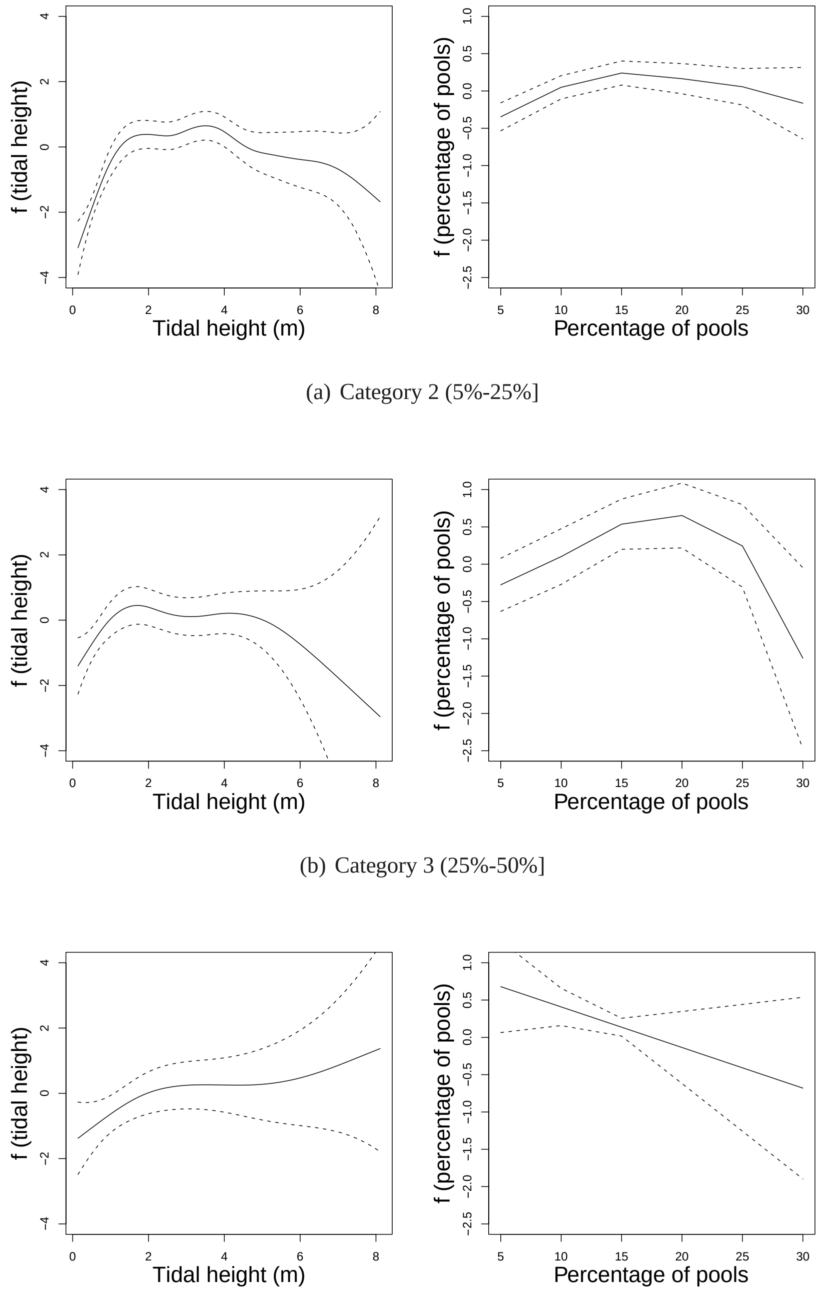
Table 1: *Estimates, standard deviations (S.D) and 95% credible confidence interval for the fixed effects. Category 1 (< 5%) is taken as reference category.*

	Estimated effects	S.D.	95%CI	
Category 2 (5%-25%]				
NN	−1.73	0.662	−3.02	−0.43
NE	−0.86	0.179	−1.21	−0.51
SE	−0.17	0.280	−0.72	0.37
NW	−0.36	0.161	−0.68	−0.04
Category 3 (25%-50%]				
NN	−0.04	0.435	−0.86	0.54
NE	−0.03	0.365	−0.74	0.68
SE	0.41	0.441	−0.44	0.44
NW	0.06	0.251	−0.42	0.25
Category 4 > 50				
NN	−0.49	0.687	−1.04	0.06
NE	−0.09	0.769	−1.60	1.41
SE	0.36	0.784	−1.16	1.90
NW	0.44	0.553	−0.64	1.53

The estimated smooth effects of tidal height and percentage of pools on the presence of mussel seed are shown in Figure 1. As can be seen, these are complex nonlinear effects. Hence, the use of purely linear models to describe such data could lead to estimations and, by extension, conclusions that were erroneous. STAR models enable flexible forms of the effects of continuous covariates to be incorporated in the response and better knowledge of the biological process so obtained.

The effect of tidal height appeared to be similar for the above three categories in the initial metres, with a much more pronounced shape for Category 2. The function increased until a tidal height of about two metres, and then decreased from 4 metres. It seems that the most suitable tidal heights for the presence of mussel seed range from 2 to 4 metres, particularly for Category 2. For Category 3, tidal height appeared to have no effect from a height of about 2 metres.

The effects of percentage of pools are depicted in the right panels of Figure 1. For the presence of Categories 2 and 3, which plotted similar patterns, sites with 15% to 20% of pools would seem to be more suitable. The presence of these categories decreased



(a) Category 2 (5%-25%]

(b) Category 3 (25%-50%]

(c) Category 4 (>50)

Figure 1: *Estimated smooth effects of tidal height (left panel) and percentage of pools (right panel),with 95% pointwise credible intervals. Category 1 (<5%) is taken as reference category.*

thereafter but posterior probabilities were non-significant above 20%. For Category 4 (very high abundance), in contrast, the presence of mussel seed decreased linearly with percentage of pools.

Figure 2 displays the spatial effects on a grey scale, after controlling for the covariates: white colour refers to a positive spatial effect signifying higher abundance of the category, while dark colour refers to a negative effect signifying lower abundance. The significance of the structured spatial effects are shown in the third column of Figure 2, with black areas indicating strictly negative credible intervals, white ones indicating strictly positive, and grey areas indicating no effect. In cases where the structured spatial effects proved non-significant, the map of posterior probabilities is not shown.

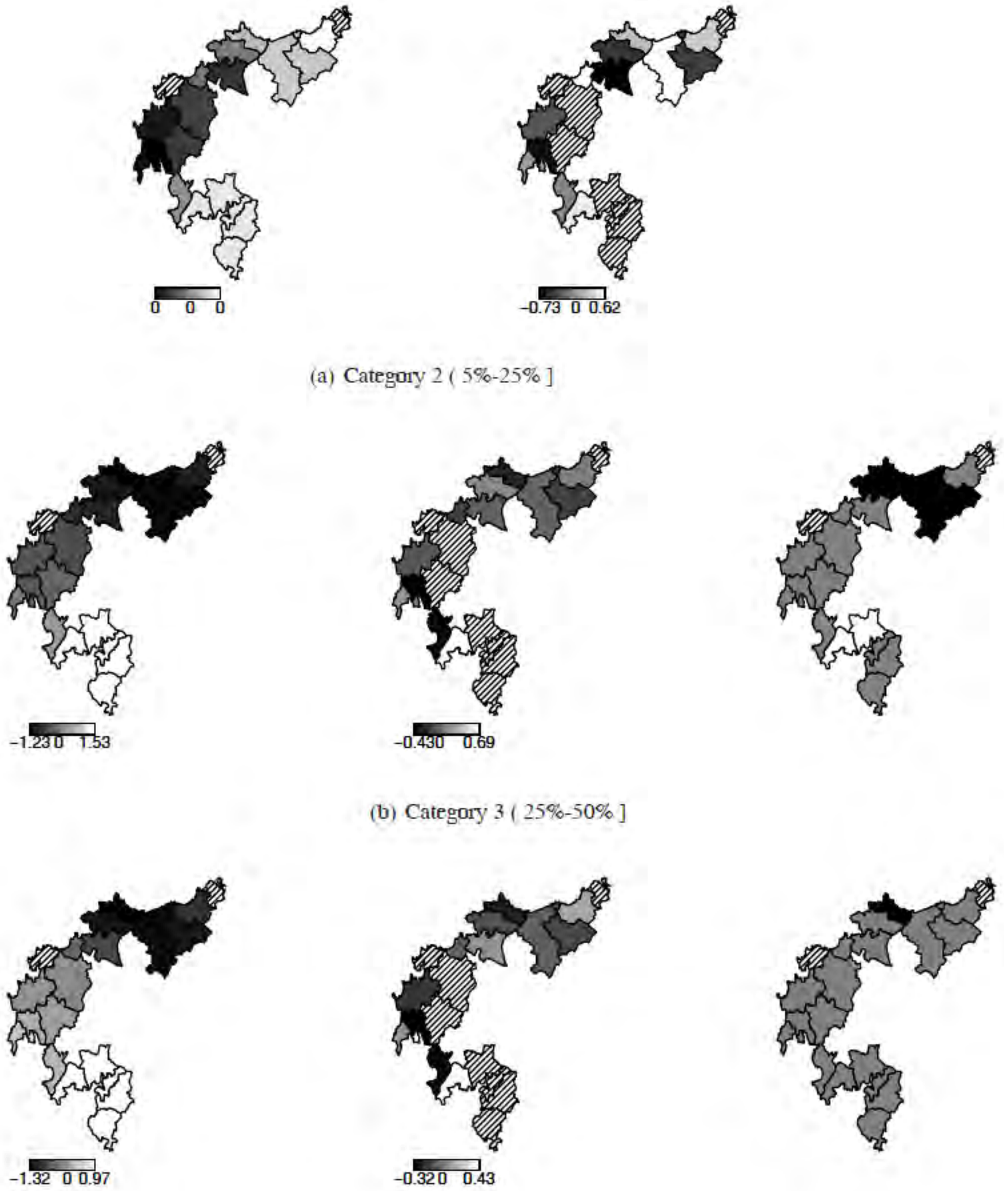
As can be seen from the maps (Figure 2, first row), the structured spatial effects for Categories 3 and 4 displayed a clear regional pattern but were not significant for Category 2. There is a descending south-north gradient, with southern regions appearing to be more appropriate and northern areas unsuitable for the presence of mussel seed. These results seem to be plausible because the northern areas are extremely exposed and steep, and even the nature of the substrate is less suitable for settlement and, by the same token, a higher abundance of mussel seed.

In the maps of unstructured effects (Figure 2, second row), the dashed areas denote regions in which unstructured effects are not estimated. No clear pattern in unstructured effects is displayed in these maps and, compared to the structured effects, the local effects were smaller. Moreover, the posterior probabilities (maps not shown) indicate that no region has a significant effect on response.

5. Conclusions

This study proposes a novel application of STAR models to the field of marine resources. The use of geoadditive multicategorical models for mussel seed data demonstrates that these models can be very useful tools for fitting this type of biological data. STAR models enable flexible non-linear effects of covariates as well spatial effects to be incorporated. Moreover, since spatial effects can be split into spatially correlated and uncorrelated parts, it becomes possible to distinguish between unobserved covariates that display a strong spatial structure and those that are only present locally. Our data revealed marked, downward, south-north spatial pattern. However, since the unstructured effects were not significant, the distribution of the mussel seed would not seem to be affected by locally present covariates.

As pointed out in Section 2 above, though the response variable was treated as nominal (with multinomial distribution) in this study, this outcome could also be considered ordinal, in which case other STAR models, such as cumulative probit/logit models, could be used in our application. Future extensions of our work include a statistical comparison of categorical versus cumulative models for fitting mussel seed distribution.



(a) Category 2 (5%-25%]

(b) Category 3 (25%-50%]

(c) Category 4 (>50%)

Figure 2: *From left to right, averages estimates of the structured spatial effects (first column), unstructured spatial effects (second column) and posterior probabilities (third column). Category 1 (<5%) is taken as reference category. For category 2 the map of posterior probabilities is not shown since the structured spatial effects proved non-significant.*

Finally, an additional advantage of using STAR models for fitting ecological data lies in the flexibility of incorporating temporal effects in a simple manner, something that makes it possible to offer flexible spatio-temporal models, which are of great interest in many biomedical fields.

Acknowledgements

The authors would like to express their gratitude for the support received in the form of the Spanish MEC Grant MTM2008-01603 and the Galician Regional Authority (Xunta de Galicia) projects INCITE08PXIB208113PR and 07MMA001200PR. We are also grateful to the referee for her/his valuable comments and suggestions, which served to make a substancial improvement to this paper.

References

Austin, M. P. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, 157, 101-118.

Austin, M. P. (2007). Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, Review, 200, 1-19.

Belitz, C., Brezger, A., Kneib, T. and Lang, S. (2009): BayesX - Software for Bayesian inference in structured additive regression models. Version 2.0.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1-59.

Boeck, P. and Wilson, M. eds. (2006). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer, New York.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing using B-splines and penalties (with comments and rejoinder). *Statistical Science*, 11, 89-121.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive models based on Markov random field priors. *Applied Statistics*, 50 (2), 201-220.

Fahrmeir, L., Kneib, T. and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, 14, 731-761.

Guisan, A., Edwards, T. C. and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, 157, 89-100.

Kammann, E. E. and Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society C*, 52, 1-18.

Kneib, T. (2006). Mixed model based inference in structured additive regression. PhD thesis, Dr.Hut-Verlag.

Kneib, T. and Fahrmeir, L. (2006). Structured additive regression for categorical space-time data: a mixed model approach. *Biometrics*, 62, 109-118.

Kneib, T., Müller, J. and Hothorn, T. (2008). Spatial smoothing techniques for the assessment of habitat suitability. *Environmental and Ecological Statistics*, 15, 343-364.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13, 183-212.

McCullagh, P., Nelder, J. A. (1997). *Generalized Linear Models*, second ed. Chapman and Hall, London.

Underwood, A. J., Chapman, M. G. and Connell, S. D. (2000). Observation s in ecology: you can't make progress on processes without understanding the patterns. *Journal of Experimental Marine Biology and Ecology*, 250, 97-115.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. CRC Press, Boca Raton, FL.

SORT 34 (1) January-June 2010, 79-94

New aging properties of the Clayton-Oakes model based on multivariate dispersion*

José Pablo Arias-Nicolás[1], Julio Mulero[2],
Olga Núñez-Barrera and Alfonso Suárez-Llorens[3]

[1] *Departamento de Matemáticas Universidad de Extremadura*

[2] *Departamento de Estadística e I.O. Universidad de Alicante*

[3] *Departamento de Estadística e I.O. Universidad de Cádiz*

Abstract

In this work we present a recent definition of Multivariate Increasing Failure Rate (MIFR) based on the concept of multivariate dispersion. This new definition is an extension of the univariate characterization of IFR distributions under dispersive ordering of the residual lifetimes. We apply this definition to the Clayton-Oakes model. In particular, we provide several conditions to order in the multivariate dispersion sense the residual lifetimes of random vectors with a dependence structure given by the Clayton-Oakes survival copula. We illustrate our results with a graphical method.

MSC: 62N05, 90B25, 62N86, 60K10 60E15

Keywords: IFR distributions, multivariate increasing failure rate, multivariate dispersion, survival copula, truncation, Clayton-Oakes model.

1. Introduction

We use the following notations throughout the paper. For every random variable or vector Z and an event A, let $[Z \mid A]$ denote a random variable or vector whose distribution is the conditional distribution of Z given A. For a random variable Z with distribution function F_Z we will denote by $\bar{F}_Z(t) = 1 - F_Z(t)$ the survival function and by $Q_Z(p) \equiv \inf\{x : F_Z(x) \geq p\}$ the quantile function. When we refer to $=_{st}$, we mean

*Corresponding author: Alfonso Suárez-Llorens Tel: (+34) 956015481, Fax: (+34) 956015378, Facultad de CC. EE. y Empresariales. C/ Duque de Nájera 8, 11002 Cádiz (Spain). Email: <alfonso.suarez@uca.es>.

Received: November 2009
Accepted: May 2010

equality in law. For every matrix $A \in M_{n\times m}$ we denote by A^t the transpose matrix. We will denote in bold all entities concerned with more than one dimension. We will assume that all multivariate distribution functions are absolutely continuous functions.

The following univariate stochastic orders are common in Stochastic Order Theory. Let X and Y be two random variables with distribution functions F and G. The random variable X is said to be smaller than Y in the univariate dispersive ordering, denoted by $X \leq_{disp} Y$, if $Q_X(q) - Q_X(p) \leq Q_Y(q) - Q_Y(p)$ for all $0 < p \leq q < 1$. In other words, if any pair of quantiles of Y are more widely separated than the corresponding of X. Let us consider now $\mathbf{X}$ and $\mathbf{Y}$ be two random vectors in $\mathbb{R}^n$. The random vector $\mathbf{X}$ is said to be smaller than $\mathbf{Y}$ in the usual stochastic ordering, denoted by $\mathbf{X} \leq_{st} \mathbf{Y}$, if $\mathbb{E}(h(\mathbf{X})) \leq \mathbb{E}(h(\mathbf{Y}))$ for any increasing function $h : \mathbb{R}^n \mapsto \mathbb{R}$ for which the expectations exist. Note that if X and Y are two random variables, $n = 1$, then $X \leq_{st} Y$ if and only if $F_X(t) \geq F_Y(t)$ for every t. Roughly speaking, $\mathbf{X} \leq_{st} \mathbf{Y}$ if $\mathbf{X}$ is less likely than $\mathbf{Y}$ to take on large values. For more details about these stochastic orders the reader may see Shaked and Shanthikumar (2007).

Univariate notions of aging constitute a well established core of reliability theory. We focus on the definition of the IFR notion. Let T be a nonnegative random variable which represents the lifetime of a unit or system. For a survival time t such that $\bar{F}_T(t) > 0$, the conditional residual lifetime distribution is given by $T_t = [T - t \mid T > t]$. Then the random variable T (or its distribution) is said to be IFR [increasing failure rate] if the survival function of the residual lifetime is decreasing when t increases that is,

$$\Pr\{T_t > h\} = \frac{\bar{F}_T(t+h)}{\bar{F}_T(t)} \text{ is decreasing in } 0 < t < \infty \text{ for all } h \geq 0. \qquad (1)$$

If the density function $f_T(t)$ exists, a straightforward computation leads to the following characterization:

$$T \text{ is IFR} \quad \Leftrightarrow \quad r_T(t) = \frac{f_T(t)}{\bar{F}_T(t)} \text{ is increasing in } t \geq 0. \qquad (2)$$

The function $r_T(t)$ given in (2) is the well known concept of failure or hazard rate, and can be interpreted as the “probability” of instant failure for a unit or a system with survival time t. Therefore the IFR notion means that the probability of instant failure or death is increasing in the survival time, see Barlow and Proschan (1975) for more details.

The IFR univariate definition has a clear interpretation and provides the basis for many useful results, which apply when dealing with the analysis of a single unit or of several units with stochastically independent lifetimes. It is worth to mention that most of the units that are alive at time t will inexorably have the IFR aging property when the time passes. Among other results, the IFR aging class can be characterized by dispersive comparisons of residual lifetimes. It holds that

$$T \text{ is IFR} \quad \Leftrightarrow \quad T_{t'} \leq_{disp} T_t, \tag{3}$$

whenever $0 \leq t \leq t'$. We can find (3) in Belzunce, Candel and Ruiz (1996) and Pellerey and Shaked (1997). A more detailed explanation for these topics can be found in Arias-Nicolás *et al.* (2009) and Belzunce and Shaked (2007). Note that expression (3) reflects the effect of the time over the dispersion of the residual lifetimes.

We also note that the definition of the DFR [Decreasing Failure Rate] aging class follows by replacing decreasing by increasing in (1), increasing by decreasing in (2) and reversing the inequality in (3).

On the other hand, multivariate IFR notions are rather controversial. In fact, starting from the univariate definition, several types of multivariate extensions can be defined. Harris (1970), Brindley and Thompson (1972), Basu (1971), Marshall (1975), Block (1977a) and (1977b), Johnson and Kotz (1975), Savits (1985), Arjas (1981) and Shaked and Shanthikumar (1991) yield different point of view which are useful in different contexts.

Arias-Nicolás *et al.* (2009) present a new concept of MIFR [Multivariate Increasing Failure Rate] based on a natural generalization of (3) via the multivariate dispersion order defined in Fernández-Ponce and Suárez-Llorens (2003), denoted by disp-MIFR. They study the main properties of this new multivariate aging concept and apply it to some well known families of multivariate distributions. They also study the relationships with the other multivariate extensions. The main purpose of this paper is the study of this new notion in the context of Clayton-Oakes model. The paper is organized as follows. In Section 2, we recall the definition of multivariate dispersion order and provide a new property which relates dispersion and copula. In Section 3, we consider the multivariate aging notion defined by Arias-Nicolás *et al.* (2009) in the context of the Clayton-Oakes model. In Section 4, we provide a graphical tool in order to clarify the exposition.

2. The multivariate dispersion order

Several attempts have been made in the literature to extend the univariate dispersion order to the multivariate case. Important contributions have been made by Oja (1983) and Giovagnoly and Wynn (1995). These authors define multivariate dispersion orders through the existence of a multivariate function k which stochastically maps a random vector $\mathbf{X}$ to another random vector $\mathbf{Y}$, i.e., $\mathbf{Y} =_{st} k(\mathbf{X})$. Shaked and Shantikumar (2007) summarize two multivariate dispersion concepts based on a particular transformation by means of the standard construction, viz., the multivariate dispersion orders defined in Shaked and Shanthikumar (1998) and Fernández-Ponce and Suárez-Llorens (2003). Recently Belzunce, Ruiz and Suárez-Llorens (2008) consider another multivariate dispersion order also based on the standard construction and study the relationship with the other definitions. We recall here the multivariate dispersion order defined in Fernández-

Ponce and Suárez-Llorens (2003). We want to emphasize that this order has desirable properties when we compare two random vectors with the same dependence structure, i.e., with the same copula.

Let $\mathbf{X}$ be a random vector and let $\mathbf{u} = (u_1, \ldots, u_n)$ in $[0,1]^n$. The standard construction for $\mathbf{X}$, denoted by

$$\hat{\mathbf{x}}(\mathbf{u}) = (\hat{x}_1(u_1), \hat{x}_2(u_1,u_2), \ldots, \hat{x}_n(u_1, \ldots, u_n)),$$

is defined as follows in terms of the univariate quantile function Q

$$\begin{aligned} &\hat{x}_1(u_1) = Q_{X_1}(u_1) \\ &\hat{x}_i(u_1, \ldots, u_i) = Q_{[X_i | \bigcap_{j=1}^{i-1} X_j = \hat{x}_j(u_1, \ldots, u_j)]}(u_i), \text{ for } i = 2, \ldots, n. \end{aligned}$$

This well known construction is widely used in simulation theory and plays the role of the quantile function in the multivariate case. It is well known that $\hat{\mathbf{x}}(\mathbf{U}) =_{st} \mathbf{X}$ where $\mathbf{U}$ is a random vector with n independent uniform components in $[0,1]$.

Let $\mathbf{X}$ and $\mathbf{Y}$ be two random vectors in $\mathbb{R}^n$. We say that $\mathbf{X}$ is less than $\mathbf{Y}$ in the multivariate dispersion order, denoted by $\mathbf{X} \leq_{disp} \mathbf{Y}$, if

$$\| \hat{\mathbf{x}}(\mathbf{v}) - \hat{\mathbf{x}}(\mathbf{u}) \|_2 \quad \leq \quad \| \hat{\mathbf{y}}(\mathbf{v}) - \hat{\mathbf{y}}(\mathbf{u}) \|_2,$$

for all $\mathbf{u}$ and $\mathbf{v}$ in $(0,1)^n$, where $\| \cdot \|_2$ means the Euclidean norm.

Fernández-Ponce and Suárez-Llorens (2003) showed that the $\leq_{disp}$ order is equivalent to verifying whether the multivariate function $\Phi = (\Phi_1, \ldots, \Phi_n)$, defined as

$$\begin{aligned} &\Phi_1(x_1) = Q_{Y_1}(F_{X_1}(x_1)) \\ &\Phi_i(x_1, \cdots, x_i) = Q_{[Y_i | \bigcap_{j=1}^{i-1} Y_j = \Phi_j(x_1, \cdots, x_j)]}(F_{[X_i | \bigcap_{j=1}^{i-1} X_j = x_j]}(x_i)), \text{ for } i = 2, \ldots, n, \end{aligned} \tag{4}$$

which satisfies that $\mathbf{Y} =_{st} \Phi(\mathbf{X})$, is an expansion function. Recall from Giovagnoly and Wynn (1995) that a function $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ is called an *expansion* if

$$\|\Phi(\mathbf{x}_2) - \Phi(\mathbf{x}_1)\|_2 \geq \|\mathbf{x}_2 - \mathbf{x}_1\|_2 \quad \text{for all } \mathbf{x}_2 \text{ and } \mathbf{x}_1 \text{ in } \mathbb{R}^n,$$

or equivalently if $J_\Phi(\mathbf{x})^t J_\Phi(\mathbf{x}) - I_n$ is non-negative for all $\mathbf{x} \in \mathbb{R}^n$, where $J_\Phi(\mathbf{x})$ and I_n denote the Jacobian and the identity matrix, respectively.

Fernández-Ponce and Suárez-Llorens (2003), Arias-Nicolás *et al.* (2005), Belzunce *et al.* (2008) and Arias-Nicolás *et al.* (2009) provide many properties of the $\leq_{disp}$ order and study the relationship with other well known multivariate dispersion concepts. For the purpose of our study, we are interested in recalling the relationship between the multivariate dispersion order and the notion of a copula.

A copula is a function that links univariate marginals to their multivariate distribution. Copulas were introduced in the context of probabilistic metric spaces, but the copula method for understanding multivariate distributions has a relatively short history in the statistics literature. In fact, most of the statistical applications have arisen in the last ten years. Given an n-dimensional distribution $\mathbf{F}(x_1,\dots,x_n)$, with marginals $F_1,\dots,F_n$, there exists an n-dimensional distribution function C, with marginals uniformly distributed over the interval $[0,1]$, such that

$$\mathbf{F}(x_1,\dots,x_n) = C(F_1(x_1),\dots,F_n(x_n)),$$

for all $(x_1,\dots,x_n)\in\mathbb{R}^n$. Moreover, this copula representation is unique if the margins are continuous. As we can see, this fact allows to separate the marginal feature and the dependence structure which is represented by the copula. For more details about the notion of copula see Nelsen (1999).

From the above, a natural question arises about comparing in dispersion two random vectors with the same dependence structure. Belzunce *et al.* (2008) provide some results concerning both copula and dispersion. The following result can be found in Arias-Nicolás *et al.* (2005). Let $\mathbf{X}=(X_1,\dots,X_n)$ and $\mathbf{Y}=(Y_1,\dots,Y_n)$ be n-dimensional random vectors with the same copula. Then

$$\mathbf{X}\leq_{disp}\mathbf{Y} \text{ if and only if } X_i\leq_{disp} Y_i, \text{ for all } i=1,\dots,n. \tag{5}$$

Note that in case of a common copula we only have to be care about the comparison of the marginal distributions. Next we provide a new result concerning the multivariate dispersion ordering that will be used later on.

Theorem 1 *Let $\mathbf{X}=(X_1,\dots,X_n)$ and $\mathbf{Y}=(Y_1,\dots,Y_n))$ be two random vector sharing the same copula. If the marginal distributions X_i and Y_i have the same finite left endpoint for $i=1,\dots,n$. If $\mathbf{X}\leq_{disp}\mathbf{Y}$ then*

$$\mathbf{X}\leq_{st}\mathbf{Y}$$

$$\left[(X_i,\dots,X_n)\,\middle|\,\bigcap_{j=1}^{i-1}X_j=\hat{x}_j(u_1,\dots,u_j)\right]\leq_{st}\left[(Y_i,\dots,Y_n)\,\middle|\,\bigcap_{j=1}^{i-1}Y_j=\hat{y}_j(u_1,\dots,u_j)\right], \tag{6}$$

for $i=2,\dots,n$ and $\mathbf{u}\in[0,1]^n$.

Proof For random vectors sharing the same copula, Arias-Nicolás *et al.* (2005) showed that the function Φ, defined in (4), can be expressed as

$$\begin{aligned}\Phi_i(x_1,\dots,x_i) &= Q_{[Y_i|\bigcap_{j=1}^{i-1}Y_j=\Phi_j(x_1,\cdots,x_j)]}(F_{[X_i|\bigcap_{j=1}^{i-1}X_j=x_j]}(x_i)),[0.2cm]\\ &= Q_{Y_i}(F_{X_i}(x_i)).\end{aligned}$$

for $i = 1, \dots, n$. On the other hand, by construction the function Φ maps the standard construction of $\mathbf{X}$ to the corresponding one of $\mathbf{Y}$, see Fernández-Ponce and Suárez-Llorens (2003), then it is clear that

$$F_{X_i}(\hat{x}_i(u_1, \dots, u_i)) = F_{Y_i}(\hat{y}_i(u_1, \dots, u_i)), \text{ for } i = 1, \dots, n. \tag{7}$$

By hypothesis assumption and using (5), $X_i \leq_{disp} Y_i$ holds for $i = 1, \dots, n$. It is well known that the univariate dispersive order implies the stochastic order when we compare distribution functions having the same left endpoint in their supports, see Shaked and Shantikumar (2007). Hence we obtain that $X_i \leq_{st} Y_i$, for $i = 1, \dots, n$, which trivially implies that $Q_{X_i}(u) \leq Q_{Y_i}(u)$ for all $u \in [0, 1]$. From the expression (7), it easily holds that $\hat{x}_i(u_1, \dots, u_i)$ and $\hat{y}_i(u_1, \dots, u_i)$ represents the same univariate quantile for the marginal distributions X_i and Y_i, respectively. Therefore

$$\hat{\mathbf{x}}_i(u_1, \dots, u_i) \leq \hat{\mathbf{y}}_i(u_1, \dots, u_i), \text{ for } i = 1, \dots, n.$$

From the mentioned fact that $\hat{\mathbf{x}}(\mathbf{U}) =_{st} \mathbf{X}$ and $\hat{\mathbf{y}}(\mathbf{U}) =_{st} \mathbf{Y}$ where $\mathbf{U}$ is a random vector with n independent uniform components in $[0, 1]$ and using Theorem 6.B.1 in Shaked and Shanthikumar (2007) we obtain that $\mathbf{X} \leq_{st} \mathbf{Y}$. Taking in account that $(\hat{x}_i(u_1, \dots, u_i), \dots, \hat{x}_n(u_1, \dots, u_n))$ represents the standard construction evaluated at $(u_i, \dots, u_n)$ for the conditional random vector

$$\left[(X_i, \dots, X_n) \middle| \bigcap_{j=1}^{i-1} X_j = \hat{x}_j(u_1, \dots, u_j) \right],$$

the rest of the proof follows directly with an equivalent argument. □

3. The Disp-MIFR notion in the context of Clayton-Oakes model

Arias-Nicolás *et al.* (2009) generalize condition (3) via the multivariate dispersion ordering. Let $\mathbf{T} = (T_1, \dots, T_n)$ be a nonnegative random vector with an absolutely continuous distribution function which represent the lifetimes of n individuals in some system. Given a vector $\mathbf{t} = (t_1, \dots, t_n)$ on $[0, \infty)^n$, the residual lifetime of $\mathbf{T}$ conditional on the observed survival data $\mathbf{t}$ is given by $\mathbf{T_t} = [\mathbf{T} - \mathbf{t} \mid \mathbf{T} > \mathbf{t}]$. Note that in the general case the residual lifetime of $\mathbf{T}$ takes into account different ages for the individuals. In this paper we restrict our study to a particular survival data, $\mathbf{t} = (t, \dots, t)$, where all individuals have the same age. The following definition can be found in Arias-Nicolás *et al.* (2009).

Definition 1 *Let $\mathbf{T} = (T_1, \dots, T_n)$ be a non-negative absolutely continuous random vector and let $\mathbf{t} = (t, \dots, t)$ and $\mathbf{t}' = (t', \dots, t')$ be two observations of survival data* *such that* $0 \leq t \leq t'$*. We will say that* $\mathbf{T}$ *is disp3-MIFR (disp3-MDFR) if*

$$\mathbf{T}'_{\mathbf{t}} = [\mathbf{T} - \mathbf{t}' \mid \mathbf{T} > \mathbf{t}'] \quad \leq_{disp} (\geq_{disp}) \quad \mathbf{T}_{\mathbf{t}} = [\mathbf{T} - \mathbf{t} \mid \mathbf{T} > \mathbf{t}]. \tag{8}$$

It is worth to mention that Arias-Nicolás *et al.* (2009) also studied the disp-MIFR and disp2-MIFR definitions for $\mathbf{t} = (t_1, \ldots, t_n)$, $\mathbf{t}' = (t'_1, \ldots, t'_n)$ and $\mathbf{t} = (t_1, \ldots, t_n)$, $\mathbf{t}' = (t_1 + t, \ldots, t_n + t)$, respectively. As we have mentioned, disp3-MIFR could be appropriated in situations where all individuals have the same age. For instance, the well known problems for twins or left-eye and right-eye.

In many types of applications, the dependence structure of a random vector $\mathbf{T}$ is given by the Clayton-Oakes survival copula:

$$\bar{C}(u_1, \ldots, u_n) = \left(\sum_{i=1}^{n} u_i^{1-\theta} - (n-1) \right)^{\frac{1}{1-\theta}}, \tag{9}$$

where, $\theta > 1$. A survival copula is a copula which yields the value of the joint survival function in terms of the values of the marginal survival functions. A multivariate distribution function $\mathbf{F}$ has the above survival copula if

$$\bar{\mathbf{F}}(x_1, \ldots, x_n) = \bar{C}(\bar{F}_1(x_1), \ldots, \bar{F}_n(x_n))$$

holds for all $\mathbf{x}$ in $\mathbb{R}^n$. Two multivariate distribution functions have the same survival copula if and only if they have the same copula, for more details see Nelsen (1999).

The family given by (9) has been widely studied in the biostatistics literature. Cook and Johnson (1981) used it to model hydro geochemical data, it is used to generalize the multivariate Pareto distribution and has been used in survival analysis, where it is generally referred to as the gamma frailty model, see Clayton (1978). In epidemiological and actuarial studies there is strong empirical evidence that supports the dependence of mortality on pairs of individuals. This type of copula is useful not only for detecting dependency but also for fitting multivariate data. In the literature we can find examples in medicine, see Sun, Wang and Sun (2006), Bogaerts and Lessafre (2008a) and (2008b) or hidrology, see De Michele *et al.* (2005) and Genest and Favre (2007).

One of the reasons why Clayton-Oakes survival copula becomes so important in biostatistics is the truncation-invariance property. If the random vector $\mathbf{T}$ has a Clayton-Oakes survival copula, then the residual lifetime $\mathbf{T}_{\mathbf{t}}$ has also the same copula. This property characterizes the Clayton-Oakes survival copula, see Sungur (1999) and (2002), Oakes (2005), Charpentier and Juri (2006) and Ahamadi Javid (2008).

From the truncation-invariance property and using (5), if $\mathbf{T}$ has a Clayton-Oakes survival copula, then $\mathbf{T}$ is disp3-MIFR (disp3-MDFR) if and only if

$$[T_i - t' \mid \mathbf{T} \geq \mathbf{t}'] \leq_{disp} (\geq_{disp}) [T_i - t \mid \mathbf{T} \geq \mathbf{t}], \tag{10}$$

for all $\mathbf{t} = (t, \ldots, t)$ and $\mathbf{t}' = (t', \ldots, t')$ such that $0 \le t \le t'$, $i = 1, \ldots, n$. This fact was pointed out in Proposition 8 in Arias-Nicolás *et al.* (2009) for the general definition disp-MIFR. The next proposition presents a result for the stochastic comparison of the residual lifetimes.

Proposition 1 *Let* $\mathbf{T} = (T_1, \ldots, T_n)$ *be a non-negative absolutely continuous random vector having a Clayton-Oakes survival copula and let* $\mathbf{t} = (t, \ldots, t)$ *and* $\mathbf{t}' = (t', \ldots, t')$ *such that* $0 \le t \le t'$. *If* $\mathbf{T}$ *is disp3-MIFR (disp-3MDFR) then*

$$\mathbf{T}_{\mathbf{t}'} \le_{st} (\ge_{st}) \mathbf{T}_{\mathbf{t}}$$

$$\left[(T_i - t', \ldots, T_n - t') \middle| \bigcap_{j=1}^{i-1} T_j = t', \bigcap_{j=i}^{n} T_j > t' \right] \le_{st} (\ge_{st}) \left[(T_i - t, \ldots, T_n - t) \middle| \bigcap_{j=1}^{i-1} T_j = t, \bigcap_{j=i}^{n} T_j > t \right],$$

for all $i = 2, \ldots, n$.

Proof From the truncation-invariance property of the Clayton-Oakes survival copula the random vectors $\mathbf{T}_{\mathbf{t}'}$ and $\mathbf{T}_{\mathbf{t}}$ share a common copula. Let us denote by $\hat{\mathbf{h}'}_{\mathbf{t}'}(\mathbf{u})$ and $\hat{\mathbf{h}}_{\mathbf{t}}(\mathbf{u})$ the standard constructions of $\mathbf{T}_{\mathbf{t}'}$ and $\mathbf{T}_{\mathbf{t}}$, respectively. The proof follows directly from Theorem 1. It is only necessary to note that $\hat{\mathbf{h}'}_{\mathbf{t}'}(0, \ldots, 0) = \hat{\mathbf{h}}_{\mathbf{t}}(0, \ldots, 0) = (0, \ldots, 0)$ for all $\mathbf{t}$ and $\mathbf{t}'$. □

Proposition 1 implies that the disp3-MIFR (disp-3MDFR) property for a random vector with a Clayton-Oakes survival copula is a sufficient condition for the aging property studied in Mulero and Pellerey (2008) based on the stochastic order of the residual lifetimes.

As a common practice in Biostatistics we will consider now a random vector $\mathbf{T}$ having an exchangeable distribution function, i.e. symmetric permutation. Some examples of this last assumption can be found in clinical trials which involve randomizing clusters or groups of subjects or units into two or more treatment arms, see Manatunga and Chen (2000).

With those settings we provide the main result of the paper. We also need a technical result before about establishing the dispersive order among members of a parametric family of univariate probability distributions.

Theorem 2 (Saunders and Moran (1978)) *Let* X_a *be a univariate random variable with distribution function* F_a *for each* $a \in \mathbb{R}$ *such that:*

1. F_a *is supported on some interval* $(X_-^{(a)}, X_+^{(a)}) \subseteq (-\infty, +\infty)$,
2. F_a *has density* f_a *which does not vanish on any subinterval of* $(X_-^{(a)}, X_+^{(a)})$, *and*
3. *the derivative of* F_a *with respect to a exists and is denoted by* $\frac{d}{da}F_a(x)$.

Then,

$$X_a \ge_{disp} X_{a^*} \text{ for } a, a^* \in \mathbb{R}, a > a^*, \tag{11}$$

if and only if

$$\frac{\frac{d}{da}F_a(x)}{f_a(x)} \textit{ is decreasing in } x. \tag{12}$$

Remark 1 Although Saunders and Moran (1978) did not mention this explicitly, it is immediate to observe, just considering the parameter $a' = 1/a$, that Theorem 2 is also valid replacing simultaneously $\leq_{disp}$ for $\geq_{disp}$ in (11) and increasing for decreasing in (12).

Theorem 3 (The main result) *Let* $\mathbf{T}$ *be a non-negative absolutely continuous random vector having a Clayton-Oakes survival copula. If* $\mathbf{T}$ *has an exchangeable distribution with margins having a common distribution* F_T*, then* $\mathbf{T}$ *is disp3-MIFR (disp3-MDFR) if and only if the function* $\phi(s)$ *defined by*

$$\phi(s) = \frac{n-(n-1)\bar{F}_T(t+s)^{\theta-1}}{r_T(t+s)} \tag{13}$$

is decreasing (increasing) in s.

Proof Without lack of generality, we will prove the result just for disp3-MIFR. Due to the fact that $\mathbf{T}$ has an exchangeable distribution with a Clayton-Oakes survival copula and using (10), $\mathbf{T}$ is disp3-MIFR if and only if

$$[T_1 - t \mid \mathbf{T} \geq \mathbf{t}] \leq_{disp} [T_1 - t' \mid \mathbf{T} \geq \mathbf{t}'], \tag{14}$$

for all $\mathbf{t} = (t,\ldots,t)$ and $\mathbf{t}' = (t',\ldots,t')$ such that $0 \leq t \leq t'$.

Note that the first component of the residual lifetime $\mathbf{T_t}$, denoted by $[\mathbf{T_t}]_1 \equiv [T_1 - t \mid \mathbf{T} \geq \mathbf{t}]$, can be considered a parametric class of univariate probability distributions depending on parameter t. Then using Theorem 2 and Remark 1, it is clear that inequality (14) holds if and only if the function

$$\frac{\frac{d}{dt}F_{[\mathbf{T_t}]_1}(s)}{f_{[\mathbf{T_t}]_1}(s)} \tag{15}$$

is increasing in s.

The conditional distribution $[\mathbf{T_t}]_1$ is given by the expression

$$F_{[\mathbf{T_t}]_1}(s) = 1 - \frac{\bar{\mathbf{F}}_\mathbf{T}(t+s,t,\ldots,t)}{\mathbf{F_T}(t,\ldots,t)},$$

where

$$\bar{\mathbf{F}}_\mathbf{T}(t_1,t_2,\ldots,t_n) = \left(\bar{F}_T(t_1)^{1-\theta} + \bar{F}_T(t_2)^{1-\theta} + \ldots + \bar{F}_T(t_n)^{1-\theta} - (n-1)\right)^{\frac{1}{1-\theta}}.$$

Therefore, a straightforward computation leads to

$$f_{[\mathbf{T_t}]_1}(s)=\frac{d}{ds}F_{[\mathbf{T_t}]_1}(s)=\left(\frac{\bar{\mathbf{F}}_{\mathbf{T}}(t+s,t,\ldots,t)}{\bar{F}_T(t+s)}\right)^{\theta}\frac{f_T(t+s)}{\bar{\mathbf{F}}_{\mathbf{T}}(t,\ldots,t)}.$$

Now if we take the partial derivative of $F_{[\mathbf{T_t}]_1}(s)$ with respect to the parameter t we obtain

$$\begin{aligned}\frac{d}{dt}F_{[\mathbf{T_t}]_1}(s) &= f_{[\mathbf{T_t}]_1}(s)+(n-1)\left(\frac{\bar{\mathbf{F}}_{\mathbf{T}}(t+s,t,\ldots,t)}{\bar{F}_T(t)}\right)^{\theta}\frac{f_T(t)}{\bar{\mathbf{F}}_{\mathbf{T}}(t,\ldots,t)}\\ &\quad -n\left(\frac{\bar{\mathbf{F}}_{\mathbf{T}}(t,\ldots,t)}{\bar{F}_T(t)}\right)^{\theta}\frac{f_T(t)\bar{\mathbf{F}}_{\mathbf{T}}(t+s,t,\ldots,t)}{\bar{\mathbf{F}}_{\mathbf{T}}(t,\ldots,t)^2}.\end{aligned}$$

Hence we have to study the expression

$$\begin{aligned}\frac{\frac{d}{dt}F_{[\mathbf{T_t}]_1}(s)}{f_{[\mathbf{T_t}]_1}(s)} &= 1+\left((n-1)-n\left(\frac{\bar{\mathbf{F}}_{\mathbf{T}}(t+s,t,\ldots,t)}{\bar{\mathbf{F}}_{\mathbf{T}}(t,\ldots,t)}\right)^{1-\theta}\right)\left(\frac{\bar{F}_T(t+s)}{\bar{F}_T(t)}\right)^{\theta}\frac{f_T(t)}{f_T(t+s)}\\ &= 1-\left(\frac{\bar{\mathbf{F}}_{\mathbf{T}}(t+s,\ldots,t+s)}{\bar{\mathbf{F}}_{\mathbf{T}}(t,\ldots,t)}\right)^{1-\theta}\left(\frac{\bar{F}_T(t+s)}{\bar{F}_T(t)}\right)^{\theta}\frac{f_T(t)}{f_T(t+s)}.\end{aligned}$$

Therefore it is clear that (15) is increasing in s, if and only if the function

$$\frac{\bar{\mathbf{F}}_{\mathbf{T}}(t+s,\ldots,t+s)^{1-\theta}\bar{F}_T(t+s)^{\theta}}{f_T(t+s)}=\frac{n-(n-1)\bar{F}_T(t+s)^{\theta-1}}{f_T(t+s)/\bar{F}_T(t+s)}$$

is decreasing in s. □

Bassan and Spizzichino (2005) pointed out the importance of studying the relations among univariate aging, multivariate aging and dependence structure for multivariate lifetimes. Note that Theorem 3 relates the new concept of MIFR-disp aging with the survival function and the hazard rate function of the margins for a particular dependence structure. We emphasize in those relations in the following results.

Corollary 1 *Let* **T** *be a non-negative absolutely continuous random vector having a Clayton-Oakes survival copula. If* **T** *has an exchangeable distribution with margins having a common distribution F_T and a non-increasing hazard rate function, then* **T** *is disp3-MDFR.*

Proof From Theorem 3 we only have to prove that the function $\phi(s)$ given by the expression (13) is increasing in s. From the hypothesis assumption for margins, the function $r_T(t+s)$ is non-increasing in s. The proof is immediate just noting that $n-(n-1)\bar{F}(t+s)^{\theta-1}$ is always increasing in s. □

Corollary 2 *Let* $\mathbf{T}$ *be a non-negative absolutely continuous random vector having a Clayton-Oakes survival copula with* $\theta \leq \frac{n}{n-1}$. *If* $\mathbf{T}$ *has an exchangeable distribution with margins having a common convex distribution* F_T, *then* $\mathbf{T}$ *is disp3-MIFR.*

Proof From Theorem 3 we only have to prove that the function $\phi(s)$ given by the expression (13) is decreasing in s. If we take the logarithm of $\phi(s)$ we obtain

$$\log(\phi(s)) = \log(n-(n-1)\bar{F}_T(t+s)^{\theta-1}) + \log \bar{F}_T(t+s) - \log f_T(t+s).$$

If F_T is convex it is clear that $-\log f_T(t+s)$ is decreasing. Now, if we take the derivative of the first and second term of $\log(\phi(s))$ with respect to s, we obtain that

$$\begin{aligned}
\frac{d}{ds}\left(\log\left(n-(n-1)\bar{F}(t+s)^{\theta-1}\right)+\log \bar{F}(t+s)\right) &\leq 0 \Longleftrightarrow \\
f_T(t+s)\frac{\theta(n-1)\bar{F}_T(t+s)^{\theta-1}-n}{(n-(n-1)\bar{F}_T(t+s)^{\theta-1})\bar{F}_T(t+s)} &\leq 0 \Longleftrightarrow \\
\theta(n-1)\bar{F}_T(t+s)^{\theta-1}-n &\leq 0 \Longleftrightarrow \\
(n-1)(\theta-1)\bar{F}(t+s,\ldots,t+s)^{\theta-1} &\leq 1. \qquad (16)
\end{aligned}$$

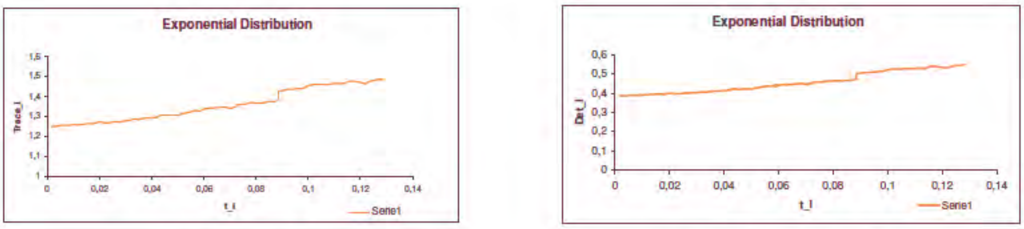
The proof concludes just observing that $\theta \leq \frac{n}{n-1}$ is a sufficient condition for inequality (16). □

4. A graphical example

In this section we only provide a graphical tool which can help to evaluate the disp3-MIFR (disp3-DMFR) notion from a practical point of view. Arias-Nicolás *et al.* (2009) pointed out the $\leq_{disp}$ order preserves many classical multivariate dispersion measures given in the literature. In particular, if $\mathbf{X} \leq_{disp} \mathbf{Y}$ then $\mathrm{trace}[\mathrm{Cov}(\mathbf{X}))] \leq \mathrm{trace}[\mathrm{Cov}(\mathbf{Y}))]$ and $\det[\mathrm{Cov}(\mathbf{X}))] \leq \det[\mathrm{Cov}(\mathbf{Y}))]$, where Cov($\mathbf{X}$) means the variance-covariance matrix of $\mathbf{X}$ and the same for $\mathbf{Y}$. Both dispersion measures based on the trace and the determinant of the variance-covariance matrix are well known in the literature and easy to estimate. The first one is known as the Total Variance, and the second one as the Wilk's Generalized Variance. Based on these properties authors provide a graphical tool to evaluate the dispersion of the multivariate residual lifetimes. Let $\mathbf{T} = (T_1,\cdots,T_n)$ be a random vector and let t be a real number in $[0,\infty)$. We denote by f_1 and f_2 the following real functions:

$$\begin{aligned}
f_1 : [0,\infty) \mapsto [0,\infty),\ f_1(t) &= \mathrm{trace}\left(\mathrm{Cov}\left(\left[T_1-t,\ldots,T_n-t|\bigcap_{i=1}^{n} T_i > t\right]\right)\right) \\
f_2 : [0,\infty) \mapsto [0,\infty),\ f_2(t) &= \det\left(\mathrm{Cov}\left(\left[T_1-t,\ldots,T_n-t|\bigcap_{i=1}^{n} T_i > t\right]\right)\right)
\end{aligned}$$

From the above discussion, it is clear that if $\mathbf{T}$ is disp3-MIFR (disp3-DMFR), then the functions f_1 and f_2 are decreasing (increasing) when times increases. In practice, the functions f_1 and f_2 can be easily estimated from the well-known non-parametric estimator of the variance-covariance matrix based on the empirical distribution. From a practical point of view we can use the non-parametric estimation of these functions to detect aging properties. We illustrate this method with two simulated examples in the bivariate case.



(a) Estimation of f_1

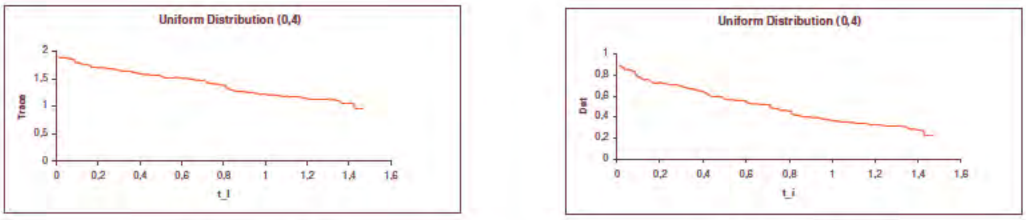
(b) Estimation of f_2

Figure 1: For $\mathbf{T} = (T_1, T_2)$ *where* $T_1 =_{st} T_2 =_{st} \text{Exp}(0,5)$

Let $\mathbf{T} = (T_1, T_2)$ be a bivariate random vector having a Clayton-Oakes survival copula. Let us consider an i.i.d. sample of size n of $\mathbf{T}$ denoted by (t_{1j}, t_{2j}), $j = 1, \ldots, n$, where simulation is done using the well known algorithm proposed by Marshall and Olkin (1988), i.e. we first generate a bivariate sample (u_{1j}, u_{2j}), $j = 1, \ldots, n$, from the Clayton-Oakes copula for a particular parameter θ and secondly we consider $t_{1j} = F_{T_1}^{-1}(u_{1j})$ and $t_{2j} = F_{T_2}^{-1}(u_{2j})$, $j = 1, \ldots, n$, where T_i, $i = 1, 2$, represent the marginal distributions. Observe that for exchangeable vectors we will consider identical marginal distributions, $T_1 =_{st} T_2 =_{st} T$. From the data, it is easy to observe that the sets

$$\{p_j = (a_j, \text{trace}(\hat{\text{Cov}}([T_1 - a_j, T_2 - a_j \mid T_1 > a_j, T_2 > a_j,]))), for\ \ j = 1, \ldots, m\}, \quad (17)$$

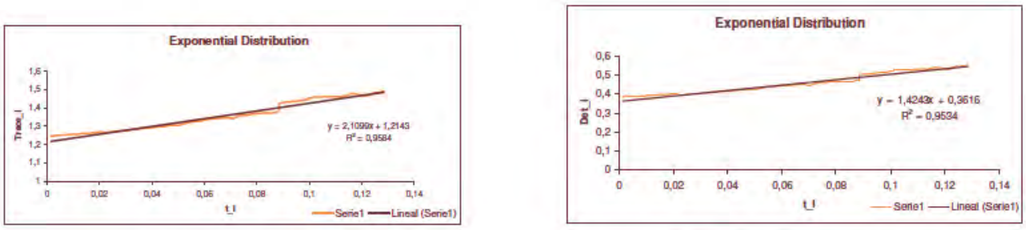
$$\{q_j = (a_j, \det(\hat{\text{Cov}}([T_1 - a_j, T_2 - a_j \mid T_1 > a_j, T_2 > a_j,]))), for\ \ j = 1, \ldots, m\} \quad (18)$$



(c) Estimation of f_1

(d) Estimation of f_2

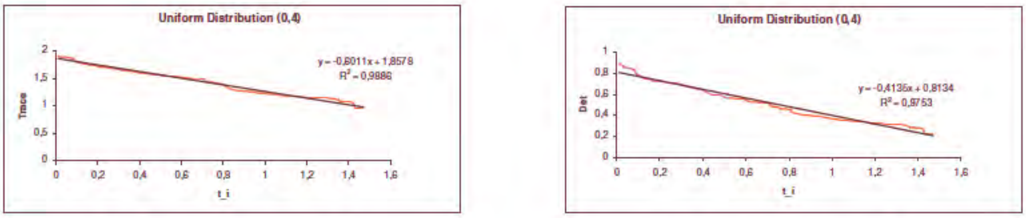
Figure 2: For $\mathbf{T} = (T_1, T_2)$ *where* $T_1 =_{st} T_2 =_{st} U(0,4)$



(e) Testing of tendency of f_1

(f) Testing of tendency of f_2

Figure 3: *Fitting a regression model for* $\mathbf{T} = (T_1, T_2)$ *where* $T_1 =_{st} T_2 =_{st} \mathrm{Exp}(0,5)$



(g) Testing of tendency of f_1

(h) Testing of tendency of f_2

Figure 4: *Fitting a regression model for* $\mathbf{T} = (T_1, T_2)$ *where* $T_1 =_{st} T_2 =_{st} U(0,4)$

provide a non-parametric estimation of the graph of f_1 and f_2 by a family of m points, $m < n$. Note that $\hat{\mathrm{Cov}}$ represents the non-parametric estimator of the variance-covariance matrix based on the empirical distribution and a_j, $j = 1, \ldots m$ are univariate sample values included in the support of T.

From the expression (17) and (18), we have simulated two estimations based on exponential and uniform marginal distributions for $n = 100$. Figure 1 represents the non-parametric estimation of f_1 and f_2 for a bivariate vector $\mathbf{T}$ having a Clayton-Oakes survival copula with parameter $\theta = 1.7$ and identical components that are exponentially distributed with mean 2 and analogously for Figure 2 but considering $\theta = 1.5$ and identical components that are uniformly distributed on $(0,4)$. Observe that we have considered $m = 60$ to guarantee a good estimation of the variance-covariance matrix. It is well known that the exponential distribution has a constant hazard rate. Hence, using Corollary 1, it easily holds that the vector $\mathbf{T} = (T_1, T_2)$ with exponential margins is disp3-MDFR. On the other hand, the uniform distribution has a convex distribution function. Hence using Corollary 2, where $\theta = 1.5 < 2$, it is clear that the vector $\mathbf{T} = (T_1, T_2)$, having a Clayton-Oakes survival copula with parameter $\theta = 1.5$, with uniform margins is disp3-MIFR.

From a practical point view, these graphs are not difficult to compute and interpret. We can graphically see the dispersion of the residual lifetimes. If the graph decreases (increases) when the time increases, we could expect a behaviour less (more) dispersive when the time increases, which is closely related to the increase (decrease) of the capacity for predicting an imminent failure. To finalize we can also fit a classical regression model to evaluate the significance of the tendency of f_1 and f_2. Figures 3 and 4 show the result of fitting some classical regression models. Note that the R^2 coefficient is larger than 0.95 in all cases.

Acknowledgements

We thank the reviewers for the careful reading of the paper and for their comments which enhanced the paper. A. Suárez-Llorens and J. P. Arias-Nicolás research has been partially supported by Ministerio de Ciencia e Innovación, grant MTM2009-08326 and grant TIN2008-06796-C04-03 respectively.

References

Ahamadi Javid, A. (2008). Copulas with truncation-invariance property. *Communications in Statistics, Theory and Methods.* DOI:10.1080/03610920802133301

Arias-Nicolás, J. P., Belzunce, F., Núñez-Barrera, O. and Suárez-Llorens, A. (2009). A multivariate IFR notion based on the multivariate dispersive ordering. *Applied Stochastic Models in Business and Industry*, 25, 339-358.

Arias-Nicolás, J. P., Fernández-Ponce, J. M., Luque-Calvo, P. and Suárez-Llorens, A. (2005). The multivariate dispersion order and the notion of copula applied to the multivariate t-distribution. *Probability in the Engineering and Informational Science*, 19, 363-375.

Arjas, E. (1981). The failure and hazard processes in multivariate reliabiblity system. *Mathematics of Operations Research*, 6, 551-562.

Barlow, R. E. and Proschan F. (1975). *Statistical Theory of Reliability and Life Testing: Probability Models*. Holt, Rinehart and Winston, New York.

Basu, A. P. (1971). Bivariate failure rate. *Journal of the American Statistical Association*, 66, 103-104.

Bassan, B. and Spizzichino, F. (2005). Relations among univariate aging, bivariate aging and dependence for exchangeable lifetimes. *Journal of Multivariate Analysis*, 93, 313-339.

Belzunce, F., Candel, J. and Ruiz, J. M. (1996). Dispersive ordering and characterizations of aging classes. *Statistics & Probability Letters*, 28, 321-327.

Belzunce, F., Ruiz, J. M. and Suárez-LLorens, A. (2008). On multivariate dispersion orderings based on the standard construction. *Statistics & Probability Letters*, 78, 271-281.

Belzunce, F. and Shaked, M. (2007). Stochastic orders and aging notions. In *Encyclopedia of Statistics in Quality and Reliability*, edited by F. Ruggeri, F. Faltin and R. Kenett, Wiley, London, 1931-1935.

Bogaerts, K. and Lesaffre, E. (2008a). Estimating local and global measures of association for bivariate interval censored data with a smooth estimate of the density. *Statistics in Medicine*, 27, 5941-5955.

Bogaerts, K. and Lesaffre, E. (2008b). Modelling the association of bivariate interval-censored data using the copula approach. *Statistics in Medicine*, 27, 6379-6392.

Block, H. W. (1977a). Monotone failure rates for multivariate distributions. *Naval Research Logistic Quarterly*, 24, 627-637.

Block, H. W. (1977b). Multivariate reliabiblity classes. *Applications in Statistics*. P. R. Krinshnaiah, edited by North Holland, 79-88.

Brindley, E. C. and Thompson, W. A., Jr. (1972). Dependence and aging aspects of multivariate survival. *Journal of the American Statistical Association*, 67, 822-830.

Charpentier, A. and Juri, A. (2006). Limiting dependence structures for tail events, with applications to credit derivatives. *Journal of Applied Probability*, 43, 563-586.

Clayton, D. G. (1978). A model for association in bivariate life tables and its applications in epidemiological studies of familiar tendency in chronic disease incidence. *Biometrika*, 65, 141-151.

Cook, R. D. and Johnson, M. E. (1981). A family of distributions for modelling non-elliptically symmetric multivariate data. *Journal of the Royal Statistical Society, Series B*, 43, 210-219.

De Michele, C., Salvadori, G., Canossi, M., Petaccia, A. and Rosso, R. (2005). Bivariate statistical approach to check adequacy of dam spillway. *Journal of Hydrologic Engineering ASCE*, 10, 50-57.

Fernández-Ponce, J. M. and Suárez-Llorens, A. (2003). A multivariate dispersion order based on quantiles more widely separated. *Journal of Multivariate Analysis*, 85, 40-53.

Genest, C. and Favre, A. C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12, 347-368.

Giovagnoli, A. and Wynn, H. P. (1995). Multivariate dispersion orderings. *Statistics & Probability Letters*, 22, 325-332.

Harris, R. (1970). A multivariate definition for increasing hazard rate distributions functions. *Annals of Mathematical Statistics*, 37, 713-717.

Johnson, N. L. and Kotz, S. (1975). A vector multivariate hazard rate. *Journal of Multivariate Analysis*, 5, 53-66.

Manatunga, A. K. and Chen, S. (2000). Sample size estimation for survival outcomes in cluster-randomized studies with small cluster sizes. *Biometrics*, 56, 616-626.

Marshall, A. W. (1975). Multivariate distributions with monotone hazard rate. *Reliability and Fault Tree Analysis. SIAM Philadelphia*, 259-284.

Marshall, A. W. and Olkin, I. (1988). Families of multivariate distributions. *Journal of the American Statistical Association*, 83, 834-841.

Mulero J. and Pellerey, F. (2010). Bivariate aging properties under archimedean dependence structures. *Communications in Statistics: Theory and Methods.* In press.

Nelsen, R. B. (1999). An introduction to copulas. *Lectures Notes in Statistics*, 139, Springer-Verlag, New York.

Oakes, D. (2005). On the preservation of copula structure under truncation. *The Canadian Journal of Statistics*, 33, 465-468.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1, 327-332.

Pellerey, F. and Shaked, M. (1997). Characterizations of the IFR and DFR aging notions by means of the dispersive order. *Statistics & Probability Letters*, 33, 389-393.

Saunders, I. and Moran, P. (1978). On the Quantiles of the Gamma and F Distributions. *Journal of Applied Probability*, 15, 426-432.

Savits, T. H. (1985). A Multivariate IFR distributions. *Journal of Applied Probability*, 22, 197-204.

Shaked, M. and Shanthikumar, J. G. (1991). Dinamic multivariate aging notions in Reliability Theory. *Sthocastic Processes and Their Appications*, 38, 85-97.

Shaked, M. and Shanthikumar, J. G. (1998). Two variability orders. *Probability in the Engineering and Informational Sciences*, 12, 1-23.

Shaked, M. and Shanthikumar, J. G. (2007). Sthocastic Orders. *Springer Series in Statistics*.

Sun, L, Wang, L. and Sun J. (2006). Estimation of the association for bivariate interval-censored failure time data. *Scandinavian Journal of Statistics*, 33, 637-649.

Sungur, E. A. (1999). Truncation invariant dependence structures. *Communications in Statistics: Theory and Methods*, 28, 2553-2568.

Sungur, E. A. (2002). Some results on truncation dependence invariant class of copulas. *Communications in Statistics: Theory and Methods*, 31, 1399-1422.

Book review

Antedependence Models for Longitudinal Data

Dale Zimmerman and Vicente Núñez Antón

Chapman & Hall / CRC, 2010, 270 pp

Longitudinal data analysis has become an extremely important area of statistics, and this is reflected both in the diversity of the real life applications and in the depth of the methods developed for this type of data. This book is the first one dedicated entirely to antedependece models, which are a very rich family of conditional independence models for serially correlated data. Following the ideas of Diggle et al. (2002), the inference problem in longitudinal data is approached first by selecting an appropriate covariance structure (which in most applications is considered as a set of nuisance parameters) and then making inferences on the mean parameters (which generally are the ones of primary interest).

The material presented is well organized. The first chapter presents motivational examples and the concept of antedependence models. The second chapter characterizes more formally the unstructured antedependence models, and shows equivalent ways of defining them. The structured models are presented and discussed in chapter 3, together with related models for serially correlated data. Chapter 4 presents exploratory techniques to identify these models, and shows examples of both numerical and graphical diagnostic tools. Formal likelihood-based inference is presented in chapters 5-7. Chapter 5 introduces estimation techniques for different mean models, both for complete and incomplete data (including several patterns of missing data common in longitudinal data). Chapter 6 presents hypothesis tests and related tools (such as penalized likelihood criteria) for the covariance structure, and chapter 7 approaches the problem of testing hypotheses on the mean parameters. The use of most of the techniques presented is illustrated in the four data sets presented in chapter 1, and chapter 8 integrates these examples into case studies for each of the data sets. Chapter 9 presents some additional topics and extensions. Several matrix results and more technical proofs are presented as appendices.

Several relevant R functions are available for download from the first author's webpage (http://www.stat.uiowa.edu/∼dzimmer), and the authors are making available and documenting more in the near future. This is very important for applications, since there is little software developed for these models, and applied researchers will be very satisfied with the availability of R functions to fit antedependence models to their data.

The book is geared towards statisticians (both theoretical and applied) and other researchers in application areas (medicine, epidemiology, animal science, forestry, ecology) with experience in linear models, multivariate analysis, and/or longitudinal data analysis. For graduate students in statistics or biostatistics, the book is very useful as a supplementary material in a course on longitudinal data analysis, or it could be the basis of a special topics course on antedependence models.

In summary, this book is a welcome addition to the bibliography of longitudinal data, combining rigorous statistical presentation with interesting real life examples.

References

Diggle, P. J., Heagerty, P. J., Liang, K. Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. 2nd ed., New York: Oxford University Press.

Raúl E. Macchiavelli
University of Puerto Rico Mayagüez

Information for authors and subscribers

Information for authors and subscribers

Submitting articles to SORT

Guidelines for submitting articles

SORT accepts for publication only original articles that have not been submitted to any other journal in the areas of Statistics, Operations Research, Official Statistics and Biometrics. Articles may be either theoretical or applied, and may include computational or educational elements. Publication will be exclusively in English, with an abstract in Catalan (abstract translation into Catalan can be done by the journal's staff).

All articles submitted to thematic sections will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board, except for those articles specifically invited by the journal or reprinted with permission. Reviewer comments will be sent to the primary author if changes are requested in form or content.

When an original article is received, the journal sends the author a dated receipt; this date will appear in the publication as "reception date". The date in which its final version is received will be the "acceptance date".

To submit an article, the author must send it in PDF format to the electronic address of the journal (sort@idescat.cat) especifying his e-mail address and/or a phone number. Once the article has been approved, the author must send a final electronic version, following the instructions of the editor responsible for that article. It is recommended that this final version be submitted using a $\LaTeX\,2_{\varepsilon}$.

In any case, upon request the journal secretary will provide authors with $\LaTeX\,2_{\varepsilon}$ templates and appropriate references to the MSC2000 classification of the American Mathematical Society. For more information about the criteria for using this classification system, authors may consult directly the Mathematics Subject Classification 2000 (MSC2000) at the website of the Institut d'Estadística de Catalunya (http://www.idescat.es/sort/Normes.stm).

Publishing rights and authors' opinions

The publishing rights for SORT belong to the Institut d'Estadística de Catalunya since 1992 through express concession by the Technical University of Catalonia. The journal's current legal registry can be found under the reference number DL B-46.085-1977 and its ISSN is 1696-2281. Partial or total reproduction of this publication is expressly forbidden, as is any manual, mechanical, electronic or other type of storage or transmission without prior written authorisation from the Institut d'Estadística de Catalunya.

Authored articles represent the opinions of the authors; the journal does not necessarily agree with the opinions expressed in authored articles.

Formal guidelines for articles

The first page of the article must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (100-200 words) followed by MSC2000 classification of the American Mathematical Society and the keywords. Before submitting an article, the author(s) would be well advised to ensure that the text uses correct English.

Bibliographic references have to be listed alphabetically at the end of the article, according to the following examples:

Citations
Mahalanobis (1936), Rao (1982b)

Journal articles
Pfeffermann, D. and Barnard, C. H. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business & Economic Statistics*, 9, 73-84.

Books
Platek, R., Rao, J. N. K., Särndal, C. E. and Singh, M. P. (eds.) (1987). *Small Area Statistics: an International Symposium*. New York: John Wiley and Sons.

Parts of books
Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J.P. Klein and P.K. Goel (eds.), Kluwer Academic Publishers, 237-247.

Web files or "pages"
Nielsen, S. F. (2001). *Proper and improper multiple imputation*
http://www.stat.ku.dk/˜feodor/publications/ (10th May 2003).

Multiple publications by a single author are to be listed chronologically. Explanatory notes should be numbered sequentially and placed at the bottom of the corresponding page. Tables and figures should also be numbered sequentially and will be reproduced directly from the submitted originals if it is impossible to include them in the electronic text.

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel.: +34-93 412 09 24 or +34-93 412 00 88
Fax: +34-93 412 31 45
E-mail: sort@idescat.cat

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form
SORT *(Statistics and Operations Research Transactions)*

Name ____

Organisation ____

Street Address ______________________________________

Zip/Postal code ____________________ City ____________________

State/Country ____________________ Tel. ____________________

Fax ______________ NIF/VAT Registration Number ______________

E-mail ____

Date ______________________

Signature

I wish to subscribe to **SORT** ***(Statistics and Operations Research Transactions)*** for the year 2010 (volume 34)

Annual subscription rates:

– Spain: €22 (4% VAT included)
– Other countries: €25 (4% VAT included)

Price for individual issues (current and back issues):

– Spain: €15/issue (4% VAT included)
– Other countries: €17/issue (4% VAT included)

Method of payment:

□ Bank transfer to account number 2013-0100-53-0200698577

□ Automatic bank withdrawal from the following account number
□□□□ □□□□ □□ □□□□□□□□□□

□ Check made payable to the Institut d´Estadística de Catalunya

Please send this subscription form (or a photocopy) to:

SORT ***(Statistics and Operations Research Transactions)***
Institut d´Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45

Bank copy

Authorisation for automatic bank withdrawal in payment for

SORT ***(Statistics and Operations Research Transactions)***

The undersigned ____

authorises Bank/Financial institution ______________________________

located at (Street Address) ___________________________________

Zip/postal code ____________________ City ____________________

Country ____

to draft the subscription to **SORT** ***(Statistics and Operations Research Transactions)*** from my account

number □□□□ □□□□ □□ □□□□□□□□□□

Date ________________________

Signature

SORT ***(Statistics and Operations Research Transactions)***
Institut d´Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45