ISSN: 1696-2281
SORT 30 (2) July-December (2006)

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

SORT

Volume 30 (2), July-December 2006

Formerly Qüestiió

Contents

FOREWORD

In July 2006, a new editorial board took up the task of promoting the journal SORT, selecting manuscripts and evaluating them by the peer-review methodology commonly used in scientific publications. The new editorial team is dedicated to maintaining the high standards of SORT and to make the journal even more widely recognized among scholars working in Statistics and Operations Research.

On behalf of the editorial board, I wish to extend sincere thanks to the past editor in chief, Prof. Carles M. Cuadras, for his wonderful and enormous job during many years, and for his continual enthusiasm for the journal. He devoted a great deal of time and energy to the journal, and we believe that future generations of statisticians in Catalonia and the general readers of SORT will continue to benefit from his outstanding work.

One of our immediate goals is to make sure that SORT is abstracted in the citation reports that are increasingly relevant to the scientific community. We believe that authors would appreciate the journal even more when they know that it tries to distribute their articles to the largest number of readers possible.

I would like to introduce to you the objectives and goals of the new editorial team and some information about the type of manuscripts we would like to see submitted for publication. Given that the audience of SORT is very broad, the new editorial team is especially encouraging the submission of manuscripts that can show innovation in the practical use of Statistics and Operations Research, in fields such as biology, medicine, psychology, economics and marketing, amongst others. Survey articles showing new techniques and their usefulness will also be considered. We feel that SORT has contributed significantly to the dissemination of information about progress in our field, but we also believe that it can offer much more to our community. We welcome your input and any comments you wish to send to us should be directed to sort@idescat.net.

We acknowledge all authors, associated editors and referees who have been involved in making SORT a respected journal in the past and who wish to do so also in the future. Finally, we also want to express our indebtedness to the Statistical Institute of Catalonia (Idescat) for their ongoing support for the journal since 1992, first publishing it as QÜESTIIÓ and now as SORT.

Let me wish you enjoyable reading!

Montserrat Guillén
Editor in chief

SORT 30 (2) July-December 2006, 125-146

What does intrinsic mean in statistical estimation?*

Gloria García[1] and Josep M. Oller[2]

[1] *Pompeu Fabra University, Barcelona, Spain* [2] *University of Barcelona, Barcelona, Spain.*

Abstract

In this paper we review different meanings of the word *intrinsic* in statistical estimation, focusing our attention on the use of this word in the analysis of the properties of an estimator. We review the intrinsic versions of the bias and the mean square error and results analogous to the Cramér-Rao inequality and Rao-Blackwell theorem. Different results related to the Bernoulli and normal distributions are also considered.

MSC: 62F10, 62B10, 62A99.

Keywords: Intrinsic bias, mean square Rao distance, information metric.

1 Introduction

Statistical estimation is concerned with specifying, in a certain framework, a plausible probabilistic mechanism which explains observed data. The inherent nature of this problem is inductive, although the process of estimation itself is derived through mathematical deductive reasoning.

In parametric statistical estimation the probability is assumed to belong to a class indexed by some parameter. Thus the inductive inferences are usually in the form of point or region estimates of the probabilistic mechanism which has generated some specific data. As these estimates are provided through the estimation of the parameter, a label of the probability, different estimators may lead to different methods of induction.

*This research is partially sponsored by CGYCIT, PB96-1004-C02-01 and 1997SGR-00183 (Generalitat de Catalunya), Spain.

Address for correspondence: J. M. Oller, Departament d'Estadística, Universitat de Barcelona, Diagonal 645, 08028-Barcelona, Spain. **e-mail:** joller@ub.edu

Received: April 2006

Under this approach an estimator should not depend on the specified parametrization of the model: this property is known as the *functional invariance* of an estimator. At this point, the notion of intrinsic estimation is raised for the first time: an estimator is *intrinsic* if it satisfies this functional invariance property, and in this way is a real probability measure estimator. On the other hand, the bias and the mean square error (MSE) are the most commonly accepted measures of the performance of an estimator. Nevertheless these concepts are clearly dependent on the model parametrization and thus unbiasedness and uniformly minimum variance estimation are *non-intrinsic*.

It is also convenient to examine the goodness of an estimator through *intrinsic* conceptual tools: this is the object of the *intrinsic analysis of statistical estimation* introduced by Oller & Corcuera (1995) (see also Oller (1993b) and Oller (1993a)). These papers consider an intrinsic measure for the bias and the square error taking into account that a parametric statistical model with suitable regularity conditions has a natural Riemannian structure given by the information metric. In this setting, the square error loss is replaced by the square of the corresponding Riemannian distance, known as the *information distance* or the *Rao distance*, and the bias is redefined through a convenient vector field based on the geometrical properties of the model. It must be pointed out that there exist other possible intrinsic losses but the square of the Rao distance is the most natural intrinsic version of the square error.

In a recent paper of Bernardo & Juárez (2003), the author introduces the concept of intrinsic estimation by considering the estimator which minimizes the Bayesian risk, taking as a loss function a symmetrized version of Kullback-Leibler divergence (Bernardo & Rueda (2002)) and considering a reference prior based on an information-theoretic approach (Bernardo (1979) and Berger & Bernardo (1992)) which is independent of the model parametrization and in some cases coincides with the Jeffreys uniform prior distribution. In the latter case the prior, usually improper, is proportional to the Riemannian volume corresponding to the information metric (Jeffreys (1946)). This estimator is intrinsic as it does not depend on the parametrization of the model.

Moreover, observe that both the loss function and the reference prior are derived just from the model and this gives rise to another notion of intrinsic: an estimation procedure is said to be *intrinsic* if it is formalized only in terms of the model. Observe that in the framework of information geometry, a concept is *intrinsic* as far as it has a well-defined geometrical meaning.

In the present paper we review the basic results of the above-mentioned intrinsic analysis of the statistical estimation. We also examine, for some concrete examples, the intrinsic estimator obtained by minimizing the Bayesian risk using as an intrinsic loss the square of the Rao distance and as a reference prior the Jeffrey's uniform prior. In each case the corresponding estimator is compared with the one obtained by Bernardo & Juárez (2003).

2 The intrinsic analysis

As we pointed out before, the bias and mean square error are not intrinsic concepts. The aim of the *intrinsic analysis* of the *statistical estimation*, is to provide intrinsic tools for the analysis of intrinsic estimators, developing in this way a theory analogous to the classical one, based on some natural geometrical structures of the statistical models. In particular, intrinsic versions of the Cramér–Rao lower bound and the Rao–Blackwell theorem have been established.

We first introduce some notation. Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space and Θ be a connected open set of $\mathbb{R}^n$. Consider a map $f : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}$ such that $f(x,\theta) \geq 0$ and $f(x,\theta)\mu(dx)$ defines a probability measure on $(\mathcal{X}, \mathcal{A})$ to be denoted as P_θ. In the present paper a *parametric statistical model* is defined as the triple $\{(\mathcal{X}, \mathcal{A}, \mu)\,;\, \Theta\,;\, f\}$. We will refer to μ as the *reference measure of the model* and to Θ as the *parameter space*.

In a general framework Θ can be any manifold modelled in a convenient space as $\mathbb{R}^n$, $\mathbb{C}^n$, or any Banach or Hilbert space. So even though the following results can be written with more generality, for the sake of simplicity we consider the above-mentioned form for the parameter space Θ. In that case, it is customary to use the same symbol (θ) to denote points and coordinates.

Assume that the parametric statistical model is identifiable, i.e. there exists a one-to-one map between parameters θ and probabilities P_θ; assume also that f satisfies the regularity conditions to guarantee that the Fisher information matrix exists and is a strictly positive definite matrix. In that case Θ has a natural Riemannian manifold structure induced by its information metric and the parametric statistical model is said to be *regular*. For further details, see Atkinson & Mitchel (1981), Burbea (1986), Burbea & Rao (1982) and Rao (1945), among others.

As we are assuming that the model is identifiable, an *estimator $\mathcal{U}$ of the true probability measure* based on a k-size random sample, $k \in \mathbb{N}$, may be defined as a measurable map from $\mathcal{X}^k$ to the manifold Θ, which induces a probability measure on Θ known as the *image measure* and denoted as ν_k. Observe that we are viewing Θ as a manifold, not as an open set of $\mathbb{R}^n$.

To define the bias in an intrinsic way, we need the notion of mean or expected value for a random object valued on the manifold Θ. One way to achieve this purpose is through an affine connection on the manifold. Note that Θ is equipped with Levi–Civita connection, corresponding to the Riemannian structure supplied by the information metric.

Next we review the exponential map definition. Fix θ in Θ and let $T_\theta\Theta$ be the tangent space at θ. Given $\xi \in T_\theta\Theta$, consider a geodesic curve $\gamma_\xi : [0,1] \to \Theta$, starting at θ and satisfying $\frac{d\gamma_\xi}{dt}\big|_{t=0} = \xi$. Such a curve exists as far as ξ belongs to an open star-shaped neighbourhood of $0 \in T_\theta\Theta$. In that case, the exponential map is defined as $\exp_\theta(\xi) = \gamma_\xi(1)$. Hereafter, we restrict our attention to the Riemannian case, denoting by $\|\cdot\|_\theta$ the

norm at $T_\theta\Theta$ and by ρ the Riemannian distance. We define

$$\Im_\theta = \{\xi \in T_\theta\Theta : \|\xi\|_\theta = 1\} \subset T_\theta\Theta$$

and for each $\xi \in \Im_\theta$ we define

$$c_\theta(\xi) = \sup\{t > 0 : \rho\,(\theta, \gamma_\xi(t)) = t\}\,.$$

If we set

$$\mathfrak{D}_\theta = \{t\xi \in T_\theta\Theta : 0 \le t < c_\theta(\xi)\ ;\ \xi \in \Im_\theta\} \quad \text{and} \quad D_\theta = \exp_\theta(\mathfrak{D}_\theta)\,,$$

it is well known that $\exp_\theta$ maps $\mathfrak{D}_\theta$ diffeomorphically onto D_θ. Moreover, if the manifold is complete the boundary of $\mathfrak{D}_\theta$ is mapped by the exponential map onto the boundary of D_θ, called the *cut locus of* θ *in* Θ. For further details see Chavel (1993).

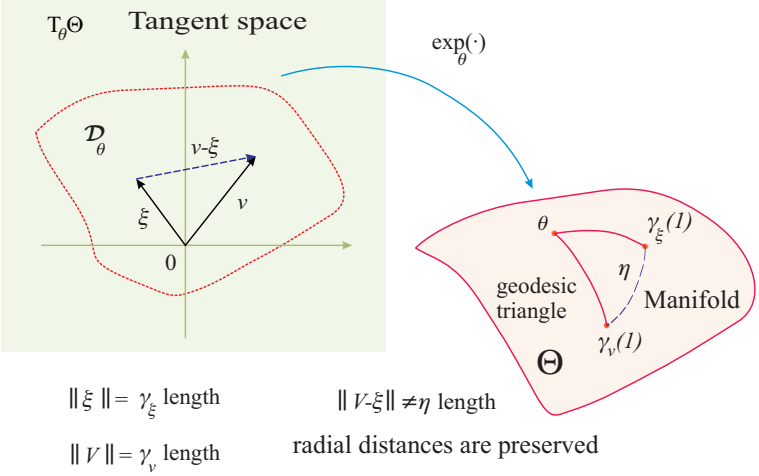

Figure 1: *The exponential map*

For the sake of simplicity, we shall assume that $\nu_k(\Theta \setminus D_\theta) = 0$, whatever true probability measure in the statistical model is considered. In this case, the inverse of the exponential map, $\exp_\theta^{-1}$, is defined ν_k–almost everywhere. For additional details see Chavel (1993), Hicks (1965) or Spivak (1979).

For a fixed sample size k, we define the *estimator vector field* A as

$$A_\theta(x) = \exp_\theta^{-1}(\mathcal{U}(x))\,, \quad \theta \in \Theta\,.$$

which is a C^∞ random vector field (first order contravariant tensor field) induced on the manifold through the inverse of the exponential map.

For a point $\theta \in \Theta$ we denote by E_θ the expectation computed with respect to the probability distribution corresponding to θ. We say that θ is a *mean value* of $\mathcal{U}$ if and

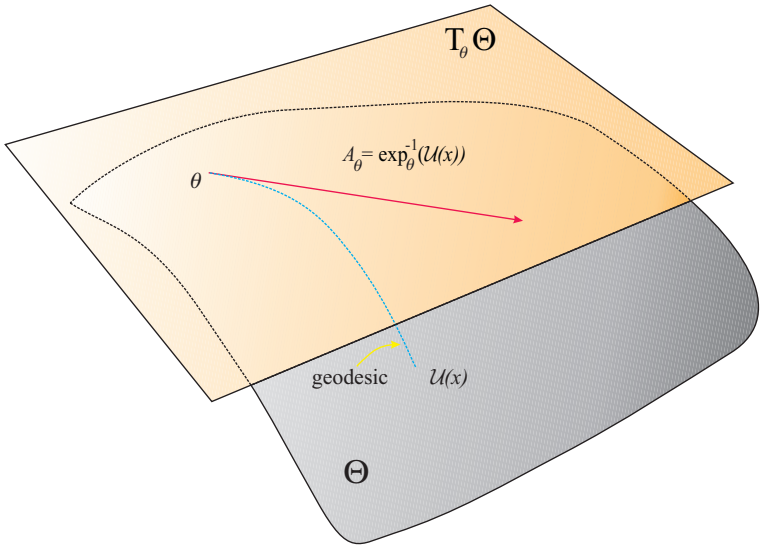

Figure 2: *Estimator vector field*

only if $E_\theta(A_\theta) = 0$. It must be pointed out that if a *Riemannian centre of mass* exists, it satisfies the above condition (see Karcher (1977) and Oller & Corcuera (1995)).

We say that an estimator $\mathcal{U}$ is *intrinsically unbiased* if and only if its mean value is the true parameter. A tensorial measure of the bias is the *bias vector field B*, defined as

$$B_\theta = E_\theta(A_\theta)\,, \quad \theta \in \Theta\,.$$

An *invariant bias measure* is given by the scalar field $\|B\|^2$ defined as

$$\|B_\theta\|^2_\theta\,, \quad \theta \in \Theta\,.$$

Notice that if $\|B\|^2 = 0$, the estimator is intrinsically unbiased.

The estimator vector field A also induces an intrinsic measure analogous to the mean square error. The *Riemannian risk of* $\mathcal{U}$, is the scalar field defined as

$$E_\theta\left(\|A_\theta\|^2_\theta\right) = E_\theta\left(\rho^2(\mathcal{U},\theta)\right)\,, \quad \theta \in \Theta\,.$$

since $\|A(x)\|^2_\theta = \rho^2(\mathcal{U}(x),\theta)$. Notice that in the Euclidean setting the Riemannian risk coincides with the mean square error using an appropriate coordinate system.

Finally note that if a mean value exists and is unique, it is natural to regard the expected value of the square of the Riemannian distance, also known as the *Rao distance*, between the estimated points and their mean value as an intrinsic version of the variance of the estimator.

To finish this section, it is convenient to note the importance of the selection of a loss function in a statistical problem. Let us consider the estimation of the probability of success $\theta \in (0,1)$ in a binary experiment where we perform independent trials until the first success. The corresponding density of the number of is given by

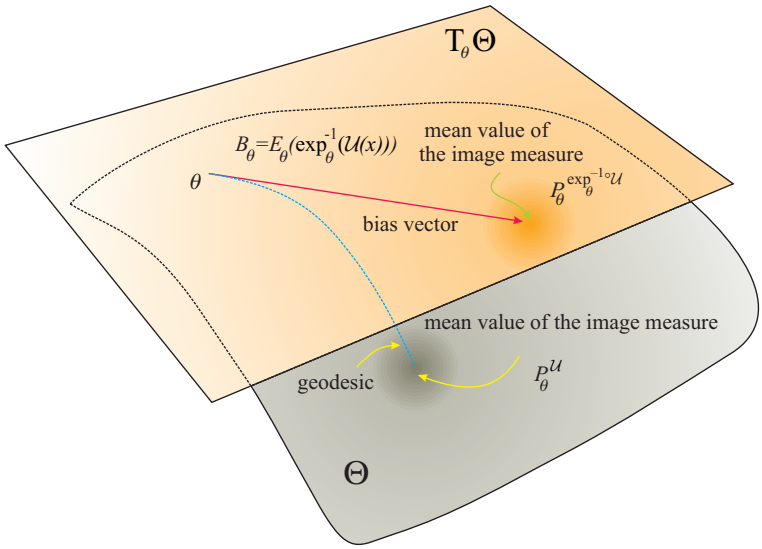

Figure 3: *Bias vector field*

$$f(k;\theta) = (1-\theta)^k\,\theta\,; \quad k = 0, 1, \ldots$$

If we restrict our attention to the class of unbiased estimators, a (classical) unbiased estimator U of θ, must satisfy

$$\sum_{k=0}^{\infty} U(k)\,(1-\theta)^k\,\theta = \theta, \quad \forall\theta \in (0,1),$$

where it follows that $\sum_{k=0}^{\infty} U(k)\,(1-\theta)^k$ is constant for all $\theta \in (0,1)$. So $U(0) = 1$ and $U(k) = 0$ for $k \geq 1$. In other words: when the first trial is a success, U assigns θ equal to 1; otherwise θ is taken to be 0.

Observe that, strictly speaking, there is no (classical) unbiased estimator for θ since U takes values in the boundary of the parameter space $(0,1)$. But we can still use the estimator U in a wider setting, extending both the sample space and the parameter space. We can then compare U with the maximum likelihood estimator, $V(k) = 1/(k+1)$ for $k \geq 0$, in terms of the mean square error. After some straightforward calculations, we obtain

$$\begin{aligned} E_\theta\left((U-\theta)^2\right) &= \theta - \theta^2 \\ E_\theta\left((V-\theta)^2\right) &= \theta^2 + \left(\theta\,Li_2(1-\theta) + 2\theta^2\ln(\theta)\right)/(1-\theta) \end{aligned}$$

where Li_2 is the dilogarithm function. Further details on this function can be found in Abramovitz (1970), page 1004. The next figure represents both mean square error of U and V.

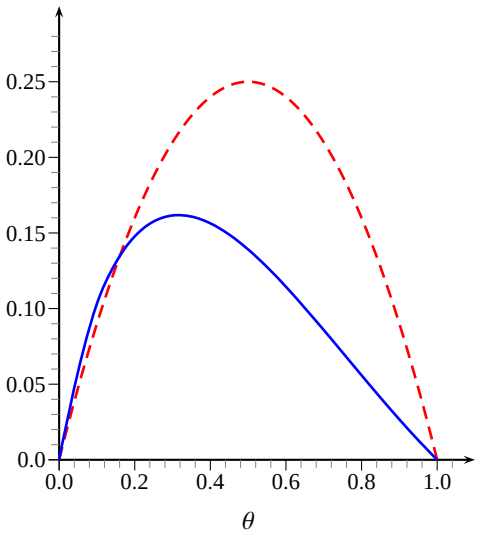

Figure 4: *MSE of U (dashed line) and V (solid line).*

It follows that there exist points in the parameter space for which the estimator U is preferable to V since U scores less risk; precisely for $\theta \in (0, 0.1606)$ where the upper extreme has been evaluated numerically. This admissibility contradicts the common sense that refuses U: this estimator assigns θ to be 0 even when the success occurs in a finite number of trials. This points out the fact that the MSE criterion is not enough to distinguish properly between estimators.

Instead of using the MSE we may compute the Riemannian risk for U and V. In the geometric model, the Rao distance ρ is given by

$$\rho(\theta_1, \theta_2) = 2 \left| \operatorname{arg\,tanh}\left(\sqrt{1-\theta_1}\right) - \operatorname{arg\,tanh}\left(\sqrt{1-\theta_2}\right) \right|, \ \theta_1, \theta_2 \in (0,1)$$

which tends to $+\infty$ when θ_1 or θ_2 tend to 0. So $E_\theta\left(\rho^2(U,\theta)\right) = +\infty$ meanwhile $E_\theta\left(\rho^2(V,\theta)\right) < +\infty$. The comparison in terms of Riemannian risk discards the estimator U in favour of the maximum likelihood estimator V, as is reasonable to expect.

Furthermore we can observe that the estimator U, which is classically unbiased, has infinite norm of the bias vector. So U is not even intrinsically unbiased, in contrast to V which has finite bias vector norm.

3 Intrinsic version of classical results

In this section we outline a relationship between the unbiasedness and the Riemannian risk obtaining an intrinsic version of the Cramér–Rao lower bound. These results are obtained through the comparison theorems of Riemannian geometry, see Chavel (1993)

and Oller & Corcuera (1995). Other authors have also worked in this direction, such as Hendricks (1991), where random objects on an arbitrary manifold are considered, obtaining a version for the Cramér–Rao inequality in the case of unbiased estimators. Recent developments on this subject can be found in Smith (2005).

Hereafter we consider the framework described in the previous section. Let $\mathcal{U}$ be an estimator corresponding to the regular model $\{(\chi, \mathcal{A}, \mu)\,;\,\Theta\,;\,f\}$, where the parameter space Θ is a n–dimensional real manifold and assume that for all $\theta \in \Theta$, $\nu_k(\Theta \setminus D_\theta) = 0$.

Theorem 3.1. *[Intrinsic Cramér–Rao lower bound] Let us assume that* $E\left(\rho^2(\mathcal{U},\theta)\right)$ *exists and the covariant derivative of* $E(A)$ *exists and can be obtained by differentiating under the integral sign. Then,*

1. *We have*
$$E\left(\rho^2(\mathcal{U},\theta)\right) \geq \frac{\left(\operatorname{div}(B) - E(\operatorname{div}(A))\right)^2}{kn} + \|B\|^2\,,$$
where $\operatorname{div}(\cdot)$ *stands for the divergence operator.*
2. *If all the sectional Riemannian curvatures K are bounded from above by a non-positive constant* $\mathcal{K}$ *and* $\operatorname{div}(B) \geq -n$*, then*
$$E\left(\rho^2(\mathcal{U},\theta)\right) \geq \frac{\left(\operatorname{div}(B) + 1 + (n-1)\sqrt{-\mathcal{K}}\,\|B\|\coth\left(\sqrt{-\mathcal{K}}\,\|B\|\right)\right)^2}{kn} + \|B\|^2\,.$$
3. *If all sectional Riemannian curvatures K are bounded from above by a positive constant* $\mathcal{K}$ *and* $d(\Theta) < \pi/2\sqrt{\mathcal{K}}$*, where* $d(\Theta)$ *is the diameter of the manifold, and* $\operatorname{div}(B) \geq -1$*, then*
$$E\left(\rho^2(\mathcal{U},\theta)\right) \geq \frac{\left(\operatorname{div}(B) + 1 + (n-1)\sqrt{\mathcal{K}}\,d(\Theta)\cot\left(\sqrt{\mathcal{K}}\,d(\Theta)\right)\right)^2}{kn} + \|B\|^2\,.$$

In particular, for intrinsically unbiased estimators, we have:

4. *If all sectional Riemannian curvatures are non-positive, then*
$$E\left(\rho^2(\mathcal{U},\theta)\right) \geq \frac{n}{k}$$
5. *If all sectional curvatures are less or equal than a positive constant* $\mathcal{K}$ *and* $d(\Theta) < \pi/2\sqrt{\mathcal{K}}$*, then*
$$E\left(\rho^2(\mathcal{U},\theta)\right) \geq \frac{1}{kn}$$

The last result shows up the effect of the Riemannian sectional curvature on the precision which can be attained by an estimator.

Observe also that any one–dimensional manifold corresponding to one–parameter family of probability distributions is always Euclidean and $\operatorname{div}(B) = -1$; thus part 2 of

Theorem (3.1) applies. There are also some well known families of probability distributions which satisfy the assumptions of this last theorem, such as the multinomial, see Atkinson & Mitchel (1981), the negative multinomial distribution, see Oller & Cuadras (1985), or the extreme value distributions, see Oller (1987), among many others.

It is easy to check that in the n-variate normal case with known covariance matrix Σ, where the Rao distance is the Mahalanobis distance, the sample mean based on a sample of size k is an estimator that attains the intrinsic Cramér–Rao lower bound, since

$$\begin{aligned} E\left(\rho^2(\overline{X},\mu)\right) &= E\left((\overline{X}-\mu)^{\mathsf{T}}\Sigma^{-1}(\overline{X}-\mu)\right) = \\ &= E\left(\mathrm{tr}(\Sigma^{-1}(\overline{X}-\mu)(\overline{X}-\mu)^{\mathsf{T}})\right) = \\ &= \mathrm{tr}\left(\Sigma^{-1}E\left((\overline{X}-\mu)(\overline{X}-\mu)^{\mathsf{T}}\right)\right) = \mathrm{tr}(\frac{1}{k}\,I) = \frac{n}{k} \end{aligned}$$

where v^{T} is the transpose of a vector v.

Next we consider a tensorial version of the Cramér-Rao inequality. First we define the *dispersion tensor* corresponding to an estimator $\mathcal{U}$ as:

$$S_\theta = E_\theta(A_\theta \otimes A_\theta) \qquad \forall\theta \in \Theta$$

Theorem 3.2. *The dispersion tensor S satisfies*

$$S \geq \frac{1}{k}\mathrm{Tr}^{\,2,4}[\mathrm{G}^{\,2,2}[(\nabla B - E(\nabla A)) \otimes (\nabla B - E(\nabla A))]] + B \otimes B$$

where $\mathrm{Tr}^{\,i,j}$ *and* $\mathrm{G}^{\,i,j}$ *are, respectively, the contraction and raising operators on index i, j and ∇ is the covariant derivative. Here the inequality denotes that the difference between the right and the left hand side is non-negative definite.*

Now we study how we can decrease the mean square Rao distance of a given estimator. Classically this is achieved by taking the conditional mean value with respect to a sufficient statistic; we shall follow a similar procedure here. But now our random objects are valued on a manifold: we need to define the conditional mean value concept in this case and then obtain an intrinsic version of the Rao–Blackwell theorem.

Let $(\mathcal{X}, \mathcal{A}, P)$ be a probability space. Let M be a n–dimensional, complete and connected Riemannian manifold. Then M is a complete separable metric space (a Polish space) and we will have a regular version of the conditional probability of any M–valued random object f with respect to any σ-algebra $\mathcal{D} \subset \mathcal{A}$ on $\mathcal{X}$. In the case where the mean square of the Riemannian distance ρ of f exists, we can define

$$E(\rho^2(f,m)|\mathcal{D})(x) = \int_M \rho^2(t,m) P_{f|\mathcal{D}}(x,dt)\,,$$

where $x \in \mathcal{X}$, B is a Borelian set in M and $P_{f|\mathcal{D}}(x,B)$ is a regular conditional probability of f given $\mathcal{D}$.

If for each $x \in \mathcal{X}$ there exists a unique mean value $p \in M$ corresponding to the conditional probability $P_{f|\mathcal{D}}(x, B)$, i.e. a point $p \in M$ such that

$$\int_M \exp_p^{-1}(t) P_{f|\mathcal{D}}(x, dt) = 0_p,$$

we have a map from $\mathcal{X}$ to M that assigns, to each x, the mean value corresponding to $P_{f|\mathcal{D}}(x, B)$.

Therefore, if f is a random object on M and $\mathcal{D} \subset \mathcal{A}$ a σ–algebra on $\mathcal{X}$, we can define the conditional mean value of f with respect $\mathcal{D}$, denoted by $\mathfrak{M}(f|\mathcal{D})$, as a $\mathcal{D}$-measurable map, Z, such that

$$E(\exp_Z^{-1}(f(\cdot))|\mathcal{D}) = 0_Z$$

provided it exists. A sufficient condition to assure that the mean value exists and is uniquely defined, is the existence of an open geodesically convex subset $N \subset M$ such that $P\{f \in N\} = 1$. Finally, it is necessary to mention that $\mathfrak{M}(\mathfrak{M}(f|\mathcal{D})) \neq \mathfrak{M}(f)$, see for instance Kendall (1990).

Let us apply these notions to statistical point estimation. Given the regular parametric statistical model $\{(\mathcal{X}, \mathcal{A}, \mu); \Theta; f\}$, we assume that Θ is complete or that there exist a metric space isometry with a subset of a complete and connected Riemannian manifold. We recall now that a real valued function h on a manifold, equipped with an affine connection, is said to be *convex* if for any geodesic γ, $h \circ \gamma$ is a convex function. Then we have the following result.

Theorem 3.3. (Intrinsic Rao–Blackwell) *Let $\mathcal{D}$ be a sufficient σ–algebra for the statistical model. Consider an estimator $\mathcal{U}$ such that $\mathfrak{M}(\mathcal{U}|\mathcal{D})$ is well defined.*

If θ is such that $\rho^2(\theta, \cdot)$ is convex then

$$E_\theta\left(\rho^2(\mathfrak{M}(\mathcal{U}|\mathcal{D}), \theta)\right) \leq E_\theta(\rho^2(\mathcal{U}, \theta)) \ .$$

The proof is based on Kendall (1990). Sufficient conditions for the hypothesis of the previous theorem are given in the following result

Theorem 3.4. *If the sectional curvatures of N are at most 0, or $\mathcal{K} > 0$ with $d(N) < \pi/2\sqrt{\mathcal{K}}$, where $d(N)$ is the diameter of N, then $\rho^2(\theta, \cdot)$ is convex $\forall \theta \in \Theta$.*

It is not necessarily true that the mean of the square of the Riemannian distance between the true and estimated densities decreases when conditioning on $\mathcal{D}$. For instance, if some of the curvatures are positive and we do not have further information about the diameter of the manifold, we cannot be sure about the convexity of the square of the Riemannian distance.

On the other hand, the efficiency of the estimators can be improved by conditioning with respect to a sufficient σ-algebra $\mathcal{D}$ obtaining $\mathfrak{M}(\mathcal{U}|\mathcal{D})$. But in general the bias is

not preserved, in contrast to the classical Rao-Blackwell theorem; in other words, even if $\mathcal{U}$ were intrinsically unbiased, $\mathfrak{M}(\mathcal{U}|\mathcal{D})$ would not be in general intrinsically unbiased since,

$$\mathfrak{M}(\mathfrak{M}(\mathcal{U}|\mathcal{D})) \neq \mathfrak{M}(\mathcal{U}).$$

However the norm of the bias tensor of $\mathfrak{M}(\mathcal{U}|\mathcal{D})$ is bounded: if we let $B^{\mathfrak{M}(\mathcal{U}|\mathcal{D})}$ be the bias tensor, by the Jensen inequality,

$$||B_\theta^{\mathfrak{M}(\mathcal{U}|\mathcal{D})}||_\theta^2 \leq E_\theta\left(\rho^2(\mathfrak{M}(\mathcal{U}|\mathcal{D}), \theta)\right) \leq E_\theta(\rho^2(\mathcal{U}, \theta)).$$

4 Examples

This section is devoted to examine the goodness of some estimators for several models. Different principles apply in order to select a convenient estimator; here we consider the estimator that minimizes the Riemannian risk for a prior distribution proportional to the Riemannian volume. This approach is related to the ideas developed by Bernardo & Juárez (2003), where the authors consider as a loss function a symmetrized version of the Kullback-Leibler divergence instead of the square of the Rao distance and use a reference prior which, in some cases, coincides with the Riemannian volume. Once that estimator is obtained, we examine its intrinsic performance: we compute the corresponding Riemannian risk and its bias vector, precisely the square norm of the intrinsic bias. We also compare this estimator with the maximum likelihood estimator.

4.1 Bernoulli

Let $X_1, \ldots, X_k$ be a random sample of size k from a Bernoulli distribution with parameter θ, that is with probability density $f(x\,;\theta) = \theta^x(1-\theta)^{1-x}$, for $x \in \{0, 1\}$. In that case, the parameter space is $\Theta = (0, 1)$ and the metric tensor is given by

$$g(\theta) = \frac{1}{\theta(1-\theta)}$$

We assume the prior distribution π for θ be the Jeffreys prior, that is

$$\pi(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}}$$

The corresponding joint density of θ and $(X_1, \ldots, X_k)$ is then proportional to

$$\frac{1}{\sqrt{\theta(1-\theta)}}\ \theta^{\sum_{i=1}^k X_i}\ (1-\theta)^{k-\sum_{i=1}^k X_i} = \theta^{\sum_{i=1}^k X_i - \frac{1}{2}}\ (1-\theta)^{k-\sum_{i=1}^k X_i - \frac{1}{2}}$$

which depends on the sample through the sufficient statistic $T = \sum_{i=1}^k X_i$. When $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$ put $T = t$. since,

$$\int_0^1 \theta^{\,t-\frac{1}{2}}(1-\theta)^{\,k-t-\frac{1}{2}}\,d\theta = \mathrm{Beta}\Big(t+\frac{1}{2},\ k-t+\frac{1}{2}\Big)$$

the posterior distribution $\pi(\cdot \mid t)$ based on the Jeffreys prior is as follows

$$\pi(\theta \mid t) = \frac{1}{\mathrm{Beta}\left(t+\frac{1}{2},\ k-t+\frac{1}{2}\right)}\,\theta^{t-\frac{1}{2}}\ (1-\theta)^{k-t-\frac{1}{2}}$$

where Beta is the Euler beta function.

The Bayes estimator related to the loss function given by the square of the Rao distance ρ^2 is

$$\theta^b(s) = \arg\min_{\theta^e\in(0,1)} \int_0^1 \rho^2(\theta^e,\theta)\ \ \pi(\theta \mid t)\,d\theta$$

Since an intrinsic estimation procedure is invariant under reparametrization, we perform the change of coordinates defined through the equation

$$1 = \left(\frac{d\theta}{d\xi}\right)^2 \frac{1}{\xi(1-\xi)}$$

in order to obtain a metric tensor equal to 1: the Riemannian distance expressed via this coordinate system, known as *Cartesian coordinate system*, will coincide with the Euclidean distance between the new coordinates. If we solve this differential equation, with the initial conditions equal to $\xi(0) = 0$, we obtain $\xi = 2\arcsin(\sqrt{\theta})$ and $\xi = -2\arcsin(\sqrt{\theta})$; we only consider the first of these two solutions. After some straightforward computations we obtain

$$\rho\,(\theta_1,\theta_2) = 2\arccos\Big(\sqrt{\theta_1\ \theta_1} + \sqrt{(1-\theta_1)\ (1-\theta_2)}\ \Big) = |\xi_1 - \xi_2| \tag{1}$$

for $\xi_1 = 2\arcsin(\sqrt{\theta_1})$ and $\xi_2 = 2\arcsin(\sqrt{\theta_2})$ and $\theta_1, \theta_2 \in \Theta$.

In the Cartesian setting, the Bayes estimator $\xi^b(s)$ is equal to the expected value of ξ with respect to the posterior distribution

$$\pi(\xi \mid t) = \frac{1}{\mathrm{Beta}\left(t+\frac{1}{2},\ k-t+\frac{1}{2}\right)}\left(\sin^2\left(\frac{\xi}{2}\right)\right)^t\ \left(1-\sin^2\left(\frac{\xi}{2}\right)\right)^{k-t}$$

Once we apply the change of coordinates $\theta = \sin^2\left(\frac{\xi}{2}\right)$, the estimator $\xi^b(s)$ is

$$\xi^b(t) = \frac{1}{\mathrm{Beta}\left(t+\frac{1}{2},\ k-t+\frac{1}{2}\right)}\int_0^1 2\ \arcsin\Big(\sqrt{\theta}\ \Big)\ \theta^{\,t-\frac{1}{2}}\,(1-\theta)^{\,k-t-\frac{1}{2}}\,d\theta$$

Expanding $\arcsin(\sqrt{\theta}\,)$ in power series of θ,

$$\arcsin(\sqrt{\theta}) = \frac{1}{\sqrt{\pi}} \sum_{j=0}^{\infty} \frac{\Gamma\left(j+\frac{1}{2}\right)}{j!\,(2j+1)}\,\theta^{\,j+\frac{1}{2}}$$

where Γ is the Euler gamma function. After some computations, we obtain

$$\xi^b(t) = 2\,\frac{\Gamma(k+1)\;\Gamma(t+1)}{\Gamma\left(k+\frac{3}{2}\right)\;\Gamma\left(t+\frac{1}{2}\right)}\;{}_3F_2\left(\frac{1}{2},\frac{1}{2},t+1;k+\frac{3}{2},\frac{3}{2};1\right) \tag{2}$$

where ${}_3F_2$ denotes a generalized hypergeometric function. Further details on the gamma, beta and hypergeometric functions can be found on Erdélyi et al. (1955). Finally the Bayes estimator $\theta^b(t)$ of θ is given by

$$\theta^b(t) = \sin^2\left(\frac{\Gamma(k+1)\;\Gamma(t+1)}{\Gamma\left(k+\frac{3}{2}\right)\;\Gamma\left(t+\frac{1}{2}\right)}\;{}_3F_2\left(\frac{1}{2},\frac{1}{2},t+1;k+\frac{3}{2},\frac{3}{2};1\right)\right)$$

It is straightforward to prove that

$$\theta^b(k-t) = 1-\theta^b(t)$$

and can be approximated by

$$\theta^a(t) = \frac{t}{k} + \left(\frac{1}{2}-\frac{t}{k}\right)\left(\frac{0.63}{k}-\frac{0.23}{k^2}\right)$$

with relative errors less that 3.5% for any result based on sample size $k \le 100$.

The behaviour of these estimators, for different values of k and for small t, is shown in the following table.

	$\theta^b(0)$	$\theta^b(1)$	$\theta^b(2)$	$\theta^a(0)$	$\theta^a(1)$	$\theta^a(2)$
$k=1$	0.20276	0.79724	-	0.20000	0.80000	-
$k=2$	0.12475	0.50000	0.87525	0.12875	0.50000	0.87125
$k=5$	0.05750	0.23055	0.40995	0.05840	0.23504	0.41168
$k=10$	0.03023	0.12109	0.21532	0.03035	0.12428	0.21821
$k=20$	0.01551	0.06207	0.11037	0.01546	0.06392	0.11237
$k=30$	0.01043	0.04173	0.07420	0.01037	0.04301	0.07566
$k=50$	0.00630	0.02521	0.04482	0.00625	0.02600	0.04575
$k=100$	0.00317	0.01267	0.02252	0.00314	0.01308	0.02301

Observe that these estimators do not estimate θ as zero when $t=0$, similarly to the estimator obtained by Bernardo & Juárez (2003), which is particularly useful when we are dealing with rare events and small sample sizes.

The Riemannian risk of this intrinsic estimator has been evaluated numerically and is represented in Figure . Note that the results are given in terms of the Cartesian coordinates ξ^b, in order to guarantee that the physical distance in the plots is proportional to the Rao distance. The Riemannian risk of θ^b is given by

$$E_\theta\left(\rho^2(\theta^b,\theta)\right) = E_\xi\left((\xi^b-\xi)^2\right) = \sum_{t=0}^{k}(\xi^b(t)-\xi)^2\binom{k}{t}\sin^{2t}\left(\frac{\xi}{2}\right)\cos^{2(k-t)}\left(\frac{\xi}{2}\right)$$

which can be numerically computed through expression (2). This can be compared with the numerical evaluation of the Riemannian risk of the maximum likelihood estimator $\theta^* = t/k$, given by

$$\begin{aligned} E_\theta\left(\rho^2(\theta^*,\theta)\right) &= E_\xi\left((\xi^*-\xi)^2\right) \\ &= \sum_{t=0}^{k}\left(2\ \arcsin\left(\sqrt{\frac{t}{k}}\right)-\xi\right)^2\binom{k}{t}\sin^{2t}\left(\frac{\xi}{2}\right)\cos^{2(k-t)}\left(\frac{\xi}{2}\right) \end{aligned}$$

as we can see in Figure 5.
We point out that the computation of the Riemannian risk for the maximum likelihood estimator requires the extension by continuity of the Rao distance given in (1) to the closure of the parameter space Θ as θ^* takes values on $[0, 1]$.

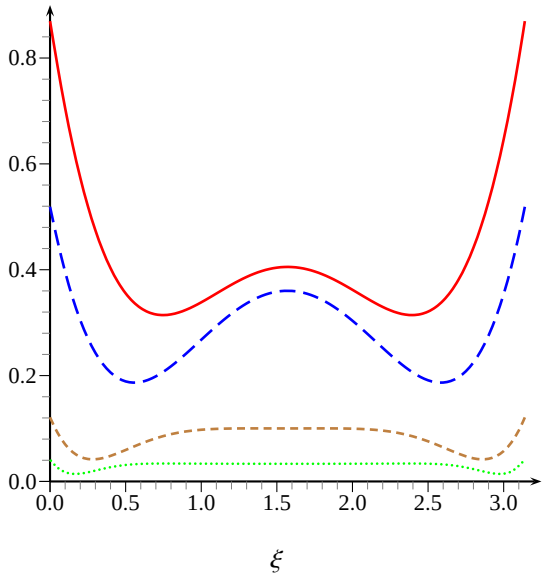

Figure 5: *Riemannian risk of ξ^b, for $k = 1$ (solid line), $k = 2$ (long dashed line), $k = 10$ (short dashed line) and $k = 30$ (dotted line).*

For a fixed sample size, observe that the Riemannian risk corresponding to ξ^b is lower than the Riemannian risk corresponding to ξ^* in a considerable portion of the parameter space, as it is clearly shown in Figure .

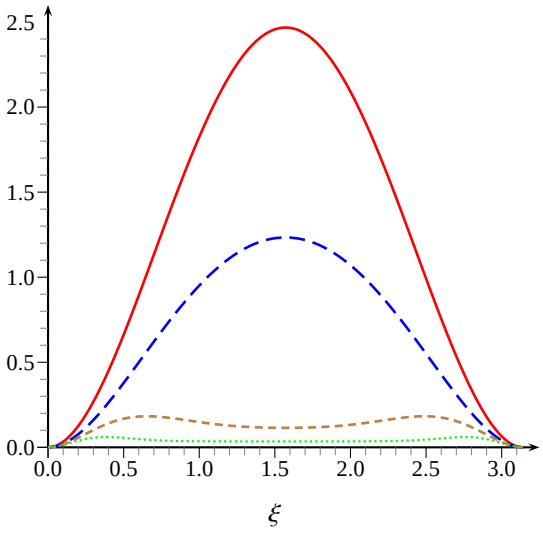

Figure 6: *Riemannian risk of ξ^*, for $k = 1$ (solid line), $k = 2$ (long dashed line), $k = 10$ (short dashed line) and $k = 30$ (dotted line).*

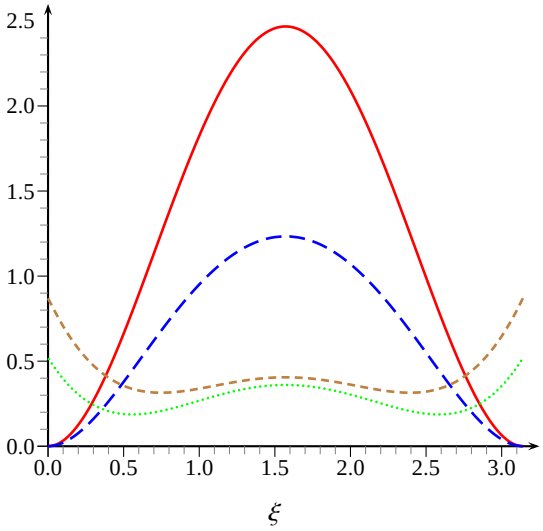

Figure 7: *Riemannian risk corresponding to θ^*, for $k = 1$ (solid line), $k = 2$ (long dashed line) and corresponding to θ^b, for $k = 1$ (short dashed line) and $k = 2$ (dotted line).*

Note that part of the Riemannian risk comes up through the bias of an estimator. Next the square of the norm of the bias vector B^b for θ^b and B^* for θ^* is evaluated numerically. Formally, in the Cartesian coordinate system ξ

$$B^b_\xi = E_\xi\left(\xi^b\right) - \xi = \sum_{t=0}^{k}(\xi^b(t) - \xi)\binom{k}{t}\sin^{2t}\left(\frac{\xi}{2}\right)\cos^{2(k-t)}\left(\frac{\xi}{2}\right)$$

$$B^*_\xi = E_\xi\left(\xi^*\right) - \xi = \sum_{t=0}^{k}\left(2\ \arcsin\left(\sqrt{\frac{t}{n}}\right) - \xi\right)\binom{k}{t}\sin^{2t}\left(\frac{\xi}{2}\right)\cos^{2(k-t)}\left(\frac{\xi}{2}\right)$$

The squared norm of the bias vector B^b and of B^* are represented in Figures and respectively.

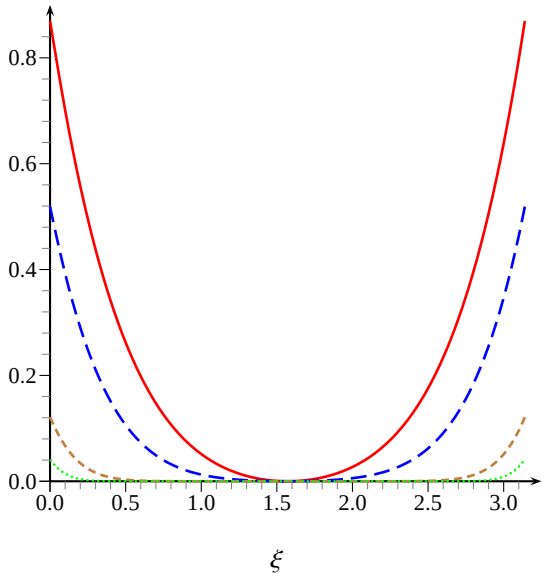

Figure 8: $\|B^b\|^2$ *for* $k = 1$ *(solid line),* $k = 2$ *(long dashed line),* $k = 10$ *(short dashed line) and* $k = 30$ *(dotted line).*

Now, when the sample size is fixed, the intrinsic bias corresponding to ξ^b is greater than the intrinsic bias corresponding to ξ^* in a wide range of values of the model parameter, that is the opposite behaviour showed up by the Riemannian risk.

4.2 Normal with mean value known

Let $X_1, \ldots, X_k$ be a random sample of size k from a normal distribution with known mean value μ_0 and standard deviation σ. Now the parameter space is $\Theta = (0, +\infty)$ and the metric tensor for the $N(\mu_0, \sigma)$ model is given by

$$g(\sigma) = \frac{2}{\sigma^2}$$

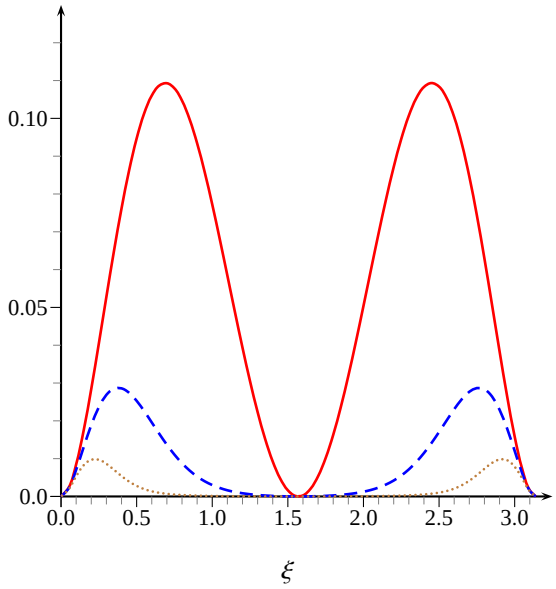

Figure 9: $\|B^*\|^2$ $k = 1, 2$ *(solid line) (the same curve),* $k = 10$ *(dashed line) and* $k = 30$ *(dotted line).*

We shall assume again the Jeffreys prior distribution for σ. Thus the joint density for σ and $(X_1, \ldots, X_k)$ is proportional to

$$\frac{1}{\sigma^{k+1}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{k}(X_i - \mu_0)^2\right)$$

depending on the sample through the sufficient statistic $S^2 = \frac{1}{k}\sum_{i=1}^{k}(X_i - \mu_0)^2$. When $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$ put $S^2 = s^2$. As

$$\int_0^\infty \frac{1}{\sigma^{k+1}} \exp\left(-\frac{k}{2\sigma^2}s^2\right) d\sigma = \frac{2^{\frac{k}{2}-1}}{(ks^2)^{\frac{k}{2}}}\,\Gamma\left(\frac{k}{2}\right)$$

the corresponding posterior distribution $\pi(\cdot \mid s^2)$ based on the Jeffreys prior satisfies

$$\pi(\sigma \mid s^2) = \frac{\left(ks^2\right)^{\frac{k}{2}}}{2^{\frac{k}{2}-1}\,\Gamma\left(\frac{k}{2}\right)} \frac{1}{\sigma^{k+1}} \exp\left(-\frac{k}{2\sigma^2}s^2\right)$$

Denote by ρ the Rao distance for the $N(\mu_0, \sigma)$. As we did in the previous example, instead of directly determining

$$\sigma^b(s) = \arg\min_{\sigma^e \in (0+\infty)} \int_0^{+\infty} \rho^2(\sigma^e, \sigma)\ \pi(\sigma \mid s^2)\, d\sigma$$

we perform a change of coordinates to obtain a Cartesian coordinate system. Then we compute the Bayes estimator for the new parameter's coordinate θ; as the estimator obtained in this way is intrinsic, we finish the argument recovering σ from θ. Formally, the

change of coordinates for which the metric tensor is constant and equal to 1 is obtained by solving the following differential equation:

$$1 = \left(\frac{d\sigma}{d\theta}\right) \frac{2}{\sigma^2}$$

with the initial conditions $\theta(1) = 0$. We obtain $\theta = \sqrt{2}\ \ln(\sigma)$ and $\theta = -\sqrt{2}\ \ln(\sigma)$; we only consider the first of these two solutions. We then obtain

$$\rho(\sigma_1, \sigma_2) = \sqrt{2}\ \left|\ln\left(\frac{\sigma_1}{\sigma_2}\right)\right| = |\theta_1 - \theta_2| \tag{3}$$

for $\theta_1 = \sqrt{2}\ln\sigma_1$ and $\theta_2 = \sqrt{2}\log\sigma_2$.

In the Cartesian setting, the Bayes estimator $\theta^b(s^2)$ for θ is the expected value of θ with respect to the corresponding posterior distribution, to be denoted as $\pi(\cdot \mid s^2)$. The integral can be solved, after performing the change of coordinates $\theta = \sqrt{2}\ \ln(\sigma)$ in terms of the gamma function Γ and the digamma function Ψ, that is the logarithmic derivative of Γ. Formally,

$$\begin{aligned}\theta^b(s^2) &= \frac{\left(ks^2\right)^{\frac{k}{2}}}{2^{\frac{n-3}{2}}\ \Gamma\left(\frac{k}{2}\right)} \int_0^{+\infty} \frac{\ln(\sigma)}{\sigma^{k+1}}\ \exp\left(-\frac{k}{2\sigma^2}s^2\right) d\sigma \\ &= \frac{\sqrt{2}}{2}\ \left(\ln\left(\frac{k}{2}\ s^2\right) - \Psi\left(\frac{k}{2}\right)\right)\end{aligned}$$

The Bayes estimator for σ is then

$$\sigma^b(s^2) = \sqrt{\frac{k}{2}}\ \exp\left(-\frac{1}{2}\Psi\left(\frac{k}{2}\right)\right)\ \sqrt{s^2}$$

Observe that this estimator, σ^b is a multiple of the maximum likelihood estimator $\sigma^* = \sqrt{s^2}$. We can evaluate the proportionality factor, for some values of n, obtaining the following table.

n	$\sqrt{\frac{k}{2}}\ \exp\left(-\frac{1}{2}\Psi\left(\frac{k}{2}\right)\right)$	n	$\sqrt{\frac{k}{2}}\ \exp\left(-\frac{1}{2}\Psi\left(\frac{k}{2}\right)\right)$
1	1.88736	10	1.05302
2	1.33457	20	1.02574
3	1.20260	30	1.01699
4	1.14474	40	1.01268
5	1.11245	50	1.01012
6	1.09189	100	1.00503
7	1.07767	250	1.00200
8	1.06725	500	1.00100
9	1.05929	1000	1.00050

The Riemannian risk of σ^b is given by

$$E_\sigma\left(\rho^2(\sigma^b,\sigma)\right) = E_\theta\left((\theta^b-\theta)^2\right) = \frac{1}{2}\Psi'(\frac{k}{2})$$

where Ψ' is the derivative of the digamma function. Observe that the Riemannian risk is constant on the parameter space.

Additionally we can compute the square of the norm of the bias vector corresponding to σ^b. In the Cartesian coordinate system θ and taking into account that $\frac{k\,S^2}{\sigma^2}$ is distributed as χ^2_k, we have

$$\begin{aligned} B^b_\theta &= E_\theta(\theta^b)-\theta = E_\sigma\left(\sqrt{2}\ \ln\left(\sqrt{\tfrac{k}{2}\,S^2}\right)-\Psi\left(\tfrac{k}{2}\right)\right)-\ \sqrt{2}\ln\sigma \\ &= \sqrt{2}\,E_\sigma\left(\ \ln\left(\sqrt{\frac{k\,S^2}{2\,\sigma^2}}\right)\right)-\Psi\left(\frac{k}{2}\right)=0 \end{aligned}$$

That is, the estimator σ^b is intrinsically unbiased. The bias vector corresponding to σ^* is given by

$$\begin{aligned} B^*_\theta &= E_\theta(\theta^*)-\theta = E_\sigma\left(\sqrt{2}\ \ln\left(\sqrt{S^2}\right)\right)-\ \sqrt{2}\ln\sigma \\ &= \sqrt{2}E_\sigma\left(\ \ln\left(\sqrt{\frac{S^2}{\sigma^2}}\right)\right)=\frac{1}{\sqrt{2}}\left(\Psi\left(\frac{k}{2}\right)\right)+\frac{1}{\sqrt{2}}\ln\left(\frac{k}{2}\right) \end{aligned}$$

which indicates that the estimator σ^* has a non-null intrinsic bias.

Furthermore, the Bayes estimator σ^b also satisfies the following interesting property related to the unbiasedness: it is the equivariant estimator under the action of the multiplicative group $\mathbb{R}_+$ that uniformly minimizes the Riemannian risk.

We can summarize the current statistical problem to the model corresponding to the sufficient statistic S^2 which follows a gamma distribution with parameters $\frac{n}{2\sigma^2}$ and $\frac{k}{2}$. This family is invariant under the action of the multiplicative group of $\mathbb{R}_+$ and it is straightforward to obtain that the equivariant estimators of σ which are function of $S=\sqrt{S^2}$ are of the form

$$T_\lambda(S)=\lambda S\ , \qquad \lambda\in(0,+\infty)$$

a family of estimators which contains $\sigma^b=\sqrt{\frac{k}{2}}\ \exp\left(-\frac{1}{2}\Psi\left(\frac{k}{2}\right)\right)S$, the Bayes estimator, and the maximum likelihood estimator $\sigma^*=S$. In order to obtain the equivariant estimator that minimizes the Riemannian risk, observe that the Rao distance (3) is an invariant loss function with respect to the induced group in the parameter space. This, together with the fact that the induced group acts transitively on the parameter space, makes the risk of any equivariant estimator to be constant on all the parameter space, among them the risk for σ^b and for S, as it was shown before. Therefore it is enough to minimize the risk at any point of the parameter space, for instance at $\sigma=1$. We want to

determine λ^* such that

$$\lambda^* = \arg \min_{\lambda \in (0,+\infty)} E_1\left(\rho^2(\lambda S, 1)\right)$$

It is easy to obtain

$$E_1\left(\rho^2(\lambda S, 1)\right) = E_1\left(\left(\sqrt{2}\ \ln(\lambda S)\right)^2\right) = \frac{1}{2}\Psi'\left(\frac{k}{2}\right) + \frac{1}{2}\left(\Psi\left(\frac{k}{2}\right) + \ln\left(\frac{2\lambda^2}{k}\right)\right)^2$$

which attains an absolute minimum at

$$\lambda^* = \sqrt{\frac{k}{2}}\ \exp\left(-\frac{1}{2}\Psi\left(\frac{k}{2}\right)\right)$$

so that σ^b is the minimum Riemannian risk equivariant estimator.

Finally, observe that the results in Lehmann (1951) guarantee the unbiasedness of σ^b, as we obtained before, since the multiplicative group $\mathbb{R}_+$ is commutative, the induced group is transitive and σ^b is the equivariant estimator that uniformly minimizes the Riemannian risk.

4.3 Multivariate normal, Σ known

Let us consider now the case when the sample $X_1, \ldots, X_k$ comes from a n-variate normal distribution with mean value μ and known variance–covariance matrix Σ_0, positive definite. The joint density function can be expressed as

$$f(x_1, \ldots, x_k; \mu) = (2\pi)^{-\frac{nk}{2}}\ |\Sigma_0|^{-\frac{k}{2}}\ \mathrm{etr}\left(-\frac{k}{2}\Sigma_0^{-1}\left(s^2 + (\bar{x} - \mu)(\bar{x} - \mu)^{\mathsf{T}}\right)\right)$$

where $|A|$ denote the determinant of a matrix A, $\mathrm{etr}(A)$ is equal to the exponential mapping evaluated at the trace of the matrix A, $\bar{x}$ denotes $\frac{1}{k}\sum_{i=1}^k x_i$ and s^2 stands for $\frac{1}{k}\sum_{i=1}^k (x_i - \bar{x})(x_i - \bar{x})^{\mathsf{T}}$. In this case, the metric tensor G coincides with Σ_0^{-1}. Assuming the Jeffreys prior distribution for μ, the joint density for μ and $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$ is proportional to

$$|\Sigma_0|^{-\frac{1}{2}}\ \exp\left(-\frac{k}{2}\left((\bar{x} - \mu)^{\mathsf{T}}\Sigma_0^{-1}(\bar{x} - \mu)\right)\right)$$

Next we compute the corresponding posterior distribution. Since

$$\int_{\mathbb{R}^n} \exp\left(-\frac{k}{2}\left((\bar{x} - \mu)^{\mathsf{T}}\Sigma_0^{-1}(\bar{x} - \mu)\right)\right)\ |\Sigma_0|^{-\frac{1}{2}}\ d\mu = \left(\frac{2\pi}{k}\right)^{\frac{n}{2}}$$

the posterior distribution $\pi(\cdot \mid \bar{x})$ based on the Jeffreys prior is given by

$$\pi(\mu \mid \bar{x}) = (2\pi)^{-\frac{n}{2}}\ |\frac{1}{k}\Sigma_0|^{-\frac{1}{2}}\ \exp\left(-\frac{1}{2}(\mu - \bar{x})^{\mathsf{T}}\left(\frac{1}{k}\Sigma_0^{-1}\right)(\mu - \bar{x})\right)$$

Observe here that the parameter's coordinate μ is already Cartesian, the Riemannian distance expressed via this coordinate system coincides with the Euclidean distance between the coordinates. Therefore, the Bayes estimator for μ is precisely the sample mean $\overline{x}$.

$$\mu^b(\overline{x}) = \overline{x}$$

which coincides with the maximum likelihood estimator.

Arguments of invariance that are analogous to those in the previous example apply here, where $\mu^b = \overline{X}$ is the minimum Riemannian risk equivariant estimator under the action of the translation group $\mathbb{R}^n$. The induced group is again transitive so the risk is constant at any point of the parameter space; for simplicity we may consider $\mu = 0$. A direct computation shows that

$$E_0\left(\rho^2(\overline{X}, 0)\right) = E_0\left(\overline{X}^{\mathsf{T}}\, \Sigma_0^{-1}\, \overline{X}\right) = \frac{n}{k}$$

Following Lehmann, Lehmann (1951), and observing that the translation group is commutative, $\overline{X}$ is also unbiased, as can easily be verified.

References

Abramovitz, M. (1970). *Handbook of Mathematical Functions*. New York: Dover Publications Inc.

Atkinson, C. and Mitchel, A. (1981). Rao's distance measure. *Sankhyà*, 43, A, 345-365.

Berger, J. O. and Bernardo, J. M. (1992). Bayesian Statistics 4. (Bernardo, Bayarri, Berger, Dawid, Hackerman, Smith & West Eds.), *On the development of reference priors*, (pp. 35-60 (with discussion)). Oxford: Oxford University Press.

Bernardo, J. and Juárez, M. (2003). Bayesian Statistics 7. (Bernardo, Bayarri, Berger, Dawid, Hackerman, Smith & West Eds.), *Intrinsic Estimation*, (pp. 465-476). Berlin: Oxford University Press.

Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society B*, 41, 113-147 (with discussion) Reprinted in Bayesian Inference 1 (G. C. Tiao and N. G. Polson, eds). Oxford: Edward Elgar, 229-263.

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *International Statistical Review*, 70, 351-372.

Burbea, J. (1986). Informative geometry of probability spaces. *Expositiones Mathematicae*, 4.

Burbea, J. and Rao, C. (1982). Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. *Journal of Multivariate Analysis*, 12, 575-596.

Chavel, I. (1993). *Riemannian Geometry. A Modern Introduction.* Cambridge Tracts in Mathematics. Cambridge University Press.

Erdélyi, A., Magnus, W., Oberhettinger, F., and Tricomi, F. G. (1953-1955). *Higher Transcendental Functions*, volume 1,2,3. New York: McGraw–Hill.

Hendricks, H. (1991). A Cramér-Rao type lower bound for estimators with values in a manifold. *Journal of Multivariate Analysis*, 38, 245-261.

Hicks, N. (1965). *Notes on Differential Geometry*. New York: Van Nostrand Reinhold.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society*, 186 A, 453-461.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30, 509-541.

Kendall, W. (1990). Probability, convexity and harmonic maps with small image I: uniqueness and fine existence. *Proceedings of the London Mathematical Society*, 61, 371-406.

Lehmann, E. (1951). A general concept of unbiassedness. *Annals of Mathematical Statistics*, 22, 587-592.

Oller, J. (1987). Information metric for extreme value and logistic probability distributions. *Sankhyà*, 49 A, 17-23.

Oller, J. (1993a). Multivariate Analysis: Future Directions 2. (Cuadras and Rao Eds.), *On an Intrinsic analysis of statistical estimation* (pp. 421-437). Amsterdam: Elsevier science publishers B. V., North Holland.

Oller, J. (1993b). Stability Problems for Stochastic Models. (Kalasnikov and Zolotarev Eds.), *On an Intrinsic Bias Measure* (pp. 134-158). Berlin: Lect. Notes Math. 1546, Springer Verlag.

Oller, J. and Corcuera, J. (1995). Intrinsic analysis of statistical estimation. *Annals of Statistics*, 23, 1562-1581.

Oller, J. and Cuadras, C. (1985). Rao's distance for multinomial negative distributions. *Sankhyà*, 47 A, 75-83.

Rao, C. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37, 81-91.

Smith, S. (2005). Covariance, Subspace, and Intrinsic Cramér-Rao Bounds. *IEEE Transactions on Signal Processing*, 53, 1610-1630.

Spivak, M. (1979). *A Comprehensive Introduction to Differential Geometry*. Berkeley: Publish or Perish.

Discussion of “What does intrinsic mean in statistical estimation?” by Gloria García and Josep M. Oller

Jacob Burbea

University of Pittsburgh

burbea+@pitt.edu

In this review paper, García and Oller discuss and study the concept of intrinsicality in statistical estimation, where the attention is focused on the invariance properties of an estimator.

Statistical estimation is concerned with assigning a plausible probabilistic formalism that is supposed to explain the observed data. While the inference involved in such an estimation is inductive, the formulation and the derivation of the process of estimation are based on a mathematical deductive reasoning. In the analysis of this estimation, one likes to single out those estimates in which the associated estimator possesses a certain invariance property, known as the *functional invariance of an estimator*. Such an estimator essentially represents a well-defined probability measure, and as such it is termed as an *intrinsic estimator*. The present paper revolves around this concept within the framework of a parametric statistical estimation. In this context, the probability is assumed to belong to a family that is indexed by some parameter θ which ranges in a set Θ, known as the *parameter space of the statistical model*. In particular, the resulting inductive inferences are usually formulated in the form of a point or a region estimates which ensue from the estimation of the parameter θ. In general, however, such an estimation depend on the particular parametrization of the model, and thus different estimators may lead to different methods of induction. In contrast, and by definition, intrinsic estimators do not depend on the specific parametrization, a feature that is significant as well as desirable.

In order to develop a suitable analysis, called *intrinsic analysis*, of such a statistical estimation, it is required to assess the performance of the intrinsic estimators in terms of intrinsic measures or tools. At this stage, it is worthwhile to point out that, for example, the mean square error and the bias, which are commonly accepted measures of a performance of an estimator, are clearly dependent on the model parametrization. In particular, minimum variance estimation and unbiasedness are non intrinsic concepts. To avoid such situations, the intrinsic analysis exploits the natural geometrical structures that the statistical models, under some regularity conditions, possess to construct quantities which have a well-defined geometrical meaning, and hence also intrinsic.

More explicitly, let $(\chi, \mathcal{A}, \mu)$ be a measure space and consider the map $f : \chi \times \Theta \to \mathbb{R}^+$ such that $f(\cdot|\theta)d\mu$ defines a probability measure on the measure space $(\xi, \mathcal{A})$, to be denoted as P_θ, for each $\theta \in \Theta$. In this setting, the family $(P_\theta)_{\theta\in\Theta}$ is referred to as the *statistical model under the parameter space* Θ with μ as its *reference measure*. In general, the parameter space Θ is any smooth manifold that is modeled on some Banach space. It is also assumed that the mapping $\theta \to P_\theta$ of Θ onto $(P_\theta)_{\theta\in\Theta}$ is, at least locally, injective. Moreover, since the ensuing analysis is essentially local in nature, it is sufficient, as well as customary, to use the same symbol θ to denote points in θ and their coordinates in the Hilberian tangent space of the point. Some additional regularity conditions, which are quite mild, have to be imposed on the density f to guarantee that the Fisher information matrix of the model exists and is positive-definite on Θ. In this case, Θ admits a natural Riemannian manifold structure in which the Fisher information matrix serves as a metric tensor of a differential metric, known as the *information metric* of the model, and we say that the parametric statistical model $(P_\theta)_{\theta\in\Theta}$ is *regular*. In particular, such a model may well be identified with the now Riemannian manifold Θ, and hence we have at our disposal numerous intrinsic geometrical quantities as the Riemannian information metric, its Riemannian volume elements, its indicatrice, its Levi-Civita connection, its geodesics, and its Riemannian and sectional curvatures. The geodesic distance associated with the information metric is called the *information distance*, or the *Rao distance*, of the statistical model, and is usually denoted by f. We refer to Burbea (Burbea, 1986) for additional details.

As noted, this information geometrical structure enables the intrinsic analysis of a statistical estimation to develop intrinsic quantities and concepts that are parallel to the non-intrinsic ones. For example, the square error loss is replaced by the square ρ^2 of the information distance ρ of the statistical model. There exist, of course, other possible intrinsic losses, some of which even admit a simple expression in terms of easily computed quantities of the model. In contrast, and as a disadvantage, the information distance ρ does not, in general, admit a closed form expression. However, as far as the information content of a state is concerned, the square of the information distance should be regarded as the canonical intrinsic version of the square error. In a similar fashion, the intrinsic version of the mean, or the expected value, of a random object valued on the manifold Θ, is defined in terms of an affine correction on Θ, and is said to be *canonical* when the affine connection is the Levi-Civita connection associated with the information metric on Θ. In turn, such an intrinsic version of the mean gives rise to the intrinsic version of the bias. These intrinsic concepts were first developed in Oller and Corcuera (Oller and Corcuera, 1995), where the governing analysis patterned along differential geometric lines exhibited in Karcher (Karcher, 1977). Moreover, under the assumption that Θ is a finite dimensional real manifold, Oller and Corcuera (Oller and Corcuera, 1995) were able to establish an intrinsic version of Cramér-Rao inequality. The method of proof is based on comparison theorems of Riemannian geometry (see Chavel (Chavel, 1993)). A similar result, but which a different proof, seems to appear earlier and it is due

to Hendricks (Hendricks, 1991). Recent developments on the subject matter may be founding in Smith (Smith, 2005). The obtained intrinsic Cramér-Rao inequality also a tensorial version which, on following a method of proof due to Kendall (Kendall, 1990), leads to an intrinsic version of Rao-Blackwell theorem. A more detailed account of these and related results are exposed in the present discussed paper of García and Oller.

A somewhat different approach to intrinsic estimation is obtained by considering affine connections on the manifold Θ that differ from the Levi-Civita connection associated with the information metric. In this vein, García and Oller cite a recent work of Bernardo and Juárez (Bernardo and Juárez, M., 2003) in which the concept of intrinsic estimation is based on singling out the estimator that minimizes the Bayesian risk, where the symmetrized Kullback-Leibler divergence serves as an intrinsic loss, and where the so-called *information prior* serves as a reference prior. This information prior, which is derived from information theoretical arguments, is independent of the model parametrization, and in some cases coincides with the Jeffreys uniform prior distribution, in which the, usually improper, prior is proportional to the Riemannian volume element of the information metric on Θ. As such, the obtained estimator is indeed intrinsic, for it does not depend on the model parametrization.

To illustrate and elucidate matters, García and Oller conclude their paper by exploring, for some concrete cases, the intrinsic estimator obtained by minimizing the Bayesian risk, where the square of the information distance serves as an intrinsic loss, and where the Jeffreys prior serves as a reference prior. In each case, the obtained estimator is compared with the one obtained in Bernardo and Juárez (Bernardo and Juárez, M., 2003).

In conclusion, the review of García and Oller is presented in a lucid and clear manner. It provides a virtually self contained, and quite profound, account on intrinsic estimation. As such, the review should be regarded as a solid contribution to the subject matter.

References

Abramovitz, M. (1970). *Handbook of Mathematical Functions*. New York: Dover Publications Inc.

Atkinson, C. and Mitchel, A. (1981). Rao's distance measure. *Sankhyà*, 43, A, 345-365.

Berger, J. O. and Bernardo, J. M. (1992). Bayesian Statistics 4. (Bernardo, Bayarri, Berger, Dawid, Hackerman, Smith & West Eds.), *On the development of reference priors*, (pp. 35-60 (with discussion)). Oxford: Oxford University Press.

Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society B*, 41, 113-147 (with discussion) Reprinted in Bayesian Inference 1 (G. C. Tiao and N. G. Polson, eds). Oxford: Edward Elgar, 229-263.

Bernardo, J. and Juárez, M. (2003). Bayesian Statistics 7. (Bernardo, Bayarri, Berger, Dawid, Hackerman, Smith & West Eds.), *Intrinsic Estimation*, (pp. 465-476). Berlin: Oxford University Press.

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *International Statistical Review*, 70, 351-372.

Burbea, J. (1986). Informative geometry of probability spaces. *Expositiones Mathematicae*, 4.

Burbea, J. and Rao, C. (1982). Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. *Journal of Multivariate Analysis*, 12, 575-596.

Chavel, I. (1993). *Riemannian Geometry. A Modern Introduction.* Cambridge Tracts in Mathematics. Cambridge University Press.

Erdélyi, A., Magnus, W., Oberhettinger, F., and Tricomi, F. G. (1953-1955). *Higher Transcendental Functions*, volume 1,2,3. New York: McGraw–Hill.

Hendricks, H. (1991). A Cramér-Rao type lower bound for estimators with values in a manifold. *Journal of Multivariate Analysis*, 38, 245-261.

Hicks, N. (1965). *Notes on Differential Geometry*. New York: Van Nostrand Reinhold.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society*, 186 A, 453-461.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30, 509-541.

Kendall, W. (1990). Probability, convexity and harmonic maps with small image I: uniqueness and fine existence. *Proceedings of the London Mathematical Society*, 61, 371-406.

Lehmann, E. (1951). A general concept of unbiassedness. *Annals of Mathematical Statistics*, 22, 587-592.

Oller, J. (1987). Information metric for extreme value and logistic probability distributions. *Sankhyà*, 49 A, 17-23.

Oller, J. (1993a). Multivariate Analysis: Future Directions 2. (Cuadras and Rao Eds.), *On an Intrinsic analysis of statistical estimation* (pp. 421-437). Amsterdam: Elsevier science publishers B. V., North Holland.

Oller, J. (1993b). Stability Problems for Stochastic Models. (Kalasnikov and Zolotarev Eds.), *On an Intrinsic Bias Measure* (pp. 134-158). Berlin: Lect. Notes Math. 1546, Springer Verlag.

Oller, J. and Corcuera, J. (1995). Intrinsic analysis of statistical estimation. *Annals of Statistics*, 23, 1562-1581.

Oller, J. and Cuadras, C. (1985). Rao's distance for multinomial negative distributions. *Sankhyà*, 47 A, 75-83.

Rao, C. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37, 81-91.

Smith, S. (2005). Covariance, Subspace, and Intrinsic Cramér-Rao Bounds. *IEEE Transactions on Signal Processing*, 53, 1610-1630.

Spivak, M. (1979). *A Comprehensive Introduction to Differential Geometry*. Berkeley: Publish or Perish.

Joan del Castillo

Universitat Autónoma de Barcelona, Spain

castillo@mat.uab.es

The authors are to be congratulated for summarizing material from their remarkable mathematical work over recent years in a paper that is readable to a mathematical statistician.

The paper helps us to have a clearer understanding of certain everyday concepts such as bias, mean-square error or the parametrization invariant estimator. It is important to bear in mind that bias and mean-square errors are parametrization-dependent, most particularly if we are interested in estimating probability density functions rather than parameters.

The examples given in the paper are also remarkabe. The first shows that an unbiased estimator with less mean square error than the maximum-likelihood (ML) estimator, in certain points within the parameter space, is discarded from a differential geometric point of view. Whatever the circumstances, however, the ML estimator is once again a reasonable estimator, even in a non-reasonable situation. The examples also show that the Riemannian metric introduced is a complex tool for practical applications. The Kullback-Leibler distance is often a simpler way of performing a similar analysis, as Bernardo and Juárez (2003) carried out.

In Example 4.2, corresponding to the normal distribution with a known mean value, a Bayes estimator for the standard deviation was considered from a Jeffreys prior. The estimator was found by using a Riemannian distance, solving a differential equation and some integrals. Finally, a multiple of the sample standard deviation (the ML estimator) was found. The new estimator is the equivariant estimator that uniformly minimizes the Riemannian risk. Will we now change our view that the ML estimator bias-corrected is the best estimator we can have, on the basis of this mathematical result?

From my point of view, the main question is the following: if a model is not absolutely correct, then does the Riemannian metric have any practical sense? Moreover, in applied statistical work, who really believes that a statistical model is absolutely accurate?

Wilfrid S. Kendall

w.s.kendall@warwick.ac.uk

It is a pleasure to congratulate the authors on this paper, particularly as it is gratifying to see previous work on Riemannian mean-values being put to such good statistical use! Here are three comments stimulated by the reading of this most interesting paper:

1. Another interesting statistical use of Riemannian mean-values (of a more data-analytic nature) is to be found in the work of HuiLing Le and collaborators (for example, Kume and Le 2003; Le 2004). Here one is interested in computing summary shapes to represent a whole population of shapes (often derived from biological data); one is driven to use Riemannian mean-values because there is no natural Cartesian coordinate system for shapes.
2. The original motivation for my (1990) work was entirely probabilistic, and it was natural to continue investigations using barycentres based on non-metric connections (Kendall (1991, 1992)), with close links to convexity. Non-metric connections also arise in statistical asymptotics; have the authors considered whether there is anything to be gained from using these in their context?
3. The elegant discussion in Section 3 includes the important reminder that conditional centres of mass in the Riemannian context do not obey the classic commutativity of nested conditional expectations. For workers in stochastic differential geometry this leads to the consideration of Γ-martingales, which prove of great importance in probabilistic approaches to harmonic maps. It would be interesting if the work of the current paper could be extended in this way, perhaps to take account of time-series analysis (where it would be natural to consider a whole sequence of conditional Riemannian centres of mass). The work of Darling (2002) and Srivastava and Klassen (2004) may be relevant here.

References

Darling, R. W. R. (2002). Intrinsic location parameter of a diffusion process. *Electronic Journal of Probability*, 7, 23 pp. (electronic).

Kendall, W. S. (1990). Probability, convexity, and harmonic maps with small image. I. Uniqueness and fine existence. *Proceedings of the London Mathematical Society (3)*, 61, 371-406.

Kendall, W. S. (1991). Convex geometry and nonconfluent Γ-martingales. I. Tightness and strict convexity. In *Stochastic analysis (Durham, 1990), The London Mathematical Society. Lecture Note Ser.*, 167, 163-178. Cambridge: Cambridge Univ. Press.

Kendall, W. S. (1992). Convex geometry and nonconfluent Γ-martingales. II. Wellposedness and G-martingale convergence. *Stochastics Stochastics Reports*, 38, 135-147.

Kume, A. and H. Le (2003). On Fréchet means in simplex shape spaces. *Advances in Applied Probability*, 35, 885-897.

Le, H. (2004). Estimation of Riemannian barycentres. *LMS Journal of Computation and Mathematics*, 7, 193-200 (electronic).

Srivastava, A. and E. Klassen (2004). Bayesian and geometric subspace tracking. *Advances in Applied Probability*, 36, 43-56.

Steven Thomas Smith[1]

MIT Lincoln Laboratory, Lexington, MA 02420

stsmith@ll.mit.edu

The intrinsic nature of estimation theory is fundamentally important, a fact that was realized very early in the field, as evidenced by Rao's first and seminal paper on the subject in 1945 (Rao, 1945). Indeed, intrinsic hypothesis testing, in the hands of none other than R. A. Fisher played a central role establishing one of the greatest scientific theories of the twentieth century: Wegener's theory of continental drift (Fisher, 1953). (Diaconis recounts the somewhat delightful way in which Fisher was introduced to this problem (Diaconis, 1988)). Yet in spite of its fundamental importance, intrinsic analysis in statistics, specifically in estimation theory, has in fact received relatively little attention, notwithstanding important contributions from Bradley Efron (Efron, 1975), Shun-ichi Amari (Amari, 1985), Josep Oller and colleagues (Oller, 1991), Harrie Hendriks (Hendriks, 1991), and several others. I attribute this limited overall familiarity with intrinsic estimation to three factors: (1) linear estimation theory, though in itself is implicitly intrinsic, is directly applicable to the vast majority of linear, or linearizable, problems encountered in statistics, physics, and engineering, obviating any direct appeal to the underlying coordinate invariance; (2) consequently, the number of problems demanding an intrinsic approach is limited, though in some fields, such as signal processing, nonlinear spaces abound (spheres, orthogonal and unitary matrices, Grassmann manifolds, Stiefel manifolds, and positive-definite matrices); (3) intrinsic estimation theory is really nonlinear estimation theory, which is hard, necessitating as it does facility with differential and Riemannian geometry, Lie groups, and homogeneous spaces—even Efron acknowledges this, admitting being "frustrated by the intricacies of the higher order differential geometry" [Efron, 1975, p. 1241]. García and Oller's review of intrinsic estimation is a commendable contribution to addressing this vital pedagogical matter, as well as providing many important insights and results on the application of intrinsic estimation. Their explanation of the significance of statistical invariance provides an excellent introduction to this hard subject.

1. This work was sponsored by DARPA under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the United States Government.

Intrinsic estimation concerns itself with multidimensional quantities that are invariant to the arbitrary choice of coordinates used to describe these dimensions—when estimating points on the globe, what should it matter if we choose "Winkel's Tripel" projection, or the old Mercator projection? Just like the arbitrary choice of using the units of feet or meters, the numerical answers we obtain necessarily depend upon the choice of coordinates, but the answers are invariant to them because they transform according to a change of variables formula involving the Jacobian matrix.

Yet the arbitrary choice of metric—feet versus meters—is crucial in that it specifies the performance measure in which all results will be expressed numerically. There are two roads one may take in the intrinsic analysis of estimation problems: the purely intrinsic approach, in which the arbitrary metric itself is chosen based upon an invariance criterion, or the general case in which the statistician, physicist, or engineer chooses the metric they wish to use, invariant or not, and demands that answers be given in units expressed using this specific metric. García and Oller take the purely intrinsic path. They say, "the square of the Rao distance is the most natural intrinsic version of the square error," and proceed to compute answers using this Rao (or Fisher information) metric throughout their analysis. One may well point out that this Fisher information metric is itself based upon a statistical model chosen using various assumptions, approximations, or, in the best instances, the physical properties of the estimation problem itself. Thus the Fisher information metric is natural insofar as the measurements adhere to the statistical model used for them, indicating a degree of arbitrariness or uncertainty even in this "most natural" choice for the metric. In addition to the choice of metric, the choice of score function with which to evaluate the estimation performance is also important. This Fisher score yields intrinsic Cramér-Rao bounds, and other choices of score functions yield intrinsic versions of the Weiss-Weinstein, Bhattacharyya, Barankin and Bobrovsky-Zakai bounds (Smith, Scharf and McWhorter, 2006).

Moreover, there may be legitimately competing notions of natural invariance. The invariance that arises from the Fisher metric is one kind, as recognized. But in the cases where the parameter space is a Lie group $\boldsymbol{G}$ or a (reductive) homogeneous space $\boldsymbol{G}/\boldsymbol{H}$ ($\boldsymbol{H} \subset \boldsymbol{G}$ a Lie subgroup), such as found in the examples of unitary matrices, spheres, etc. cited above, invariance to transformations by the Lie group $\boldsymbol{G}$ is typically of principal physical importance to the problem. In such cases, we may wish to analyze the square error using the unique invariant Riemannian metric on this space, e.g., the square error on the sphere would be great circle distances, not distances measured using the Fisher information metric. In most cases, this natural, intrinsic metric (w.r.t. $\boldsymbol{G}$-group invariance) is quite different from the natural, intrinsic (w.r.t. the statistical model) Fisher metric. I am aware of only one nontrivial example where these coincide: the natural $\boldsymbol{Gl}(n,\mathbb{C})$-invariant Riemannian metric for the space of positive-definite matrices $\boldsymbol{Gl}(n,\mathbb{C})/\boldsymbol{U}(n)$ is the very same as the Fisher information metric for Gaussian covariance

matrix estimation [Smith, 2005, p. 1620]:

$$g_{\mathbf{R}}(\mathbf{A}, \mathbf{A}) = \mathrm{tr}(\mathbf{A}\mathbf{R}^{-1})^2 \tag{1}$$

(ignoring an arbitrary scale factor). In this example, the square error between two covariance matrices is given by, using Matlab notation and decibel units,

$$d(\mathbf{R}_1, \mathbf{R}_2) = \mathrm{norm}(10 * \log 10(\mathrm{eig}(\mathbf{R}_1, \mathbf{R}_2))) \tag{2}$$

(i.e., the logarithm of the 2-norm of a vector of generalized eigenvalues between the covariance matrices), precisely akin to the distance between variances in Equation (3) from García and Oller's review.

Furthermore, and perhaps most importantly, the statistician or engineer may well respond "So what?" to this metric's invariance, whether it be ***G***-invariance or invariance arising from the statistical model. For example, Hendriks (Hendriks, 1991) considers the embedded metric for a parameter manifold embedded in a higher dimensional Euclidean space; the extrinsic Euclidean metric possesses very nice invariance properties, but these are typically lost within arbitrary constrained submanifolds. The choice of metric is, in fact, arbitrary, and there may be many good practical reasons to express one's answers using some other metric instead of the most natural, invariant one. Or not—the appropriateness of any proposed metric must be assessed in the context of the specific problem at hand. For these reasons, an analysis of intrinsic estimation that allows for arbitrary distance metrics is of some interest (Smith, 2005), (Smith, Scharf and McWhorter, 2006), as well the special and important special case of the Fisher information metric.

Another critical factor affecting the results obtained is the weapon one chooses from the differential geometric arsenal. García and Oller present results obtained from the powerful viewpoint of comparison theory (Cheeger and Ebin, 1975), a global analysis that uses bounds on a manifold's sectional curvature to compare its global structure to various model spaces. The estimation bounds derived using these methods possess two noteworthy properties. First, Oller and Corcuera's expressions (Oller and Corcuera, 1995) are remarkably simple! It is worthwhile paraphrasing these bounds here. Let $\boldsymbol{\theta}$ be an unknown n-dimensional parameter, $\hat{\boldsymbol{\theta}}(\mathbf{z})$ an estimator that depends upon the data $\mathbf{z}$, $\boldsymbol{A}(\mathbf{z}|\boldsymbol{\theta}) = \exp_{\boldsymbol{\theta}}^{-1} \hat{\boldsymbol{\theta}}$ the (random) vector field representing the difference between the estimator $\hat{\boldsymbol{\theta}}$ and the truth $\boldsymbol{\theta}$, $\boldsymbol{b}(\boldsymbol{\theta}) = E[\boldsymbol{A}]$ the bias vector field, "$\exp_{\boldsymbol{\theta}}$" the Riemannian exponential map w.r.t. the Fisher metric, $\bar{K} \stackrel{\mathrm{def}}{=} \max_{\boldsymbol{\theta},H} K_{\boldsymbol{\theta}}(H)$ the maximum sectional curvature of the parameter manifold over all two-dimensional subspaces H, and $D = \sup_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2} d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ be the manifold's diameter. Then, ignoring the sample size k which will always appear in the denominator for independent samples, Theorem 3.1 implies that the mean square error (MSE) about the bias—which is the random part of

the bound (add $\|\boldsymbol{b}\|^2$ for the MSE about $\boldsymbol{\theta}$)—is

$$E[\|\boldsymbol{A}-\boldsymbol{b}\|^2] \geq \begin{cases} n^{-1}(\operatorname{div} E[\boldsymbol{A}] - E[\operatorname{div}\boldsymbol{A}])^2, \\ \qquad\qquad \text{(general case);} \\ n^{-1}\Big(\operatorname{div}\boldsymbol{b} + 1 + (n-1)|\bar{K}|^{\frac{1}{2}}\|\boldsymbol{b}\|\coth(|\bar{K}|^{\frac{1}{2}}\|\boldsymbol{b}\|)\Big)^2, \\ \qquad\qquad \bar{K} \leq 0,\ \operatorname{div}\boldsymbol{b} \geq -n; \\ n^{-1}(\operatorname{div}\boldsymbol{b} + 1 + (n-1)\bar{K}^{\frac{1}{2}}D\cot(\bar{K}^{\frac{1}{2}}D))^2, \\ \qquad\qquad \bar{K} \geq 0,\ D < \pi/(2\bar{K}^{\frac{1}{2}}),\ \operatorname{div}\boldsymbol{b} \geq -1; \\ n, \qquad \bar{K} \leq 0; \\ n^{-1}, \qquad \bar{K} \geq 0,\ D < \pi/(2\bar{K}^{\frac{1}{2}}), \end{cases} \tag{3}$$

where

$$\operatorname{div}\boldsymbol{b}(\boldsymbol{\theta}) = \frac{1}{|g|^{\frac{1}{2}}}\sum_i \frac{\partial |g|^{\frac{1}{2}}\boldsymbol{b}^i(\boldsymbol{\theta})}{\partial\theta^i} \tag{4}$$

is the divergence of the vector field $\boldsymbol{b}(\boldsymbol{\theta})$ w.r.t. the Fisher metric $\mathbf{G}(\boldsymbol{\theta})$ and a particular choice of coordinates $(\theta^1, \theta^2, \ldots, \theta^n)$, and $|g|^{\frac{1}{2}} = |\det\mathbf{G}(\boldsymbol{\theta})|^{\frac{1}{2}}$ is the natural Riemannian volume form, also w.r.t. the Fisher metric.

Compare these relatively simple expressions to the ones I obtain for an arbitrary metric [Smith, 2005, Theorem 2]. In these bounds, the covariance $\mathbf{C}$ of $\boldsymbol{A}-\boldsymbol{b}$ about the bias is given by

$$\mathbf{C} \geq \mathbf{M}_{\boldsymbol{b}}\mathbf{G}^{-1}\mathbf{M}_{\boldsymbol{b}}^{\mathrm{T}} - \tfrac{1}{3}(\boldsymbol{R}_{\mathrm{m}}(\mathbf{M}_{\boldsymbol{b}}\mathbf{G}^{-1}\mathbf{M}_{\boldsymbol{b}}^{\mathrm{T}})\mathbf{G}^{-1}\mathbf{M}_{\boldsymbol{b}}^{\mathrm{T}} + \mathbf{M}_{\boldsymbol{b}}\mathbf{G}^{-1}\boldsymbol{R}_{\mathrm{m}}(\mathbf{M}_{\boldsymbol{b}}\mathbf{G}^{-1}\mathbf{M}_{\boldsymbol{b}}^{\mathrm{T}})^{\mathrm{T}}) \tag{5}$$

(ignoring negligible higher order terms), where

$$\mathbf{M}_{\boldsymbol{b}} = \mathbf{I} - \frac{1}{3}\|\boldsymbol{b}\|^2\boldsymbol{K}(\boldsymbol{b}) + \nabla\boldsymbol{b}, \tag{6}$$

$(\mathbf{G})_{ij} = g(\partial/\partial\theta^i, \partial/\partial\theta^j)$ is the Fisher information matrix, $\mathbf{I}$ is the identity matrix, $(\nabla\boldsymbol{b})^i{}_j = (\partial\boldsymbol{b}^i/\partial\theta^j) + \sum_k \Gamma^i_{jk}\boldsymbol{b}^k$ is the covariant differential of $\boldsymbol{b}(\boldsymbol{\theta})$, Γ^i_{jk} are the Christoffel symbols, and the matrices $\boldsymbol{K}(\boldsymbol{b})$ and $\boldsymbol{R}_{\mathrm{m}}(\mathbf{C})$ representing sectional and Riemannian curvature terms are defined by

$$(\boldsymbol{K}(\boldsymbol{b}))_{ij} = \begin{cases} \sin^2\alpha_i\cdot K(\boldsymbol{b}\wedge\boldsymbol{E}_i) + O(\|\boldsymbol{b}\|^3), & \text{if } i=j; \\ [\sin^2\alpha'_{ij}\cdot K(\boldsymbol{b}\wedge(\boldsymbol{E}_i+\boldsymbol{E}_j)) & \\ \quad -\sin^2\alpha''_{ij}\cdot K(\boldsymbol{b}\wedge(\boldsymbol{E}_i-\boldsymbol{E}_j))] & \text{if } i\neq j, \\ \quad + O(\|\boldsymbol{b}\|^3), & \end{cases} \tag{7}$$

α_i, α'_{ij}, and α''_{ij} are the angles between the tangent vector $\boldsymbol{b}$ and the orthonormal tangent basis vectors $\boldsymbol{E}_i = \partial/\partial\theta^i$, $\boldsymbol{E}_i + \boldsymbol{E}_j$, and $\boldsymbol{E}_i - \boldsymbol{E}_j$, respectively, and $\boldsymbol{R}_{\mathrm{m}}(\mathbf{C})$ is the mean Riemannian curvature defined by the equality

$$\langle \boldsymbol{R}_{\mathrm{m}}(\mathbf{C})\boldsymbol{\Omega}, \boldsymbol{\Omega}\rangle = E[\langle \boldsymbol{R}(\boldsymbol{X} - \boldsymbol{b}, \boldsymbol{\Omega})\boldsymbol{\Omega}, \boldsymbol{X} - \boldsymbol{b}\rangle], \tag{8}$$

where $\boldsymbol{R}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{Z}$ is the Riemannian curvature tensor. Ignoring curvature, which is reasonable for small errors and biases, as well as the intrinsically local nature of the Cramér-Rao bound itself, the intrinsic Cramér-Rao bound simplifies to the expression

$$\mathbf{C} \geq (\mathbf{I} + \nabla\boldsymbol{b})\mathbf{G}^{-1}(\mathbf{I} + \nabla\boldsymbol{b})^{\mathrm{T}}. \tag{9}$$

The trace of Equations (5) and (9) (plus the square length $\|\boldsymbol{b}\|^2$) provides the intrinsic Cramér-Rao bound on the mean square error between the estimator $\hat{\boldsymbol{\theta}}$ and the true parameter $\boldsymbol{\theta}$.

Let's take a breath and step back to compare how these bounds relate to one another, beginning with the simplest, one-dimensional Euclidean case as the basis for our comparison. The biased Cramér-Rao bound for this case is provided in an exercise from Van Trees' excellent reference [Van Trees, 1968, p. 146f]: the variance of any estimator $\hat{\theta}$ of θ with bias $b(\theta) = E[\hat{\theta}] - \theta$ is bounded from below by

$$\mathrm{Var}(\hat{\theta} - \theta - b) \geq \frac{(1 + \partial b/\partial\theta)^2}{E[(\partial \log f(\mathbf{z}|\theta)/\partial\theta)^2]}. \tag{10}$$

The denominator is, of course, the Fisher information. The numerator, $(1 + \partial b/\partial\theta)^2$, is seen in all Equations (3)–(9) above. The term $\mathrm{div}\, E[\boldsymbol{A}] - E[\mathrm{div}\, \boldsymbol{A}]$ appearing in the numerator of Equation (3) is, in this context, precisely

$$\mathrm{div}\, E[\boldsymbol{A}] - E[\mathrm{div}\, \boldsymbol{A}] = \partial b/\partial\theta - E[(\partial/\partial\theta)(\hat{\theta} - \theta)] \tag{11}$$

$$= 1 + \partial b/\partial\theta, \tag{12}$$

i.e., the one-dimensional biased CRB, as promised. Likewise, the expression $\mathbf{I} + \nabla\boldsymbol{b}$ from Equation (9) reduces to this simplest biased CRB form as well. In the general Euclidean case with a Gaussian statistical model $\log f(\mathbf{z}|\boldsymbol{\mu}) = -(\mathbf{z} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z} - \boldsymbol{\mu})$ + constants, the mean square error of any unbiased estimator of the mean $E[\mathbf{z}] = \boldsymbol{\mu}$ (with known covariance) is bounded by

$$E[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\mathrm{T}}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})] \geq \mathrm{tr}\, \mathbf{R}. \tag{13}$$

Note that this is the mean square error measured using the canonical Euclidean metric $\|\boldsymbol{\mu}\|_2^2 = \sum_i \mu_i^2$, i.e., the 2-norm. As discussed above, García and Oller use the intrinsic

Fisher metric to measure the mean square error, which results in the simpler expression

$$E[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{R}^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})] \geq \operatorname{tr}\mathbf{I} = n, \tag{14}$$

which is seen in part 4 of Equation (3), as well as in García and Oller's paper.

It may be argued that the relatively simple expressions for the bounds of Equation (3) actually conceal the true complexity found in these equations because these bounds depend on the sectional curvature and diameter of the underlying Riemannian manifold, which, especially for the case of the Fisher information metric, is typically quite involved, even for the simplest examples of Gaussian densities.

The second noteworthy property of the bounds of Equation (3) is their global, versus local, properties. It must be acknowledged that parts of these bounds, based as they are upon the global analysis of comparison theory (specifically, Bishop's comparison theorems I and II (Oller and Corcuera, 1995)), are not truly local bounds. As the square error becomes small (i.e., as either the sample size or signal-to-noise-ratio becomes large) relative to the inverse of the local curvature, the manifold appears to be locally flat (as the earth appears flat to us), and the curvature terms for a truly local Cramér-Rao bound should become increasingly negligible and approach the standard Euclidean case. This local effect is not captured by the global comparison theory, and explains why the curvature terms remain present for errors of all sizes in Equation (3). It also explains why a small change of curvature, i.e., a small change in the assumed statistical model, will result in very different numerical bounds, depending upon which of the five cases from Equation (3) applies, i.e., these bounds are not "tight"—they fall strictly below the asymptotic error of the (asymptotically efficient) maximum likelihood estimator. To see this, consider the 2-dimensional unit disk. It is flat; therefore, the fourth part of the bound applies, i.e., the lower bound on the square error is $2/k$. Now deform the unit disk a little so that it has a small amount of positive curvature, i.e., so that it is a small subset of a much larger 2-sphere. Now the fifth part of Equation (3) applies, and no matter how tiny the positive curvature, the lower bound on the square error is $1/(2k)$. This is a lower bound on the error which is discontinuous as a function of curvature. A tighter estimation bound for a warped unit disk, indeed for the case of general Riemannian manifolds, is possible. Nevertheless, the relative simplicity of these bounds lends themselves to rapid analysis, as well as insight and understanding, of estimation problems whose performance metric is Fisher information.

The tight bounds of Equations (5)–(9) can be achieved using Riemann's original local analysis of curvature (Spivak, 1999), which results in a Taylor series expansion for the Cramér-Rao bound with Riemannian curvature terms appearing in the second-order part. These are truly local Cramér-Rao bounds, in that as the square error becomes small relative to the inverse of the local curvature, the curvature terms become negligible, and the bounds approach the classical, Euclidean Cramér-Rao bound. Typical Euclidean

Cramér-Rao bounds take the form (Van Trees, 1968)

$$\mathrm{Var}(\hat{\theta} - \theta - b) \geq \frac{\mathrm{beamwidth}^2}{\mathrm{SNR}}, \tag{15}$$

where the "beamwidth" is a constant that depends upon the physical parameters of the measurement system, such as aperture and wavelength, and "SNR" denotes the signal-to-noise-ratio of average signal power divided by average noise power, typically the power in the deterministic part of the measurement divided by the power in its random part. The intrinsic Cramér-Rao bounds of (5)–(9) take the form (loosely speaking via dimensional analysis and an approximation of $\boldsymbol{A} = \exp_{\boldsymbol{\theta}}^{-1} \hat{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ using local coordinates)

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} - \boldsymbol{b}) \gtrsim \frac{\mathrm{beamwidth}^2}{\mathrm{SNR}} \left(1 - \frac{\mathrm{beamwidth}^2 \cdot \mathrm{curvature}}{\mathrm{SNR}}\right) + O(\mathrm{SNR}^{-3}). \tag{16}$$

Note that the curvature term in Equation (16), the second term in a Laurent expansion about infinite SNR, approaches zero faster than the bound itself; therefore, this expression approaches the classical result as the SNR grows large—the manifold becomes flatter and flatter and its curvature becomes negligent. The same is true of the sample size. Also note that, as the curvature is a local phenomenon, these assertions all depend precisely where on the parameter manifold the estimation is being performed, i.e., the results are local ones. In addition, the curvature term decreases the Cramér-Rao bound by some amount where the curvature is positive, as should be expected because geodesics tend to coalesce in these locations, thereby decreasing the square error, and this term increases the bound where there is negative curvature, which is also as expected because geodesics tend to diverge in these areas, thereby increasing the square error. Finally, even though this intrinsic Cramér-Rao bound is relatively involved compared to Equation (3) because it includes local sectional and Riemannian curvature terms, the formulae for these curvatures is relatively simple in the case of the natural invariant metric on reductive homogeneous spaces (Cheeger and Ebin, 1975), arguably a desirable metric for many applications, and may also reduce to simple bounds [Smith, 2005, pp. 1623–24].

These comments have focused on but a portion of the wide range of interesting subjects covered well in García and Oller's review. Another important feature of their article that warrants attention is the discussion at the end of section 4.2 about obtaining estimators that minimize Riemannian risk. It is worthwhile comparing these results to the problem of estimating an unknown covariance matrix, analyzed using intrinsic methods on the space of positive definite matrices $\boldsymbol{Gl}(n, \mathbb{C}) / \boldsymbol{U}(n)$ (Smith, 2005). As noted above, the Fisher information metric and the natural invariant Riemannian metric on this space coincide, hence García and Oller's risk minimizing estimator analysis may be applied directly to the covariance matrix estimation problem as well. Furthermore, the

development of Riemannian risk minimization for arbitrary metrics, not just the Fisher one, appears promising. This body of work points to an important, yet unanswered question in this field: Aside from intrinsic estimation bounds using various distance metrics, does the intrinsic approach yield practical and useful results useful to the community at large? Proven utility will drive greater advances in this exciting field.

References

Amari, S. (1985). Differential-Geometrical Methods in Statistics. *Lecture Notes in Statistics*, 28. Berlin: Springer-Verlag.

Amari, S. (1993). *Methods of Information Geometry*, Translations of Mathematical Monographs, **191** (transl. D. Harada). Providence, RI: American Mathematical Society, 2000. Originally published in Japanese as "Joho kika no hoho" by Iwanami Shoten, Publishers, Tokyo.

Cheeger, J. and Ebin, D. G. (1975). *Comparison Theorems in Riemannian Geometry*. Amsterdam: North-Holland Publishing Company.

Diaconis, P. (1988). Group Representations in Probability and Statistics. *IMS Lecture Notes-Monograph Series*, 11, ed. S. S. Gupta. Hayward, CA: Institute of Mathematical Statistics.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). *Annals of Statistics*, 3, 1189-1242.

Fisher, R. A. (1953). Dispersion on a sphere. *Proceedings of the Royal Society London A*, 217, 295-305.

Hendriks, H. (1991). A Cramér-Rao type lower bound for estimators with values on a manifold. *Journal of Multivariate Analysis*, 38, 245-261.

Oller, J. M. (1991). On an intrinsic bias measure. In *Stability Problems for Stochastic Models*, Lecture Notes in Mathematics 1546, eds. V. V. Kalashnikov and V. M. Zolotarev, Suzdal, Russia, 134–158. Berlin: Springer-Verlag.

Oller, J. M. and Corcuera, J. M. (1995). Intrinsic analysis of statistical estimation. *Annals of Statistics*, 23, 1562–1581.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37, 81–89, 1945. See also *Selected Papers of C. R. Rao*, ed. S. Das Gupta, New York: John Wiley, Inc. 1994.

Smith, S. T. (2005). Covariance, subspace, and intrinsic Cramér-Rao bounds. *IEEE Transactions Signal Processing*, 53, 1610–1630.

Smith, S. T., Scharf, L. and McWhorter, L. T. (2006).
Intrinsic quadratic performance bounds on manifolds. In *Proceedings of the 2006 IEEE Conference on Acoustics, Speech, and Signal Processing* (Toulouse, France).

Spivak, M. (1999). *A Comprehensive Introduction to Differential Geometry*, 3d ed., vol. 2. Houston, TX: Publish or Perish, Inc.

Van Trees, H. L. (1968). *Detection, Estimation, and Modulation Theory*, Part 1. New York: John Wiley, Inc.

Rejoinder

It is not frequent to have the opportunity to discuss some fundamental aspects of statistics with mathematicians and statisticians such as Jacob Burbea, Joan del Castillo, Wilfrid S. Kendall and Steven T. Smith, in particular those aspects considered in the present paper. We are really very fortunate in this respect and thus we want to take this opportunity to thank the former Editor of SORT, Carles Cuadras, for making it possible.

Some questions have arisen on the naturalness of the information metric and thus on the adequacy of the intrinsic bias and Riemannian risk in measuring the behaviour of an estimator, see the comments of Professors S.T. Smith and J. del Castillo. Therefore it might be useful to briefly reexamine some reasons to select such metric structure for the parameter space.

The distance concept is widely used in data analysis and applied research to study the dissimilarity between physical objects: it is considered a measure on the information available about their differences. The distance is to be used subsequently to set up the relationship among the objects studied, whether by standard statistical inference or descriptive methods.

In order to define a distance properly, from a methodological point of view, it seems reasonable to pay attention to the formal properties of the underlying observation process rather than taking into account the physical nature of the objects studied. The distance is a property neither of physical objects nor of an observer: it is a property of the observation process.

From these considerations, a necessary first step is to associate to every physical object an adequate mathematical object. Usually suitable mathematical objects for this purpose are probability measures that quantify the propensity to happen of the different events corresponding to the underlying observation process.

In a more general context we may assume that we have some additional information concerning the observation process which allows us to restrict the set of probability measures that represents our possible physical objects. These ideas lead us to consider parametric statistical models to describe the knowledge of all our possible universe of study.

The question is now: which is the most convenient distance between the probability measures corresponding to a statistical model? Observe that this question pressuposes that the right answer depends on the statistical model considered: if we can assume that all the possible probability measures corresponding to a statistical analysis are of a given type (at least approximately) this information should be relevant in order to quantify, in a right scale, the differences between two probability measures of the model. In other

words, we are are interested not only in a distance between two probability mesures but in equipping with a metric structure the set of all possible probability measures which can describe our data. In the parametric case this is equivalent to equipping the parameter space with a metric structure.

Furthermore, in order to answer these questions we have to take into account some logical reasonable restrictions that such a distance must satisfy. The first requeriment should be that any reasonable distance have to be not increasing under *general data transformations*: if we change the data, once the data has been already obtained, the corresponding distance should not increase since it does not add any information on the differences between the objects compared. Here, by general data transformation we mean either any modification of the original algebra or carrying out randomization of the data or simply data transformations.

Moreover, the distance should be invariant under *admissible transformations*, i.e. transformations which induce an *equivalent* statistical problem. These admissible transformations generalize the Fisher and Blackwell sufficiency. Several interesting approaches to this concept may be found in Strasser (1985), Heyer (1982) and Čencov (1982).

It is well known that f-divergences, Csiszar (1967), are global invariant dissimilarity measures that satisfy the above-mentioned property. We may use any of these indexes to quantify how different are two probability measures, although in general they are not proper distances. But if we are interested in defining a metric structure on the parameter space it is important to bear in mind that all this divergences induce the same, up to a multiplicative constant, intrinsic metric in the parameter space: the information metric, which is a Riemannian metric, see Burbea & Rao (1982). Therefore, the parameter space became not only a metric space but a metric length space. Finally notice that other global distance measures, like the Hellinger distance, are useful to confer a metric structure to the parameter space but not a metric length space structure which is indeed a desirable property.

Response to Jacob Burbea

It is a pleasure for the authors that a so well-known expert on the information metric structures has invest time and effort in such a detailed review on the intrinsic estimation problem.

First it is worthwhile to point out that Professor Burbea has observed the *disadvantage of the information distance in not having a closed form*, while other possible intrinsic losses do accept such a form and are thus more readable and attractive in practice. But somehow this is an a-posteriori problem which can be solved by using proper numerical evaluations or well-known tight approximations of the information distance.

But the above-mentioned *canonical* status of the measures with regard to the information distance is even more remarkable: the selection of other intrinsic losses in an estimation problem lead to *a somewhat different approach to intrinsic estimation, obtained by considering affine connections on the manifold that differ from the Levi-Civita connection*, that is considering other connections which are not compatible with the metric.

Finally, and once again, we would like to thank Professor Burbea for the commendable explanation to the insights of this subject.

Response to Joan del Castillo

The authors would like to thank Professor del Castillo for his careful reading and suggesting comments on the paper.

As Prof. del Castillo has pointed out, bias and mean square error are parametrization dependent measures and thus not invariant under admissible transformations of the parameters as well as of the random variables.

The authors have considered a purely intrinsic approach to the estimation problem by setting the loss function to be the square of the Rao distance. It is possible then to obtain the explicit form for the Bayes estimator in the examples considered.

The maximum likelihood estimator is often a convenient estimator but in some situations is an estimator that can be improved, in terms of performance or, even more, appears to be a non-reasonable estimator, see Le Cam (1990). This is the case of the example 4.1 where the authors would like to point out that no *direct considerations of differential geometry* are involved to discard the ML estimator: the reasons are purely of the statistical kind. The ML estimator misbehave in the sense that scores 0 when the sample statistic T does. This is not the case of the Bayes estimator θ^b obtained, as it is shown by the table of page 137.

The situation in Example 4.2 is slightly different from the above. As Professor del Castillo observes, the ML estimator bias-corrected is *the best estimator we can have* as long as the considered measures of performance are the bias and mean square error. The estimator σ^b obtained in Example 4.2 is the equivariant estimator that and uniformly minimizes the Riemannian Risk. Since the acting group, the multiplicative group $\mathbb{R}_+$, is commutative σ^b, is also intrinsically unbiased. One could interpret the obtained estimator as a ML Riemannian risk-corrected but we would then omit the very remarkable properties of σ^b of being equivariant and intrinsically unbiased. Anyhow the term *best estimator* makes sense insofar as the performance of an estimator is fixed.

On the other hand we may observe that the criticism concerning the adequacy of the information metric, because the model is not exactly true, apply to all the methods in parametric statistical inference, in particular to maximum likelihood estimation. We think that the information metric is a reasonable approximation to any convenient distance as far as the parametric model is a reasonable approximation to our knowledge on the studied objects.

The problem is then the following: assuming that a statistical model can not be absolutely accurate, are we concerned on not adding more noise to our results by selecting intrinsic measures of the performance?

Response to Wilfrid S. Kendall

The authors would like to take the opportunity in this rejoinder to thank twice Professor Kendall. On one hand his previous work on Riemannian barycentres and convexity were really important on the final form of several parts of the paper Oller & Corcuera (1995). Secondly, his comments and global vision on the subject connecting different research areas is already an inspiration for our future work.

The bias is a quantitative measure of the systematic error and thus should be measured in the same units as the error, the latter given by the distance. The mean square error is a quantitative measure of the impreciseness of the estimates and should be measured in the square of the distance units. Both measures are deeply related and although other connection could be used to define mean values, we believe that the Levi-Civita connection is the choice which better guarantees the intuitive meaning of both measures. Furthermore this election allows also a rather simple and natural extension to the Cramér-Rao inequality. In our opinion all other connections, as useful to give account of asymptotic estimator properties, should be regarded as other natural geometrical objects defined on the parameter space.

With respect to the interesting work of Darling (2002), we have to note that he defines intrinsic means by introducing other Riemannian metric than the information metric. This is an interesting possibility to explore but only in the case that this distance satisfies the previously-mentioned logical requirements that in our context any reasonable distance must satisfy.

Response to Steven T. Smith

The authors are totally grateful to Professor Smith for the time he has invest in the careful reading of the paper but in the extended review of the state-of-the-art of the

intrinsic estimation theory. The wide range of topics and illuminating examples covered with Professor Smith will surely help the authors in their future research work.

Concerning to the problem of selecting an appropriate distance when there are no known extrinsic reasons that force one or other selection, we cannot add much more to those reasons given before. We agree that there may exist other natural distances or indexes to be considered but again, and in our opinion, they appear of interest as far as they satisfy the previous logical requirements concerning admissible transformations.

Another interesting point involved in that question is to point out which are the basics settings determining the geometry of the parameter space: if we are only concerned to the estimation problem all the relevant information should already be incorporated in the model. The parameter space appears to be of consideration insofar it is regarded as a part of the statistical model and no isolated aspects on it yield of interest.

It has been extremely thought-provoking to the authors to follow the discussion on the tightness of the Cramér-Rao bounds. We agree that the intrinsic bounds in Theorem 3.1, (2)-(3)-(4)-(5) are not tight but if we are interested in obtaining intrinsic bounds we must do so for any estimator and not only on those which concentrate their probabilistic mass around the true parameter. Consequently we need a global analysis of the problem, which has led to the non-tightness. At any case we agree that there are interesting cases where a local analysis of the problem will be very rich. Some improved bounds, continuous on the curvature, could be obtained assuming further restriccions on the diameter of the manifold.

Observe also that the order of the approximations in any local analysis will be, in general, altered when we take expectations and this aspect should be taken into account in any risk approximations. Furthermore, it is necessary to be very careful when we develop approximations of intrinsic quantities based on coordinates point of view especially if we take expectations since the goodness of the approximation is highly dependent on the coordinate system used.

Let us finish this rejoinder hoping, like Prof. Smith, that the challenging question of making all these results closely useful to the scientific community will be attained in the future.

References

Burbea, J. & Rao, C. (1982). Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. *Journal of Multivariate Analysis, 12*, 575–596.

Čencov, N. (1982). *Statistical Decision Rules and Optimal Inference*. Providence: Trans. Math. Monographs, 53, Amer. Math. Soc. (English translation of the Russian book published in 1972, Nauka, Moscow).

Csiszar, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungrica*, *2*, 299–318.

Darling, R. (2002). Intrinsic location parameter of a diffusion process. *Electronic Journal of Probability*, *7*, 23 pp. (electronic).

Heyer, H. (1982). *Theory of Statistical Experiments*. New York: Springer-Verlag (English version of Mathematische Theorie statistischer Experiments, 1973, Springer-Verlag, Berlin).

Le Cam, L. (1990). Maximum likelihood: an introduction. *International Statistical Review*, *58*, 153–171.

Oller, J. & Corcuera, J. (1995). Intrinsic analysis of statistical estimation. *Annals of Statistics*, *23*, 1562–1581.

Strasser, H. (1985). *Mathematical Theory of Statistics: Statistical Esxperiments and Asymptotic Decision Theory*. Berlin, New York: Walter de Gruyter.

SORT 30 (2) July-December 2006, 171-192

Influence diagnostics in exponentiated-Weibull regression models with censored data

Edwin M. M. Ortega[1], Vicente G. Cancho[2] & Heleno Bolfarine[3]

[1] *ESALQ-USP,* [2] *ICMC-USP,* [3] *IME-USP*

Abstract

Diagnostic methods have been an important tool in regression analysis to detect anomalies, such as departures from the error assumptions and the presence of outliers and influential observations with the fitted models. The literature provides plenty of approaches for detecting outlying or influential observations in data sets. In this paper, we follow the local influence approach (Cook 1986) in detecting influential observations with exponentiated-Weibull regression models. The relevance of the approach is illustrated with a real data set, where it is shown that by removing the most influential observations, there is a change in the decision about which model fits the data better.

MSC: 62H10, 62J20, 62N01

Keywords: Exponentiated-Weibull distribution; censored data; local influence; influence diagnostic; survival data.

1 Introduction

In this paper we consider data sets representing the elapsed time until the occurrence of an event of interest such as the recurrence of a disease, death of a patient, failure of equipment, performance of a task, and so on.

Address for correspondence: Departamento de Ciências Exatas, USP Av. Pádua Dias 11 - Caixa Postal 9, 13418-900 Piracicaba - São Paulo - Brasil. edwin@esalq.usp.br

Edwin M. M. Ortega. ESALQ, University of São Paulo, Piracicaba, Brazil. edwin@carpa.ciagri.usp.br

Vicente G. Cancho. ICMC, University of São Paulo, São Carlos, São Paulo, Brazil. garibay@icmc.usp.br

Heleno Bolfarine. IME, University of São Paulo, São Paulo, Brazil. hbolfar@ime.usp.br

Received: March 2006

Accepted: October 2006

This elapsed time is generally termed *survival time* or life time. To simplify notation, the units under study are called individuals, and the data set is called survival data. It is common, in such kind of data, to have censored observations, that is, for some individuals the exact time of death is not known. It is only known that it lies beyond a certain value (censoring time). In this paper, we consider that the censoring times are random and noninformative. Moreover, in most situations, survival times can be affected by covariates (explanatory variables), such as the age at disease onset, blood pressure, cholesterol level, treatment type and many other important factors.

In this paper, we consider the exponentiated Weibull model, which includes as special cases the Weibull and exponential models. As considered by Mudhokar *et al.* (1995), it can be used in adjusting survival data with bathtub-type risk functions. Cancho *et al.* (1999) conducted a Bayesian study for exponentiated-Weibull regression models and Bolfarine and Cancho (2001) considered an exponentiated-Weibull survival model with a survival fraction.

An important step in regression analysis is to conduct a robustness study to detect influential or extreme observations that can cause important distortions on the results of the analysis. Numerous approaches have been proposed in the literature with a view to detect influential or outlying observations that can seriously affect parameter estimates. Studies of case deletion have were started with Cook (1977). Important reviews on the main approaches to detect influential observations are considered in Cook and Weisberg (1982) and Chaterjee and Hadi (1988).

A general framework to detect influence of observations was proposed by Cook (1986) and has often been applied with regression models. The method basically indicates how sensitive the analysis is when small perturbations are made to the data or the model. For instance, under the normal error, Lawrance (1988) investigated local influence applications in linear models with a response transformation parameter, Beckman *et al.* (1987) presented influence influence studies in mixed effects analysis of variance, Tsai and Wu (1992) considered first-order autoregressive models with nonconstant variances, and Paula (1993) used local influence methods with linear regression models when there are inequality constraints on the parameters. Moving away from normal models, Petit and Bin Daud (1989) investigated local influence with proportional hazard regression models, Escobar and Meeker (1992) adapted local influence methods to regression analysis with censoring, and O'Hara *et al.* (1992) and Kim (1995) applied local influence methods with multivariate regression. More recently, Galea *et al.* (1997) and Liu (2000) used local influence with elliptical linear regression models; Kwan and Fung (1998) applied the methodology to factor analysis and Gu and Fung (1998) discussed local influence in canonical correlation analysis. An interesting discussion and comparison with other influence measures is considered in Fung and Kwan (1997). An important extension of the method to assess the local influence of observations on the predictions from the fitted model was proposed by Thomas and Cook (1990).

In Sections 2 and 3, we review the exponentiated-Weibull regression model considered in Bolfarine *et al.* (2001). In Sections 4 and 5, we discuss the local influence method and local influence on predictions. Likelihood displacement is used to evaluate the influence of observations on the maximum likelihood estimators. Section 6 presents the results of an analysis with a real data set, including a residual analysis.

2 The exponentiated-Weibull distribution

The Weibull family of distributions has been widely used in the analysis of survival data specially in medical and engineering application. This family is suitable in situations where the risk function is constant or monotone. It is not, however, suitable in situations where the risk function is unimodal or presents a bathtub shape. Many parametric families have been considered for modeling survival data with a more general shape for the risk function. For example, Prentice (1974) considered the generalized F distribution; Stacy (1962) proposed the generalized gamma distribution while Mudhokar *et al.* (1995) presented an extension of the Weibull distribution, which is called the exponentiated Weibull family of distributions, and can adequately fit data sets presenting unimodal, monotone and bathtub shaped risk functions.

The exponentiated-Weibull distribution considered in Mudhokar et al. (1995) with parameters α, θ and σ considers that life time T has a density function given by

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{\exp\left\{-\exp\left[\dfrac{y_{i1}-\mathbf{x}^T\beta_\beta}{d_1}\right]\right\}}{2}$$

$$f(t;\alpha,\theta,\sigma) = \frac{\alpha\theta}{\sigma}\left[1-\exp\left(-\left(\frac{\mathrm{t}}{\sigma}\right)^{\alpha}\right)\right]\exp\left[-\left(\frac{\mathrm{t}}{\sigma}\right)^{\alpha}\right]\left(\frac{\mathrm{t}}{\sigma}\right)^{\alpha-1}, \quad \forall \mathrm{t} > 0 \tag{1}$$

where $\alpha > 0$, $\theta > 0$ are shape parameters and $\sigma > 0$ is a scale parameter. As a special cases, there is the Weibull distribution when $\theta = 1$ and the exponential distribution when $\alpha = 1$, $\theta = 1$. The survival function corresponding to random variable T with exponentiated-Weibull density is given by

$$S(t;\alpha,\theta,\sigma) = P(T \geq t) = 1 - \left[1-\exp\left(-\left(\frac{\mathrm{t_i}}{\sigma}\right)^{\alpha}\right)\right]^{\theta}. \tag{2}$$

The great flexibility of this model in fitting survival data can be depicted from its risk function, which can be monotonically decreasing if $\alpha \leq 1$ and $\alpha\theta \leq 1$, monotonically increasing if $\alpha \geq 1$ and $\alpha\theta \geq 1$ and present a bathtub shape if $\alpha > 1$ and $\alpha\theta < 1$.

Let $t_1, t_2, \ldots, t_n$ be a random sample of random variate T with exponentiated-Weibull distribution. The likelihood function corresponding to the observed sample is given by

$$L(t;\alpha,\theta,\sigma) = \alpha^r\theta^r\sigma^{-r\alpha}\exp\left[-\sum_{i\in F}\left(\frac{t_i}{\sigma}\right)^{\sigma}\right]\prod_{i\in F}t_i^{\alpha-1}\left(1-\exp\left[-\left(\frac{t_i}{\sigma}\right)^{\alpha}\right]\right)^{\theta-1} \quad (3)$$
$$\prod_{i\in C}\left[1-\left(1-\exp\left[-\left(\frac{t_i}{\sigma}\right)^{\alpha}\right]\right)^{\theta},\right]$$

where r is the observed number of failures, F denotes the set of uncensored observations and C denotes the set of censored observations. No explicit expressions are available for the maximum likelihood estimators of α, σ and θ, which are obtained by maximizing the log-likelihood numerically. One approach that can be used is the Newton-Raphson algorithm.

3 Exponentiated-Weibull Regression models

In many practical applications, lifetimes are affected by covariates such as cholesterol level, blood pressure and many others. The covariate vector is denoted by $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$ which is related to responses $Y = \log(T)$ through a regression model. It is also considered that the scale parameter σ of the exponentiated-Weibull model depends on the matrix of explanatory variables X. Considering the transformation $\sigma = \exp(\mu)$ and $\alpha = 1/\delta$, it follows that the density function of Y can be written as

$$f(y) = \frac{\theta}{\delta}\left\{1-\exp\left[-\exp\left(\frac{y-\mu}{\delta}\right)\right]\right\}^{\theta-1}\exp\left\{\left(\frac{y-\mu}{\delta}\right)-\exp\left(\frac{y-\mu}{\delta}\right)\right\} \quad (4)$$

$y > 0$, where $\alpha > 0$, $\theta > 0$, and $-\infty < \mu < \infty$. Using (4), we can write the above model as a log-linear model

$$Y = \mu + \delta Z \quad (5)$$

where variable Z follows the density

$$f(z) = \theta\{1-\exp[-\exp(z)]\}^{\theta-1}\exp[z-\exp(z)], \quad \forall -\infty < z < \infty \quad (6)$$

with survival function given by

$$S(y) = 1-\left\{1-\exp\left[-\exp\left(\frac{y-\mu}{\delta}\right)\right]\right\}^{\theta}. \quad (7)$$

We consider now the regression model based on the log-exponentiated-Weibull given in (5), relating response Y and covariate vector $\mathbf{x}$, so that the conditional distribution $Y|\mathbf{x}$ can be represents as

$$Y_i = \mathbf{x}_i^T\beta + \delta Z_i, \quad i = 1, \ldots, n, \tag{8}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$, $\delta > 0$ and $\theta > 0$ are unknown parameters, $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is the explanatory vector and Z follows the distribution in (7).

In this case, the survival function of $Y|\mathbf{x}$ is given by

$$S(y) = 1 - \left\{1 - \exp\left[- \exp\left(\frac{\mathrm{y} - \mathbf{x}^{\mathrm{T}}\beta}{\delta}\right)\right]\right\}^{\theta}. \tag{9}$$

Moreover, corresponding to sample $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$ of n observations from distribution (4), where y_i represents the logarithm of the survival time and $\mathbf{x}_i$ the covariate vector associated with the i-th individual, the log-likelihood function can be written as

$$\begin{aligned} l(\gamma) \;=\; & \mathrm{rlog}(\theta) - \mathrm{rlog}(\delta) + (\theta - 1)\sum_{i\in F} \log\left\{1 - \exp[-\exp(\mathrm{z_i})]\right\} + \\ & + \sum_{i\in F}\left[\mathrm{z}_i - \exp(\mathrm{z_i})\right] + \sum_{\mathrm{i}\in \mathrm{C}} \log\left\{1 - [1 - \exp(-\exp(\mathrm{z_i}))]^{\theta}\right\}, \end{aligned} \tag{10}$$

where r is the number of uncensored observations (failures) and $z_i = \frac{y_i - \mathbf{x}_i^T\beta}{\delta}$. Maximum likelihood estimates for the parameter vector $\gamma = (\theta, \delta, \beta^T)^T$ can be obtained by maximizing the likelihood function while Bayesian estimation is discussed by Cancho *et al.* (1999). In this paper, software Ox (MAXBFGS subroutine) (see Doornik, 1996) was used to compute maximum likelihood estimates (MLE). Covariance estimates for the maximum likelihood estimators $\widehat{\gamma}$ can also be obtained using the Hessian matrix. Confidence intervals and hypothesis testing can be conducted by using the large sample distribution of MLE which is a normal distribution with the covariance matrix as the inverse of the Fisher information as long as regularity conditions are satisfied. More specifically, the asymptotic covariance matrix is given by $\mathbf{I}^{-1}(\gamma)$ with $\mathbf{I}(\gamma) = -E[\ddot{\mathbf{L}}(\gamma)]$ such that $\ddot{\mathbf{L}}(\gamma) = \left\{\frac{\partial^2 l(\gamma)}{\partial\gamma\partial\gamma^T}\right\}$.

Since it is not possible to compute the Fisher information matrix $\mathbf{I}(\gamma)$ due to the censored observations (censoring is random and noninformative), it is possible to use in its place the matrix of second derivatives of the log likelihood, $-\ddot{\mathbf{L}}(\gamma)$, evaluated at the MLE $\gamma = \widehat{\gamma}$, which is consistent. Then

$$\ddot{\mathbf{L}}(\gamma) = \begin{pmatrix} \mathbf{L}_{\theta\theta} & \mathbf{L}_{\theta\delta} & \mathbf{L}_{\theta\beta} \\ . & \mathbf{L}_{\delta\delta} & \mathbf{L}_{\delta\beta} \\ . & . & \mathbf{L}_{\beta\beta} \end{pmatrix}$$

with the submatrices in appendix A.

4 Influence diagnostics

Let $l(\gamma)$ denote the log-likelihood function from the postulated model, where $\gamma = (\theta, \delta, \beta^T)^T$, and let ω be a $n \times 1$ vector of perturbations restricted to some open subset $\mathbf{\Omega} \subset \mathbb{R}^n$. The perturbations are made on the log-likelihood function. We will assume, in particular, the case-weights perturbation scheme such that the log-likelihood function takes the form

$$l(\gamma|\omega) = \sum_{i \in F} \omega_i \log f(y_i; \gamma) + \sum_{i \in C} \omega_i \log S(y_i; \gamma),$$

where $0 \leq \omega_i \leq 1$ and $\omega_0 = (1, 1, \ldots, 1)^T$ is the vector of no perturbation. Note that $l(\gamma|\omega_0) = l(\gamma)$. To assess the influence of the perturbations on the maximum likelihood estimate $\hat{\gamma}$, we consider the likelihood displacement

$$LD(\omega) = 2\{l(\hat{\gamma}) - l(\hat{\gamma}_\omega)\},$$

where $\hat{\gamma}_\omega$ denotes the maximum likelihood estimate under model $l(\gamma|\omega)$. The $LD(\omega)$ measures distance between $\hat{\gamma}$ and $\hat{\gamma}_\omega$ in terms of the log-likelihood difference. It is a nonnegative function with a global minimum at ω_0.

The idea of local influence (Cook, 1986) is concerned about characterizing the behaviour of $LD(\omega)$ around ω_0. The procedure consists in selecting a unit direction $\mathbf{d}$, $\|\mathbf{d}\| = 1$, and then to consider the plot of $LD(\omega_0 + a\mathbf{d})$ against a, where $a \in \mathbb{R}$. This plot is called *lifted line*. Note that, since $LD(\omega_0) = 0$, $LD(\omega_0 + a\mathbf{d})$ has a local minimum at $a = 0$. Each lifted line can be characterized by considering the normal curvature $C_{\mathbf{d}}(\gamma)$ around $a = 0$. This curvature is interpreted as the inverse radius of the best fitting circle at $a = 0$. The suggestion is to consider direction $\mathbf{d}_{max}$ corresponding to the largest curvature $C_{\mathbf{d}_{max}}(\gamma)$. The index plot of $\mathbf{d}_{max}$ may reveal those observations that, under small perturbations, exercise notable influence on $LD(\omega)$. Cook(1986) showed that normal curvature at direction $\mathbf{d}$ takes the form $C_{\mathbf{d}}(\gamma) = 2|\mathbf{d}^T \mathbf{\Delta}^T (\ddot{\mathbf{L}})^{-1} \mathbf{\Delta} \mathbf{d}|$ where $-\ddot{\mathbf{L}}$ is the observed Fisher information matrix for the postulated model ($\omega = \omega_0$) and $\mathbf{\Delta}$ is the $(p+1) \times n$ matrix with elements $\Delta_{ji} = \partial^2 L(\gamma|\omega)/\partial\theta_i \partial\omega_j$, evaluated at $\gamma = \hat{\gamma}$ and $\omega = \omega_0$, $j = 1, \ldots, p+2$ and $i = 1, \ldots, n$. Then, $C_{\mathbf{d}_{max}}$ is the largest eigenvalue of the matrix $\mathbf{B} = \mathbf{\Delta}^T (\ddot{\mathbf{L}})^{-1} \mathbf{\Delta}$, and $\mathbf{d}_{max}$ is the corresponding eigenvector. The index plot of $\mathbf{d}_{max}$ for matrix $\mathbf{\Delta}^T (\ddot{\mathbf{L}})^{-1} \mathbf{\Delta}$ may show how to perturb the log-likelihood function to obtain larger changes in the estimate of γ. We find, after some algebraic manipulation, the following expressions for the weighted log-likelihood function and for the elements of matrix $\mathbf{\Delta}$:

In this case the log-likelihood function takes the form

$$l(\gamma|\omega) = \Big[\mathrm{rlog}(\theta) - \mathrm{rlog}(\delta)\Big] \sum_{i \in F} w_i + (\theta - 1) \sum_{i \in F} w_i \log\left\{1 - \exp\left[-\exp\left(\frac{\mathrm{y_i} - \mathbf{x}_\mathrm{i}^\mathrm{T}\beta}{\delta}\right)\right]\right\} \tag{11}$$

$$+\sum_{i\in F} w_i\left[\frac{y_i-\mathbf{x}_i^T\beta}{\delta}\right]-\sum_{i\in F} w_i\exp\left[\frac{y_i-\mathbf{x}_i^T\beta}{\delta}\right]+$$
$$+\sum_{i\in C} w_i\log\left\{1-\left[1-\exp\Big(-\exp\big(\frac{y_i-\mathbf{x}_i^T\beta}{\delta}\big)\Big)\right]^{\theta}\right\}$$

Let us denote $\boldsymbol{\Delta}=(\boldsymbol{\Delta}_1,\ldots,\boldsymbol{\Delta}_{p+2})^T$.
Then the elements of vector $\boldsymbol{\Delta}_1$ take the form

$$\Delta_{1i}=\begin{cases}\dfrac{r}{\widehat{\theta}}+\log(\widehat{g}_i) & \text{if} \quad i\in F\\ -\dfrac{(\widehat{g}_i)^{\widehat{\theta}}}{\left[1-(\widehat{g}_i)^{\widehat{\theta}}\right]}\log\left[(\widehat{g}_i)^{\widehat{\theta}}\right] & \text{if} \quad i\in C\end{cases}$$

On the other hand, the elements of vector $\boldsymbol{\Delta}_2$ can be shown to be given by

$$\Delta_{2i}=\begin{cases}\dfrac{1}{\widehat{\delta}}\left\{-r-\dfrac{[\widehat{\theta}-1]\widehat{h}_i\widehat{z}_i}{\widehat{g}_i}-\widehat{z}_i\left[1-\exp\{\widehat{z}_i\}\right]\right\} & \text{if} \quad i\in F\\ \dfrac{\widehat{\theta}\widehat{z}_i\widehat{h}_i(\widehat{g}_i)^{\widehat{\theta}-1}}{\widehat{\delta}[1-(\widehat{g}_i)]^{\widehat{\theta}}} & \text{if} \quad i\in C\end{cases}$$

The elements of vector $\boldsymbol{\Delta}_j$, for $j=3,\ldots,p+2$, may be expressed as

$$\Delta_{ji}=\begin{cases}\dfrac{x_{ij}}{\widehat{\delta}}\left\{-\dfrac{(\widehat{\theta}-1)\widehat{h}-i}{\widehat{g}_i}+\exp\{\widehat{z}_i\}-1\right\} & \text{if} \quad i\in F\\ \dfrac{x_{ij}\widehat{\theta}\,\widehat{h}_i(\widehat{g}_i)^{\widehat{\theta}-1}}{\widehat{\delta}\left[1-(\widehat{g}_i)^{\widehat{\theta}}\right]} & \text{if} \quad i\in C\end{cases}$$

where

$$\widehat{h}_i=\exp\left[\widehat{z}_i-\exp\{\widehat{z}_i\}\right],\quad \widehat{g}_i=1-\exp\left[-\exp\{\widehat{z}_i\}\right]\quad \text{e}\quad \widehat{z}_i=\frac{y_i-\mathbf{x}_i^T\widehat{\beta}}{\widehat{\delta}}$$

However, if the interest is only in vector β, the normal curvature in direction $\mathbf{d}$ is given by $C_{\mathbf{d}}(\beta)=2|\mathbf{d}^T\boldsymbol{\Delta}^T(\ddot{\mathbf{L}}^{-1}-\mathbf{B}_{22})\boldsymbol{\Delta}\mathbf{d}|$ (see Cook, 1986), where

$$\mathbf{B}_{22}=\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \ddot{\mathbf{L}}_{22}^{-1}\end{pmatrix}$$

with $\ddot{\mathbf{L}}_{22}$ denoting the submatrix of $\ddot{\mathbf{L}}$ obtained according to partition

$$\ddot{\mathbf{L}}(\gamma) = \begin{pmatrix} \mathbf{L}_{11} & \mathbf{L}_{12} \\ \mathbf{L}_{21} & \mathbf{L}_{22} \end{pmatrix}$$

The index plot of the largest eigenvector of $\mathbf{\Delta}^T(\ddot{\mathbf{L}}^{-1} - \mathbf{B}_{22})\mathbf{\Delta}$ may reveal those observations most influential on $\hat{\beta}$.

On the other hand, considering the direction for the i-th individual the total local influence in that direction is given by

$$C_i = 2|\mathbf{\Delta}_i^T(\ddot{\mathbf{L}}^{-1} - \mathbf{B}_{22})\mathbf{\Delta}_i|. \tag{12}$$

5 Local influence on predictions

Let $\mathbf{z}$ a $p \times 1$ be a vector of values of the explanatory variables, for which we do not have necessarily an observed response. Then, the prediction at $\mathbf{z}$ is $\hat{\mu}(\mathbf{z}) = \sum_{j=1}^{p} z_j\hat{\beta}_j$. Analogously, the point prediction at $\mathbf{z}$ based on the perturbed model becomes $\hat{\mu}(\mathbf{z}, \omega) = \sum_{j=1}^{p} z_j\hat{\beta}_{j\omega}$, where $\hat{\beta}_\omega = (\hat{\beta}_{1\omega}, \ldots, \hat{\beta}_{p\omega})^T$ denotes the maximum likelihood estimate from the perturbed model. Thomas and Cook (1990) have investigated the effect of small perturbations on predictions at some particular point $\mathbf{z}$ in continuous generalized linear models and by assuming ϕ known or estimated separately from $\hat{\beta}$. ϕ^{-1} is defined as a dispersion parameter. For more details, see McCullagh and Nelder (1989). They defined three objective functions based on different residuals. Because the diagnostic calculations were identical for the proposed functions, they concentrated the application of the methodology on the objective function $f(\mathbf{z}, \omega) = \{\hat{\mu}(\mathbf{z}) - \hat{\mu}(\mathbf{z}, \omega)\}^2$.

Similarly, we will concentrate our study on investigating the normal curvature of the surface formed by vector ω and function $f(\mathbf{z}, \omega)$, around ω_0. The normal curvature at unit direction $\mathbf{d}$ takes, in this case, form $C_{\mathbf{d}}(\mathbf{z}) = 2 \mid \mathbf{d}^T\ddot{\mathbf{f}}\mathbf{d} \mid$, where $\ddot{\mathbf{f}} = \partial^2 f/\partial\omega\partial\omega^T$ is evaluated at ω_0 and $\hat{\beta}$. From Thomas and Cook (1990) one has that

$$\ddot{\mathbf{f}} = \mathbf{\Delta}^T(\ddot{\mathbf{L}}_{\beta\beta}^{-1}\mathbf{z}\mathbf{z}^T\ddot{\mathbf{L}}_{\beta\beta}^{-1})\mathbf{\Delta},$$

where $\mathbf{\Delta} = \partial^2 l(\gamma \mid \omega)/\partial\beta\partial\omega^T$. Consequently

$$\mathbf{d}_{max}(\mathbf{z}) \propto -\mathbf{\Delta}^T\ddot{\mathbf{L}}_{\beta\beta}^{-1}\mathbf{z}.$$

In the sequel, we discuss the calculation of $\mathbf{d}_{max}(\mathbf{z})$ under additive perturbations for the response and for each continuous explanatory variable.

5.1 Response perturbation

Consider the regression model (8) by assuming now that each y_i is perturbed as $y_i \to y_i + \sigma\omega_i = y_i^*$, $i = 1, \ldots, n$, with σ playing a role of scale parameter. Below, we give the expressions for the log-likelihood function and for the elements of matrix $\mathbf{\Delta}$, with $z_i^* = (y_i^* - \mathbf{x}_i^T\beta)/\delta$, $i = 1, \ldots, n$.

Here, the perturbed log-likelihood function becomes expressed as

$$\begin{aligned} l(\gamma|\omega) &= \Big[\mathrm{rlog}(\theta) - \mathrm{rlog}(\delta)\Big] + (\theta - 1)\sum_{i\in F}\log\left\{1 - \exp\Big[-\exp\Big(\frac{\mathrm{y_i^* - x_i^T\beta}}{\delta}\Big)\Big]\right\} \quad (13)\\ &+ \sum_{i\in F}\left[\frac{y_i^* - \mathbf{x}_i^T\beta}{\delta}\right] - \sum_{i\in F}\exp\left[\frac{\mathrm{y_i^* - x_i^T\beta}}{\delta}\right] +\\ &+ \sum_{i\in C}\log\left\{1 - \left[1 - \exp\Big(-\exp\big(\frac{\mathrm{y_i^* - x_i^T\beta}}{\delta}\big)\Big)\right]^{\theta}\right\} \end{aligned}$$

where $y_i^* = y_i + \sigma\omega_i$.

Matrix $\mathbf{\Delta} = (\mathbf{\Delta}_1, \ldots, \mathbf{\Delta}_{p+2})^T$ is given in appendix B.

Vector $\mathbf{d}_{max}(\mathbf{z})$ is constructed by taking $\mathbf{z} = \mathbf{x}_i$, which corresponds to the $n \times 1$ vector

$$\mathbf{d}_{max}(\mathbf{x}_i) \propto -\mathbf{\Delta}^T\ddot{\mathbf{L}}_{\beta\beta}^{-1}\mathbf{x}_i. \quad (14)$$

A large value for the ith component of (15), $\mathbf{d}_{max_i}(\mathbf{x}_i)$, indicates that the ith observation should have substantial local influence on $\hat{y}_i$. Then, the suggestion is to take the index plot of the $n\times 1$ vector $(\mathbf{d}_{max_1}(\mathbf{x}_1), \ldots, \mathbf{d}_{max_n}(\mathbf{x}_n))^T$ in order to identify those observations with high influence on its own fitted value.

5.2 Explanatory variable perturbation

Consider now an additive perturbation on a particular continuous explanatory variable, namely X_t, by making $x_{it\omega} = x_{it} + \omega_i S_t$, where S_t is a scaled factor. This perturbation scheme leads to the following expressions for the log-likelihood function and for the elements of matrix $\mathbf{\Delta}$.

The perturbed log-likelihood function is, in this case, expressed as

$$\begin{aligned} l(\gamma|\omega) &= \Big[\mathrm{rlog}(\theta) - \mathrm{rlog}(\delta)\Big] + (\theta - 1)\sum_{i\in F}\log\left\{1 - \exp\Big[-\exp\Big(\frac{\mathrm{y_i - x_i^{*T}\beta}}{\delta}\Big)\Big]\right\} \quad (15)\\ &+ \sum_{i\in F}\left[\frac{y_i - \mathbf{x}_i^{*T}\beta}{\delta}\right] - \sum_{i\in F}\exp\left[\frac{\mathrm{y_i - x_i^{*T}\beta}}{\delta}\right] + \end{aligned}$$

$$+\sum_{i\in C}\log\left\{1-\left[1-\exp\Big(-\exp\big(\frac{\mathrm{y_i}-\mathbf{x}_\mathrm{i}^{*\mathrm{T}}\beta}{\delta}\big)\Big)\right]^{\theta}\right\}$$

where $\mathbf{x}_i^{*T} = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_t(x_{it} + \omega_i S_t) + \cdots + \beta_p x_{ip}$.

Matrix $\mathbf{\Delta} = (\mathbf{\Delta}_1, \ldots, \mathbf{\Delta}_{p+2})^T$ is given in appendix C.

Similarly to the response perturbation case the suggestion here is to evaluate the largest curvature at $\mathbf{z} = \mathbf{x}_i$, which leads to

$$C_{max}(\mathbf{x}_i) = 2|\mathbf{d}_{max}^T \ddot{\mathbf{f}} \mathbf{d}_{max}|,$$

and consequently

$$\mathbf{d}_{max}(\mathbf{x}_i) \propto -\mathbf{\Delta}^T \ddot{\mathbf{L}}_{\beta\beta}^{-1} \mathbf{x}_i.$$

To see for which observed values of X_t the prediction is most sensitive under small changes in X_t, we can perform the plot of $C_{max}(\mathbf{x}_i)$ against x_{it}. The index plot of the $n \times 1$ vector $(\ell_{max_1}(\mathbf{x}_1), \ldots, \ell_{max_n}(\mathbf{x}_n))^T$ can indicate those observations for which a small perturbation in the value of X_t leads to a substantial change in the prediction.

6 Application

We provide an application of the results derived in the previous sections using simulated and real data. The required numerical evaluations were implemented using program Ox (see Doornik, 1996).

6.1 Simulation study

We conducted a simulation study to analyze the behaviour of the local influence on the exponentiated-Weibull model. The simulated data consisting of 30 uncensored observations generated from the exponentiated-Weibull distribution with $z_i = \frac{y_i - \beta_0 - \beta_1 x_i}{\delta}$. Parameter values considered were $\theta = 4$, $\delta = 2$, $\beta_0 = 4$ e $\beta_1 = 2$.

To illustrate the behaviour of the approach developed in the paper, we modified observation 26, that is, we changed $y_{26} \longrightarrow y_{26} + S_t$, where S_t corresponds to the standard deviation of response Y. Parameter estimates are presented in the Table 1.

Table 1: *Maximum likelihood estimates with standard error (SE) for simulated data.*

Parameter	Estimate	SE
θ	3.1879	4.6234
δ	1.7248	1.3133
β_0	4.2291	2.0272
β_1	1.9972	0.0288

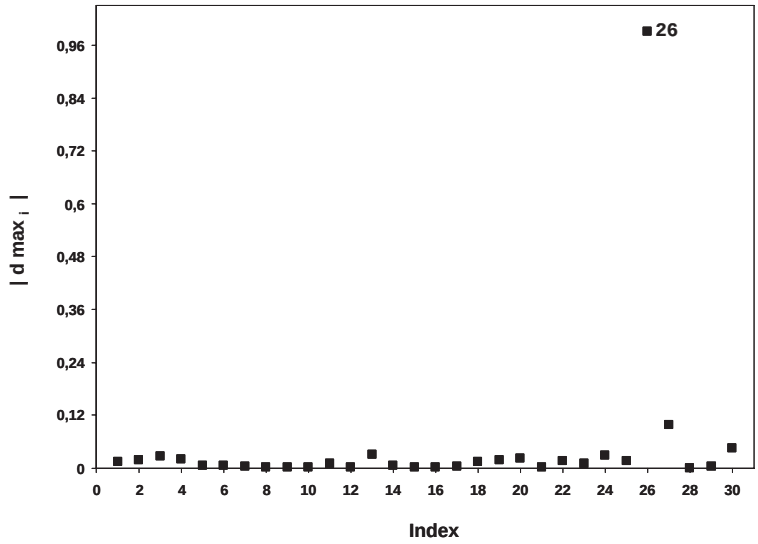

Figure 1: *Index plot of* $\mathbf{d}_{max}$ *for* γ *the simulated data (case-weights perturbation).*

Considering local influence with case weights, we obtain the maximum curvature $C_{\mathbf{d}_{\max}} = 14.677$. Vector $\mathbf{d}_{\max}$ corresponding to the direction of maximum curvature is plotted against the observation index in Figure 1, where it is clearly noted that observation 26 stands out as a possible influential observation. Similarly, Figure 2 total local influence for all observations. Observation 26 again stands out.

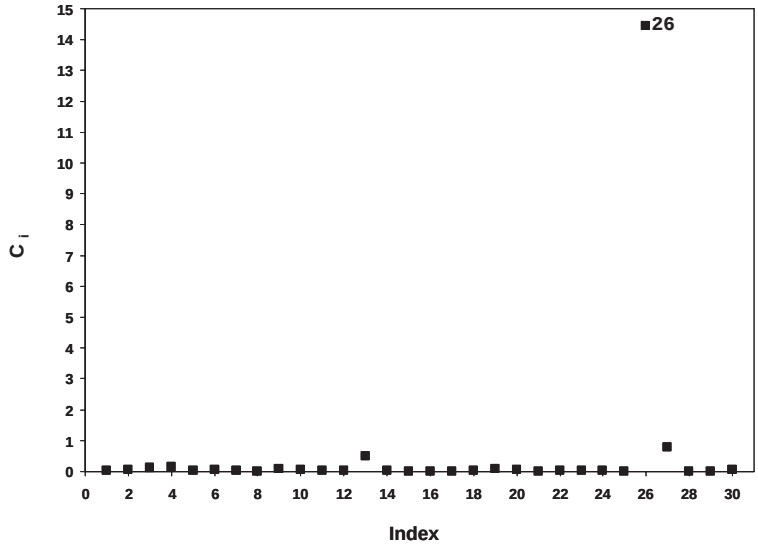

Figure 2: *Total local influence on the estimates* γ *the simulated data (case-weights perturbation).*

6.2 Golden shiner data

Survival times for the golden shiner, Notemigonus crysoleucas, were obtained from field experiments conducted in Lake Saint Pierre, Quebec, in 2005 (Laplante *et al.*, unpublished data). Individual fish were attached by means of a monofilament cord to a chronographic tethering device that allowed the fish to swim in midwater. A timer in the device was set off when the tethered fish was captured by a predator. The device was retrieved approximately 24 h after the onset of an experiment, and survival time was then obtained from the difference: time elapsed between onset of experiment and retrieval-time elapsed in device timer since predation event. The variables involved in the study were:

- y_i: survival time observed (in hours);
- $cens_i$: censoring indicator (0 = censoring, 1 = lifetime observed);
- x_{i1}: north or south bank of the lake (0 = north, 1 = south);
- x_{i2}: distance over the longitudinal axis of the lake (in km);
- x_{i3}: size of the fish (in cm);
- x_{i4}: depth of the place (in cm);
- x_{i5}: abundance index of macro-thin plants (in percentage);
- x_{i6}: transparency of the water (in cm);
- x_{i7}: initial time.

We present now results from fitting the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \delta z_i \tag{16}$$

where variable Z_i follows the log-exponentiated-Weibull distribution given in (6), $i = 1, 2, \ldots, 106$. To obtain the maximum likelihood estimates for the parameters in the model, we used the subroutine MAXBFGS in Ox, whose results are given in the following table.

Table 2: *Maximum likelihood estimates for the complete data set.*

Parameter	Estimate	SE	p-value
θ	9.4958	136.570	—
δ	4.5059	3.759000	—
β_0	−0.07456	30.482000	0.5053
β_1	2.1253	0.261380	<0.0001
β_2	0.0093338	0.0001398	0.2150
β_3	−0.12357	0.000951	<0.0001
β_4	0.033788	0.000083	<0.0001
β_5	0.022252	0.000276	0.0900
β_6	0.22427	0.040549	0.1320
β_7	−0.049872	0.025275	0.3771

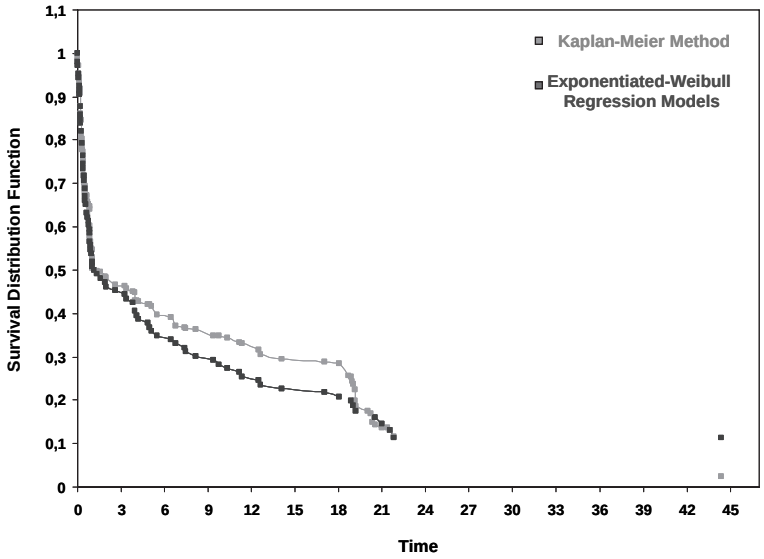

Figure 3: *Plot of the Survivor Function.*

We can see that variables x_1, x_2 and x_3 are significant for the model. We can also observe, in Figure 3, the empirical distribution function for the survival function as well as the survival function estimated by the exponentiated-Weibull regression model, where it is possible to notice a distant point in time.

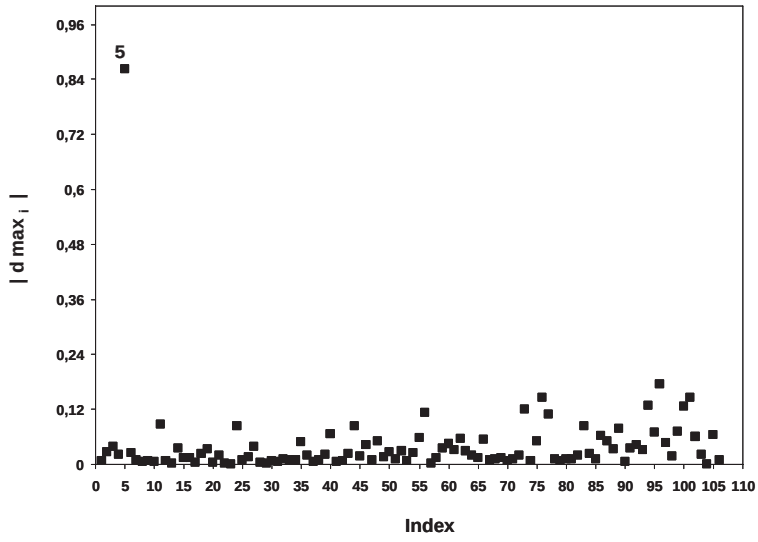

Figure 4: *Index plot of* $\mathbf{d}_{max}$ *for* γ *(case-weights perturbation).*

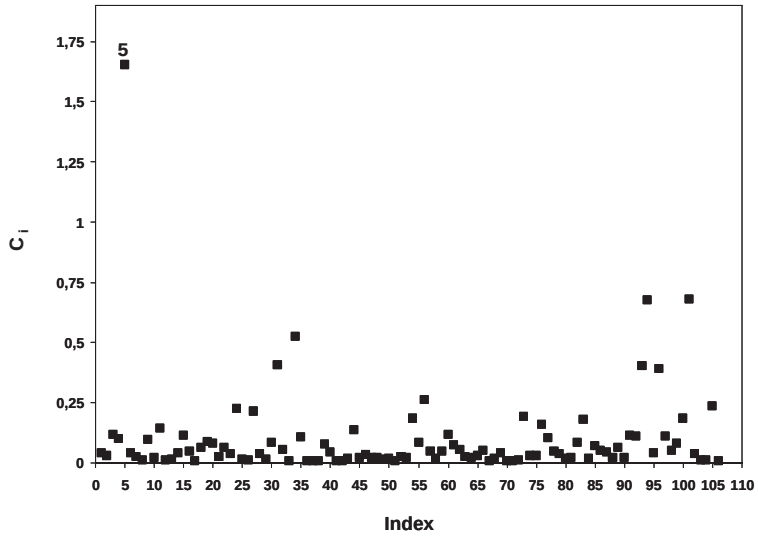

Figure 5: *Total local influence on the estimates γ (case-weights perturbation).*

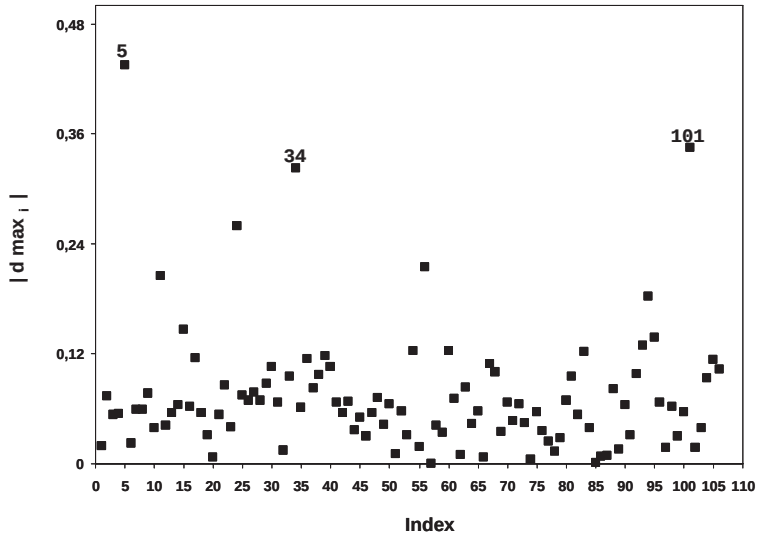

Figure 6: *Index plot of* $\mathbf{d}_{max}$ *for* γ *(response perturbation).*

6.2.1 Cases-weights perturbation

Using the exponentiated-Weibull regression model in (16), it follows that $C_{\mathbf{d}max}$ = 2.0943 with eigenvectors corresponding to $C_{\mathbf{d}max}$plotted in Figure 4 presents the plot of the eigenvector corresponding to the whole vector γ. Clearly, the most influential is

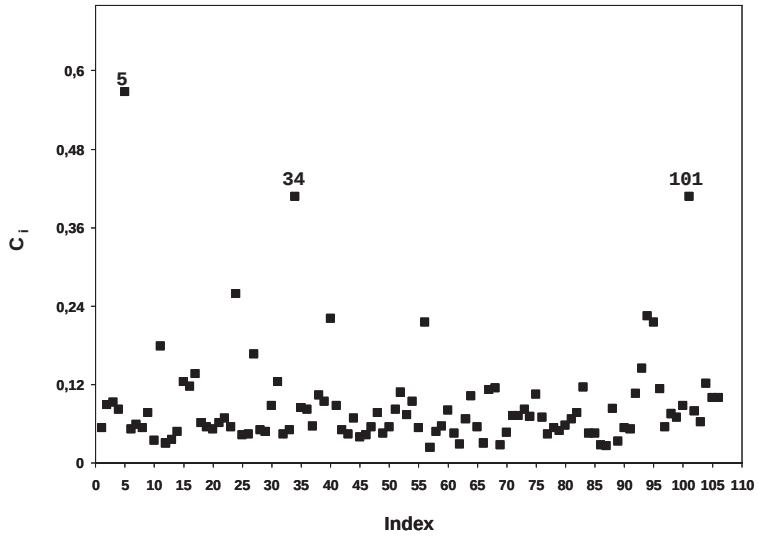

Figure 7: *Total local influence on the estimates γ (response perturbation).*

observation 5. We also use the total local influence index given in (12), whose response be found in Figure 5. We also found observation 5 as a possible influent point.

6.2.2 Prediction influence using response variable perturbation

We consider now the influence on predictions by using model (16) and the objective function proposed by Thomas and Cook(1990) as discussed in Section (5). Figure 6 and 7 present influence on the predictions by using additive perturbation in the observed response $\mathbf{y}$ ($C_{\mathbf{d}max} = 2.6454$).

6.3 Residual analysis

In order to study departures from the error assumption as well as the presence of outliers, we will first consider the martingale residual proposed by Barlow and Prentice (1988) (see also Therneau *et al.*, 1990). This residual was introduced in counting processes and can be adapted for the exponentiated-Weibull regression models as

$$r_{M_i} = \delta_i + log[S(y_i, \hat{\gamma})]$$

where $\delta_i = 0$ denotes censored observation, $\delta_i = 1$ uncensored and $S(y_i, \hat{\gamma})$ is as defined in Section 2. Due to the skewness distributional form of r_{M_i}, it has maximum value +1 and minimum value $-\infty$, transformations to achieve a more normal shaped form would be more appropriate for residual analysis. Another possibility is to use the deviance

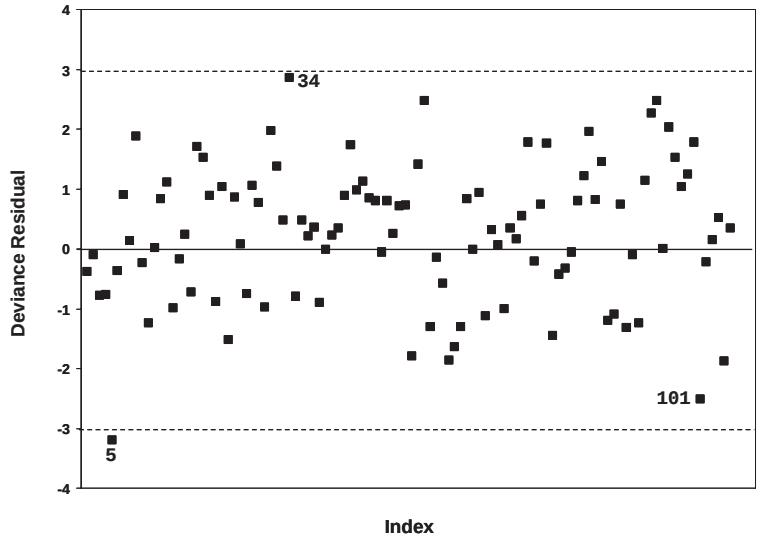

Figure 8: *Index plot of the deviance residual* r_{D_i}.

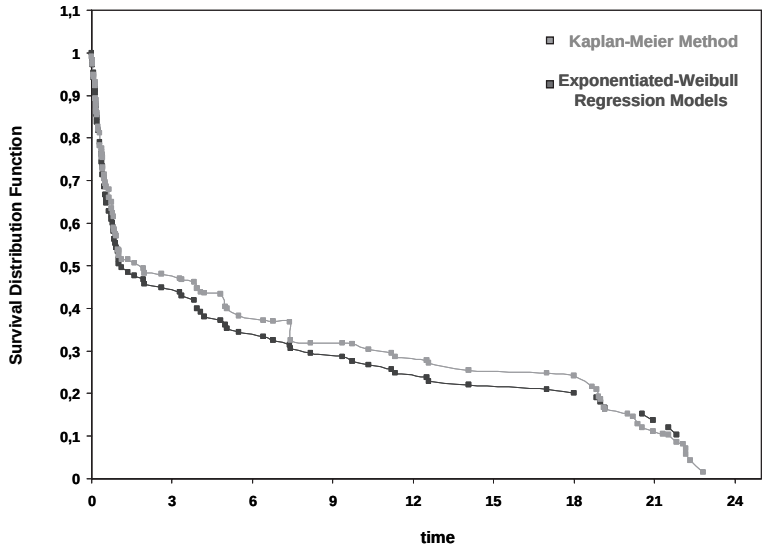

Figure 9: *Plot of the Survivor Function.*

residual (see, for instance, definition in McCullagh and Nelder, 1989, section 2.4) that has been largely applied in generalized linear models (GLMs). Various authors have investigated the use of deviance residuals in GLMs (see, for instance, Williams, 1987; Hinkley *et al.*, 1991; Paula 1995) as well as in other regression models (see, for example, Farhrmeir and Tutz, 1994). In the exponentiated-Weibull regression model, the residual

deviance can be expressed here as

$$r_{D_i} = \text{sign}(r_{M_i})\Big\{ -2\big\{r_{M_i} + \delta_i \log(\delta_i - r_{M_i})\big\}\Big\}^{\frac{1}{2}}$$

where r_{M_i} is the residual martingale corresponding to the exponentiated-Weibull regression model.

Analyzing the residual deviances, obtained after computing the residual martingales, it follows that individual 5 presented residual deviances greater than 3 (Figure 8).

6.4 Impact of the detected influential observations

To reveal the impact of the detected influential observations, we estimate the parameters again without the influential observations. Let $\hat{\gamma}$ and $\hat{\gamma}^0$ be the maximum likelihood estimates of the models that are obtained from the data sets with and without the influential observations, respectively. Lee, Lu and Song (2006) define the following two quantities to measure the difference between $\hat{\gamma}$ and $\hat{\gamma}^0$:

$$\text{TRC} = \sum_{\text{i}=1}^{\text{n}_\text{p}} \frac{|\hat{\gamma}_\text{i} - \hat{\gamma}_\text{i}^0|}{\hat{\gamma}_\text{i}} \quad \text{and} \quad \text{MRC} = \frac{\max_\text{i}|\hat{\gamma}_\text{i} - \hat{\gamma}_\text{i}^0|}{\hat{\gamma}_\text{i}}$$

where TRC is total relative changes, MRC maximum relative changes and n_p is the number of parameters.

We find that $TRC = 5.490$ and $MRC = 0.415$. In order to compare the impact of the non-influential observations, we repeat the analysis after removing the same number randomly selected from non-influential observations. We find that $TRC = 1.786$ and $MRC = 0.123$. Hence, the ML results are more sensitive to the influential observations.

Table 3: *Maximum likelihood estimates for the complete data set.*

Parameter	Estimate	SE	p-value
θ	5.556	24.761	—
δ	3.561	1.6481	—
β_0	0.28453	14.625	0.4705
β_1	2.216	0.24192	<0.0001
β_2	0.10296	0.0012952	0.0021
β_3	−0.12659	0.00083604	<0.0001
β_4	0.038063	0.0000811	<0.0001
β_5	0.0021795	0.00026622	0.4468
β_6	0.2631	0.038086	0.0888
β_7	0.028371	0.022559	0.4251

6.5 A reanalysis of golden shiner data

The model was estimated one more time, but without observation 5. Next, we present the results of the model fitting

We can observe from Table 3 that the variable x_2 became significant. The survival function was also fitted again for the exponentiated-Weibull regression model (see Figure 8) in which we can observe a good model fitting.

7 Concluding remarks

In this work, we have discussed applications of influence diagnostics in exponentiated-Weibull regression models with censored data. Appropriate matrices for assessing local influence as well as predictions on the fitted models under different perturbation schemes are obtained. Model fitting is also considered by using deviance residuals and graphs of the survival function. The approach was applied to simulated and real data sets, which clearly indicates the usefulness of the approach.

Acknowledgements

The authors acknowledge the partial financial support from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and CNPq-Brasil. The authors are thankful to two anonymous reviewers for valuable comments that substantially improved the paper.

Appendix A: Matrix of second derivatives $\ddot{\mathbf{L}}(\boldsymbol{\gamma})$

Here, we derive the necessary formulas to obtain the second order partial derivatives of the log-likelihood function. After some algebraic manipulations, we obtain

$$\mathbf{L}_{\theta\theta} = -\frac{r}{\theta^2} + \sum_{i\in C}\left\{\frac{g_i^{\theta}[\log(g_i)]^2}{(1-g_i^{\theta})^2}\right\};$$

$$\mathbf{L}_{\theta\delta} = -\frac{1}{\delta}\left\{\sum_{i\in F}\frac{z_i h_i}{g_i} + \theta\sum_{i\in C}\left[\frac{z_i h_i g_i^{\theta-1}(\log(g_i) - g_i^{\theta} + 1)}{(1-g_i^{\theta})^2}\right]\right\};$$

$$\mathbf{L}_{\theta\beta} = -\frac{1}{\delta}\left\{\sum_{i\in F}\frac{x_{ij} h_i}{g_i} + \sum_{i\in C}\left[\frac{x_{ij} h_i g_i^{\theta-1}(-g_i^{\theta} + \theta\log(g_i) + 1)}{(1-g_i^{\theta})^2}\right]\right\};$$

$$\begin{aligned}
\mathbf{L}_{\delta\delta} &= \frac{r}{\delta^2}+\frac{(\theta-1)}{\delta^2}\sum_{i\in F}\Big[\frac{g_i z_i h_i(2+h_i-h_i\exp\{z_i\})-(z_i h_i)^2}{g_i^2}\Big]+\\
&+\frac{1}{\delta^2}\sum_{i\in F}\Big[2z_i(1-\exp\{z_i\})-z_i^2\exp\{z_i\}\Big]+\\
&+\frac{\theta}{\delta^2}\sum_{i\in C}\Bigg\{\frac{z_i h_i g_i^{\theta-1}}{(1-g_i^{\theta})^2}[z_i g_i^{-1}h_i(\theta-1)+z_i(1-\exp\{z_i\})-z_i g_i^{\theta}+\\
&z_i g_i^{\theta}(g_i^{-1}h_i+\exp\{z_i\})]\Bigg\};\\
\mathbf{L}_{\delta\beta} &= -\frac{1}{\delta^2}\sum_{i\in F}\frac{(\theta-1)x_{ij}h_i}{g_i^2}\Bigg\{\Big[-g_i(1+z_i-z_i\exp\{z_i\})+z_i h_i\Big]-\\
&x_{ij}g_i^2\Big[1-\exp\{z_i\}-z_i\exp\{z_i\}\Big]\Bigg\}\\
&-\frac{\theta}{\delta^2}\sum_{i\in C}\frac{x_{ij}g_i^{\theta-1}h_i}{(1-g_i^{\theta})^2}\Bigg\{(1-g_i^{\theta-1})\Big[1-z_i g_i^{-1}h_i+z_i(1-\exp\{z_i\})\Big]+\\
&\theta z_i g_i^{-1}h_i\Bigg\};\\
\mathbf{L}_{\beta\beta} &= -\frac{(\theta-1)}{\delta^2}\sum_{i\in F}\frac{x_{ij}x_{ik}h_i\Big[g_i(-1+\exp\{z_i\})+h_i\Big]}{g_i^2}-\frac{1}{\delta^2}\sum_{i\in F}x_{ij}x_{ik}\exp\{z_i\}+\\
&+\frac{\theta}{\delta^2}\sum_{i\in C}\frac{x_{ij}x_{ik}h_i g_i^{\theta-1}\Big\{(1-g_i^{\theta-1})\Big[-1+\exp\{z_i\}-(\theta-1)h_i\Big]-\theta h_i g_i^{\theta-1}\Big\}}{(1-g_i^{\theta})^2},
\end{aligned}$$

where $h_i=\exp\Big[z_i-\exp\{z_i\}\Big]$, $\quad g_i=1-\exp\Big[-\exp\{z_i\}\Big]\quad$ and $\quad z_i=\frac{y_i-x_i^T\beta}{\delta}$.

Appendix B: Local influence on predictions: Response perturbation

Here, we provide the derivatives of elements $\mathbf{\Delta}_{ij}$ of matrix $\mathbf{\Delta}$ considering the response variables perturbation scheme. The elements of vector $\mathbf{\Delta}_1$ take the form

$$\Delta_{1i}=\begin{cases}\dfrac{\widehat{h}_i^{*}\,s}{\widehat{g}_i^{*}\,\widehat{\delta}} & \text{if}\quad i\in F\\[2ex] -\dfrac{(\widehat{g}_i^{*})^{\widehat{\theta}-1}\,\widehat{h}_i^{*}\,s}{\widehat{\delta}\,[1-(\widehat{g}_i^{*})^{\widehat{\theta}}]}\Bigg\{\widehat{\theta}\log(\widehat{g}_i^{*})\Big[\dfrac{(\widehat{g}_i^{*})^{\widehat{\theta}}}{1-(\widehat{g}_i^{*})^{\widehat{\theta}}}+1\Big]+1\Bigg\} & \text{if}\quad i\in C\end{cases}$$

On the other hand, the elements of the vector $\mathbf{\Delta}_2$ are expressed as

$$\Delta_{2i} = \begin{cases} -\dfrac{(\widehat{\theta}-1)\,\mathrm{s}\,\widehat{h}_i^*}{\widehat{\delta}^2\,\widehat{g}_i^*}\Big[-\dfrac{\widehat{z}_i^*\,\widehat{h}_i^*}{\widehat{g}_i^*}+\widehat{z}_i^*(1-\exp\{\widehat{z}_{\mathrm{i}}^*\})\Big]+\dfrac{\mathrm{s}}{\widehat{\delta}^2}\Big[\exp\{\widehat{z}_{\mathrm{i}}^*\}(\widehat{z}_{\mathrm{i}}^*+1)-1\Big] & \text{if} \quad i\in F \\ \dfrac{\mathrm{s}\,\widehat{\theta}\,\widehat{h}_i^*\,(\widehat{g}_i^*)^{\widehat{\theta}-1}}{\widehat{\delta}^2\,[1-(\widehat{g}_i^*)^{\widehat{\theta}}]}\Big\{\widehat{z}_i^*\Big[\dfrac{\widehat{h}_i^*}{\widehat{g}_i^*}\Big(\dfrac{\widehat{\theta}\,(\widehat{g}_i^*)^{\widehat{\theta}}}{1-(\widehat{g}_i^*)^{\widehat{\theta}}}+\widehat{\theta}-1\Big)-\exp\{\widehat{z}_{\mathrm{i}}^*\}+1\Big]+1\Big\} & \text{if} \quad i\in C, \end{cases}$$

while the elements of the vector $\mathbf{\Delta}_j$, $j=3,\ldots,p+2$ are expressed as

$$\Delta_{ji} = \begin{cases} \dfrac{\mathrm{s}\,x_{ij}}{\widehat{\delta}^2}\Big[-\dfrac{(\widehat{\theta}-1)\,\widehat{h}_i^*}{\widehat{g}_i^*}\Big(1-\dfrac{\widehat{h}_i^*}{\widehat{g}_i^*}-\exp\{\widehat{z}_{\mathrm{i}}^*\}\Big)+\exp\{\widehat{z}_{\mathrm{i}}^*\}\Big] & \text{if} \quad i\in F \\ \dfrac{\mathrm{s}\,\widehat{\theta}\,x_{ij}\,\widehat{h}_i^*(\widehat{g}_i^*)^{\widehat{\theta}-1}}{1-(\widehat{g}_i^*)^{\widehat{\theta}}}\Big\{\dfrac{\widehat{h}_i^*}{\widehat{g}_i^*}\Big[\dfrac{\widehat{\theta}\,(\widehat{g}_i^*)^{\widehat{\theta}}}{1-(\widehat{g}_i^*)^{\widehat{\theta}}}+\widehat{\theta}-1\Big]+1-\exp\{\widehat{z}_{\mathrm{i}}^*\}\Big\} & \text{if} \quad i\in C, \end{cases}$$

where $\widehat{h}_i^*=\exp\big[\widehat{z}_{\mathrm{i}}^*-\exp\{\widehat{z}_{\mathrm{i}}^*\}\big]$, $\widehat{g}_{\mathrm{i}}^*=1-\exp\big[-\exp\{\widehat{z}_{\mathrm{i}}^*\}\big]$ and $\widehat{z}_{\mathrm{i}}^*=\frac{y_i^*-\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}{\widehat{\delta}}$.

Appendix C: Local influence on predictions: Explanatory variable perturbation

In this appendix we provide the derivatives of elements $\mathbf{\Delta}_{ij}$ of matrix $\mathbf{\Delta}$, considering the explanatory variables perturbation scheme. The elements of vector $\mathbf{\Delta}_1$ are expressed as

$$\Delta_{1i} = \begin{cases} -\dfrac{\mathrm{s}\,\widehat{h}_i^*\,\widehat{\beta}_t}{\widehat{\delta}\,\widehat{g}_i^*} & \text{if} \quad i\in F \\ \dfrac{\mathrm{s}\,\widehat{h}_i^*\,\widehat{\beta}_t\,(\widehat{g}_i^*)^{\widehat{\theta}-1}}{\widehat{\delta}[1-(\widehat{g}_i^*)^{\widehat{\theta}}]}\Big\{1+\widehat{\theta}\log(\widehat{g}_i^*)\Big[1+\dfrac{(\widehat{g}_i^*)^{\widehat{\theta}}}{1-(\widehat{g}_i^*)^{\widehat{\theta}}}\Big]\Big\} & \text{if} \quad i\in C, \end{cases}$$

the elements of vector $\mathbf{\Delta}_2$ are expressed as

$$\Delta_{2i} = \begin{cases} \dfrac{\mathrm{s}\,\widehat{\beta}_t}{\widehat{\delta}^2}\Big\{\dfrac{(\widehat{\theta}-1)\,\widehat{h}_i^*}{\widehat{g}_i^*}\Big[\widehat{z}_i^*\Big(-\dfrac{\widehat{h}_i^*}{\widehat{g}_i^*}-\exp\{\widehat{z}_{\mathrm{i}}^*\}+1\Big)+1\Big]-\exp\{\widehat{z}_{\mathrm{i}}^*\}(1+\widehat{z}_{\mathrm{i}}^*)+1\Big\} & \text{if} \quad i\in F \\ -\dfrac{\mathrm{s}\,\widehat{\beta}_t\,\widehat{\theta}\,\widehat{h}_i^*\,(\widehat{g}_i^*)^{\widehat{\theta}-1}}{\widehat{\delta}^2\,[1-(\widehat{g}_i^*)^{\widehat{\theta}}]}\Big\{\widehat{z}_i^*\Big[\dfrac{\widehat{h}_i^*}{\widehat{g}_i^*}\Big(\dfrac{\widehat{\theta}\,(\widehat{g}_i^*)^{\widehat{\theta}}}{1-(\widehat{g}_i^*)^{\widehat{\theta}}}+\widehat{\theta}-1\Big)-\exp\{\widehat{z}_{\mathrm{i}}^*\}+1\Big]+1\Big\} & \text{if} \quad i\in C \end{cases}$$

the elements of vector $\mathbf{\Delta}_j$, for $j=1,\ldots,p$ and $j\neq t$, take the forms

$$\Delta_{ji} = \begin{cases} \dfrac{x_{ij}\,\mathrm{s}\beta_t}{\delta^2}\left\{\dfrac{(\theta-1)h_i^*}{g_i^*}\left[\dfrac{h_i^*}{g_i^*} + \exp\{\mathrm{z}_\mathrm{i}^*\} - 1\right] + \exp\{\mathrm{z}_\mathrm{i}^*\}\right\} & \text{if} \quad i \in F \\ \dfrac{\mathrm{s}\beta_t\,\theta\,h_i^*\,x_{ij}(g_i^*)^{\theta-1}}{\delta^2\,[1-(g_i^*)^{\theta}]}\left\{\dfrac{h_i^*}{g_i^*}\left[-\dfrac{\theta\,(g_i^*)^{\theta}}{1-(g_i^*)^{\theta}}\right] - \theta + 1\right\} & \text{if} \quad i \in C \end{cases}$$

the elements of vector $\boldsymbol{\Delta}_t$ are given by

$$\Delta_{ti} = \begin{cases} \dfrac{\mathrm{s}}{\delta}\left\{-\dfrac{(\theta-1)\,h_i^*}{g_i^*}\left[\dfrac{x_{it}\beta_t}{\delta}\left(\dfrac{h_i^*}{g_i^*} + \exp\{\mathrm{z}_\mathrm{i}^*\}\right) - 1\right] + \exp\{\mathrm{z}_\mathrm{i}^*\} + \left[1 - \dfrac{\beta_\mathrm{t}\,\mathrm{x}_\mathrm{it}}{\delta}\right] - 1\right\} & \text{se} \quad i \in F \\ \dfrac{\mathrm{s}\,\theta\,h_i^*\,(g_i^*)^{\theta-1}}{\delta\,[1-(g_i^*)^{\theta}]}\left\{-\dfrac{\beta_t x_{it}}{\delta}\left[\dfrac{h_i^*}{g_i^*}\left(\dfrac{\theta(g_i^*)^{\theta}}{1-(g_i^*)^{\theta}} + \theta - 1\right) - \exp\{\mathrm{z}_\mathrm{i}^*\} + 1\right] + 1\right\} & \text{se} \quad i \in C \end{cases}$$

where $\widehat{h}_i^* = \exp\left[\widehat{\mathrm{z}}_\mathrm{i}^* - \exp\{\widehat{\mathrm{z}}_\mathrm{i}^*\}\right]$, $\widehat{\mathrm{g}}_\mathrm{i}^* = 1 - \exp\left[-\exp\{\widehat{\mathrm{z}}_\mathrm{i}^*\}\right]$ and $\widehat{\mathrm{z}}_\mathrm{i}^* = \frac{\mathrm{y_i} - \mathbf{x}_\mathrm{i}^{*\mathrm{T}}}{\widehat{\delta}}$

References

Barlow, W. E., and Prentice, R. L.(1988). Residual for relative risk regression. *Biometrika*, 75, 65-74.

Beckman, R. J., Nachtsheim, C. J. and Cook, R. D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics*, 29, 413-426.

Bolfarine, H. and Cancho, V. (2001). Modelling the presence of immunes by using the exponentiated-Weibull model. *Journal of Applied Statistics*, 28, 659-671.

Cancho, V.; Bolfarine, H. and Achcar, J. A. (1999). A Bayesian analysis for the exponentiated-Weibull distribution. *Journal Applied Statistical Science*, 8, 227-242.

Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression.* New York: John Wiley.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society B*, 30, 248-275.

Collet, D. (1994). *Modelling Survival Data in Medical Research.* Chapman and Hall: London.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, 19 15-18.

Cook, R. D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society*, 48, 133-169.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* New York: Chapman and Hill.

Davison, A. C. and Gigli, A. (1989). Deviance residuals and normal scores plots. *Biometrika*, 76, 211-221.

Doornik, J. (1996). *Ox: An Object-Oriented Matrix Programming Language*. International Thomson Business Press.

Escobar, L. A. and Meeker, W. Q. (1992). Assessing influence in regression analysis with censored data. *Biometrics*, 48, 507-528.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer-Verlag: New York.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Process and Survival Analysis.* New York: John Wiley.

Fung, W. K. and Kwan, C. W. (1997). A note on local influence based on normal curvature. *Journal of the Royal Statistical Society B*, 59, 839-843.

Galea, M., Paula, G. A. and Bolfarine, H. (1997). Local influence in elliptical linear regression models. *The Statistician*, 46, 71-79.

Gu, H. and Fung, W. K. (1998). Assessing local influence in canonical analysis. *Annals of the Institute of Statistical Mathematics*, 50, 755-772.

Hinklei, D. V., Reid, N. and Snell, E. J.(1991). *Statistical Theory and Modelling-In honor of Sir David Cox*. London: Chapman and Hall.

Kim, M. G. (1995). Local influence in multivariate regression. *Communications in Statistics Theory and Methods*, 20, 1271-1278.

Kwan, C. W. and Fung, W. K. (1998). Assesing local influence for specific restricted likelihood: Applications to factor analysis. *Psychometrika*, 63, 35-46.

Lawrence, A. J. (1988). Regression transformation diagnostics using local influence. *Journal of the American Statistical Association*, 83, 1067-1072.

Lawless, J. F. (1982). *Statistical Models and Methods for lifetime data*. New York: John Wiley.

Lee, S. Y., Lu, B. and Song. X. Y. (2006). Assessing local influence for nonlinear structural equation models with ignorable missing data. *Computational Statistics and Data Analysis*, 50, 1356-1377.

Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, 54, 570-582.

Liu, S. Z. (2000). On local influence for elliptical linear models. *Statistical Papers*, 41, 211-224.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd Edition. London: Chapman and Hall.

Mudholkar, G. S.; Srivastava, D. K. and Friemer, M. (1995). The exponentiated Weibull family: A reanalysis of the bus-motor-failure data. *Technometrics*, 37, 436-445.

Nelson, W. B. (1990). *Accelerated Testing; Statistical Models, Test Plans and Data Analysis*. New York: John Wiley.

O'Hara, R. J., Lawless, J. F. and Carter, E. M. (1992). Diagnostics for a cumulative multinomial generalized linear model with application to grouped toxicological mortality data. *Journal of the American Statistical Association*, 87, 1059-1069.

Ortega, E. M. M., Bolfarine, H. and Paula G. A. (2003). Influence diagnostics in generalized log-gamma regression models. *Computational Statistics and Data Analysis*, 42, 165-186.

Paula, G. A. (1993). Assessing local influence in restricted regressions models. *Computational Statistics and Data Analysis*, 16, 63-79.

Paula, G. A. (1995). Influence residuals in restricted generalized linear models. *Journal of Statistical Computation and Simulation*, 51, 63-79.

Pettitt, A. N. and Bin Daud, I. (1989). Case-weight measures of influence for proportional hazards regression. *Applied Statistics*, 38, 51-67.

Prentice, R. L. (1974). A log-gamma model and its maximum likelihood estimation. *Biometrica*, 61, 539-544.

Stacy, E. W. (1962). A generalization of the gamma distribution. *Annals of Mathematical Statistics*, 33 1187-1192.

Thomas, W. and Cook, R. D. (1990). Assessing influence on predictions from generalized linear models. *Technometrics*, 32, 59-65.

Tsai, C. and Wu, X. (1992). Transformation-Model diagnostics. *Technometrics*, 34, 197-202.

Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77, 147-60.

Williams, D. A. (1987). Generalized linear model diagnostic using the deviance and single case deletion. *Applied Statistics*, 36, 181-191.

SORT 30 (2) July-December 2006, 193-204

Statistical models to study subtoxic concentrations for some standard mutagens in three colon cancer cell lines

Xavier Bardina[1], Laura Fernández[2], Elisabet Piñeiro[2], Jordi Surrallés[2] & Antonia Velázquez[2]

Universitat Autònoma de Barcelona

Abstract

The aim of this work is to propose models to study the toxic effect of different concentrations of some standard mutagens in different colon cancer cell lines. We find estimates and, by means of an inverse regression problem, confidence intervals for the subtoxic concentration, that is the concentration that reduces by thirty percent the number of colonies obtained in the absence of mutagen.

MSC: 62J05

Keywords: Inverse regression problem, subtoxic concentration, confidence interval.

1 Introduction

Human populations are exposed to a variety of environmental agents, including biological, chemical and physical entities, that can injure the DNA and cause adverse health consequences, such as cancer. It is therefore extremely important to detect these mutagenic agents, unravel their mechanisms of action, and define what type of injury they produce. All this information is crucial to determine the genetic risk of exposed population.

[1] Departament de Matemàtiques, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain. bardina@mat.uab.cat

[2] Departament de Genètica, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain. lfernandez@einstein.uab.es, epineiro@einstein.uab.es, jordi.surralles@blues.uab.es, avh@cc.uab.es

Received: September 2006
Accepted: November 2006

Mutagenicity assays are specifically designed to detect DNA damaging agents and to analyze their biological effects. Most of these assays are performed *in vitro* with a variety of cell types and under controlled conditions of cell growth and viability. The range of the treatment concentrations is usually determined in a previous toxicity study. The chosen concentrations for the mutagenicity study must be subtoxic in order to ensure biological effects without extreme cell death. This is why a well performed toxicity assay is absolutely required as a previous routine in all mutagenicity assays.

Tandem repeated DNA sequences of few nucleotides, the so-called microsatellites, are known to be highly unstable in colon cancer cells defective in DNA mismatch repair (MMR). In addition, expansion of tandem repeated sequences have been causally related to a number of degenerative diseases including miotonic dystrophy, fragile X syndrome and Huntington's disease. The final aim of the present investigation was to determine whether microsatellite instability is inducible *in vitro* by a set of mutagens of different mode of action. To do so, subtoxic concentrations, i.e. those inducing a reduction of thirty percent in cell viability, had to be previously determined for the following standard mutagens: bleomycin (BLEO), N-methyl-N-nitrosourea (MNU), ethoposide (ETO), mitomycin C (MMC) and ethidium bromide (EtBr). The toxicity assay was carry out in three human fibroblast cell lines derived from colon tumours: the wild-type cell line SW480, and lines HCT116 and LoVo which are both defective in MMR.

The toxicity data were obtained by the colony forming efficiency method. About 200 cells from exponentially growing cell cultures were platted in triplicate on $25cm^2$ plates (falcons). After allowing for attachment to the plate for 24 hours, the medium was replaced with fresh medium containing the test chemical at different concentration for each replica. Cell lines were maintained in these conditions for 10 days, replacing the medium every 3 days. The plates were washed with phosphate buffer saline, fixed with methanol, and stained with Giemsa. Colonies with more than 50 growing cells were counted. A reduction in the number of colonies after treatment is interpreted to be a consequence of the chemical toxicity.

In this article we will propose different types of statistical models to study the effect of the concentration of the mutagens in three colon cancer cell lines. We have considered linear and exponential models and in each case we propose the model that approximates better the data. When we consider a regression linear model, we need to assume that the errors are additive and normally distributed. When the model considered is an exponential one, we assume that the errors are multiplicative and their logarithms are normally distributed. In all the models proposed, the analysis of the residuals does not give evidence that controvert this hypothesis. For each mutagen, cell line and concentration we have three values. Then, we will consider models with weights, where the weights are calculated by the inverse of the estimated variances.

For each mutagen and each cell line we will obtain an estimation and a confidence interval for the subtoxic concentration, that is the concentration that reduces a thirty percent the initial number of colonies.

This article is organized as follows. In Section 2 we provide the mathematical justification for the calculation of the confidence intervals used in this study. In Section 3 we present some models for the standard mutagens: bleomycin, N-methyl-N-nitosourea, ethoposide, mitomycin C and ethidium bromide respectively. We also provide for each mutagen and each cell line the estimate of the subtoxic concentration and a confidence interval for this concentration. In Section 4 the use of weighted models is justified. Finally, in the appendix the data used in the study are shown.

2 Mathematical justification of the confidence intervals used in this study

In this section we will explain the method used in order to obtain a confidence interval of the subtoxic concentration, that is the concentration that reduces by thirty percent the initial number of colonies.

Consider the linear regression model

$$y = \beta_0 + \beta_1 x,$$

where y is the number of colonies and x the concentration of the mutagen.

We will estimate the concentration that reduces to 70 percent the number of colonies for $x = 0$ (that is, the concentration x such that $y = 0.7\beta_0$) by

$$\hat{x} = \frac{-0.3 \cdot \hat{\beta}_0}{\hat{\beta}_1},$$

where $\hat{\beta}' = \left(\hat{\beta}_0, \hat{\beta}_1\right)$ is the usual estimate of (β_0, β_1) obtained by the least-squares method. That is,

$$\hat{\beta} = (X'X)^{-1} X'y,$$

where

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

and $(y_1, x_1), \cdots, (y_n, x_n)$ are the data of the number of colonies and concentrations, respectively.

If we consider the vector $\lambda' = (\lambda_1, \lambda_2)$, it is well known that we can obtain a $100(1 - \alpha)$ % confidence region for $\lambda'\beta$ by using the following expression:

$$\lambda'\beta = \lambda'\hat{\beta} \pm t_{\alpha/2,n-2} S \sqrt{\lambda'\Lambda\lambda},$$

where $\Lambda = (X'X)^{-1}$, S is the estimate of the standard deviation of the errors and $t_{\alpha/2,n-2}$ is the critical point such that $P(|T| > t_{\alpha/2,n-2}) = \alpha$ where T is a Student's t distribution with $n-2$ degrees of freedom.

On the other hand, from the linear regression model, the concentration x that reduces to 70 percent the inicial number of colonies satisfies that

$$0.7\beta_0 = \beta_0 + \beta_1 x,$$

that is

$$0.3\beta_0 + x\beta_1 = 0$$

and we can write this expression as

$$\lambda'\beta = 0$$

with $\lambda' = (0.3, x)$.

So, in order to obtain a sort of $100(1-\alpha)$ % confidence interval for the concentration x that reduces a 30 % the inicial number of colonies, we can solve the following system:

$$\begin{cases} \lambda'\beta = \lambda'\hat{\beta} \pm t_{\alpha/2,n-2} S \sqrt{\lambda'\Lambda\lambda} \\ \lambda'\beta = 0 \end{cases} \tag{1}$$

whith $\lambda' = (0.3, x)$. That is, we have to find the intersections (if there exists) between the line $\lambda'\beta = 0$ and the curves given by $\lambda'\beta = \lambda'\hat{\beta} \pm t_{\alpha/2,n-2} S \sqrt{\lambda'\Lambda\lambda}$. This kind of problems are called inverse regression problems (see for example Draper and Smith, 1981).

When we use a linear model with weights, the matrix Λ now is given by

$$\Lambda = \left(X'V^{-1}X\right)^{-1}$$

and the estimate of β is

$$\hat{\beta} = \left(X'V^{-1}X\right)^{-1} X'V^{-1}y,$$

where

$$V^{-1} = \begin{pmatrix} w_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w_n \end{pmatrix}$$

and $w_1, w_2, \dots, w_n$ are the weights of the data $y_1, y_2, \dots, y_n$, respectively (see for example

Montgomery, 1992, or Draper and Smith, 1981). Observe that in our examples the confidence interval are approximated because the weights are estimated from the data.

When we consider an exponential model, that is,

$$y = e^{\beta_0+\beta_1 x},$$

we assume that the residuals are multiplicative and that their logarithms are normally distributed. That is, we suppose that if we consider the transformation

$$\ln y = \beta_0 + \beta_1 x$$

we can proceed like in a linear model. But in this case, we have to estimate the concentration x such that $y = 0.7e^{\beta_0}$. Thus, we will estimate x by

$$\hat{x} = \frac{\ln 0.7}{\hat{\beta}_1}.$$

Then, we obtain the confidence interval solving the system:

$$\begin{cases} \lambda'\beta - \ln 0.7 = \lambda'\hat{\beta} - \ln 0.7 \pm t_{\alpha/2,n-2} S \sqrt{\lambda'\Lambda\lambda} \\ \lambda'\beta - \ln 0.7 = 0, \end{cases}$$

where $\lambda' = (0,\ x)$. We remark that this confidence interval has an exact confidence level of $100(1-\alpha)$ % only when the weights are perfectly known. If the weights are estimated from the sample then the confidence levels are approximated.

3 Examples

Using the method described in Section 2 we will obtain now a confidence interval for the subtoxic concentration of the Bleomycin in the cell line LoVo. From the data given in the Appendix we have that

$$X = \begin{pmatrix} 1 & 0.0000 \\ 1 & 0.0000 \\ 1 & 0.0000 \\ 1 & 0.0001 \\ \vdots & \vdots \\ 1 & 0.0500 \end{pmatrix}, \quad y = \begin{pmatrix} 26.5 \\ 29.0 \\ 31.0 \\ 32.5 \\ \vdots \\ 6.5 \end{pmatrix}, \quad V^{-1} = \begin{pmatrix} 0.1967 & 0 & \dots & 0 \\ 0 & 0.1967 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0.2449 \end{pmatrix}.$$

Thus, with an easy computation we obtain that

$$\Lambda = \left(X'V^{-1}X\right)^{-1} = \begin{pmatrix} 1.039 & -22.082 \\ -22.082 & 1008.297 \end{pmatrix},$$

and the estimate of β is given by

$$\hat{\beta} = \left(X'V^{-1}X\right)^{-1} X'V^{-1}y = \begin{pmatrix} 27.205 \\ -440.677 \end{pmatrix}.$$

In this example the estimate of the standard deviation of the errors is $S = 1.083$ and the critical point of the Student's t is $t_{\frac{0.05}{2},11} = 2.201$.

Then in order to obtain the confidence interval explained in this section we have to solve the system (1). In this example we get

$$\begin{cases} 0.3\beta_0 + \beta_1 x = 8.161 - 440.677x \pm 2.383\sqrt{0.093 - 13.249x + 1008.297x^2} \\ 0.3\beta_0 + \beta_1 x = 0. \end{cases}$$

So, we have to find the roots of the following equation:

$$440.677x - 8.161 = \pm 2.383\sqrt{0.093 - 13.249x + 1008.297x^2},$$

that is, the roots of the quadratic equation

$$33182.804x^2 - 1253.193x + 11.634 = 0,$$

that are given by the values

$$\begin{aligned} x_1 &= 0.0164 \\ x_2 &= 0.0213. \end{aligned}$$

Then, (0.0164, 0.0213) is a confidence interval of approximately 95 % for the subtoxic concentration of the Bleomycin in the cell line Lovo.

3.1 Bleomycin (BLEO)

We have considered a regression line with weights for each cell line. Recall that y is the number of colonies and x the concentration of bleomycin. The models obtained using the SAS system are:

cell line	regression line
HCT116	$y = 67.597 - 1168.925\,x + \varepsilon$
LoVo	$y = 27.205 - 440.677\,x + \varepsilon$
SW480	$y = 53.488 - 566.464\,x + \varepsilon$

The regression coefficients R-square were 0.9285, 0.9372 and 0.9452, respectively. Finally, with these models we obtain the following estimates, and the following confidence intervals, for the subtoxic concentration in each cell line.

cell line	concentration estimate	95% confidence interval
HCT116	0.0173	(0.0152, 0.0203)
LoVo	0.0185	(0.0164, 0.0213)
SW480	0.0283	(0.0259, 0.0315)

3.2 N-methyl-N-nitrosourea (MNU)

We have considered a regression line with weights for the cell lines LoVo and SW480. In this case, the models obtained using the SAS system are the following:

cell line	regression line
LoVo	$y = 22.117 - 0.218\,x + \varepsilon$
SW480	$y = 67.826 - 0.663\,x + \varepsilon$

The regression coefficients R-square were 0.9330 and 0.9512, respectively. With these models we obtain the following estimates, and the following confidence intervals, for the subtoxic concentration in each cell line.

cell line	concentration estimate	95% confidence interval
LoVo	30.477	(29.299, 31.777)
SW480	30.708	(29.249, 32.344)

3.3 Ethoposide (ETO)

We have considered a regression line with weights for each cell line. The models obtained using the SAS system are:

cell line	regression line
HCT116	$y = 130.075 - 860.173\,x + \varepsilon$
LoVo	$y = 43.509 - 233.079\,x + \varepsilon$
SW480	$y = 104.453 - 423.569\,x + \varepsilon$

The regression coefficients R-square were 0.9032, 0.7189 and 0.8282, respectively. With these models we obtain the following estimates, and the following confidence intervals, for the subtoxic concentration in each cell line.

cell line	concentration estimate	95% confidence interval
HCT116	0.0454	(0.0380, 0.0563)
LoVo	0.0560	(0.0432, 0.0838)
SW480	0.0740	(0.0593, 0.0996)

3.4 Mitomycin C (MMC)

For this data sets we have fitted a regression line with weights for each cell line. The models obtained using the SAS system are:

cell line	regression line
HCT116	$y = 85.321 - 6334.859\,x + \varepsilon$
LoVo	$y = 13.989 - 714.984\,x + \varepsilon$
SW480	$y = 48.554 - 4174.451\,x + \varepsilon$

The regression coefficients R-square were 0.9253, 0.5364 and 0.9314, respectively. With these models we obtain the following estimates, and the following confidence intervals, for the subtoxic concentration in each cell line.

cell line	concentration estimate	95% confidence interval
HCT116	0.0040	(0.0036, 0.0046)
LoVo	0.0059	(0.0040, 0.0115)
SW480	0.0035	(0.0032, 0.0039)

3.5 Ethidium bromide (EtBr)

In this situation we have considered a exponential model with weights for the cell lines HCT116 and SW480, i.e., a regression linear model with weights for the logarithms of the data. The models obtained using the SAS system are the following:

cell line	model
HCT116	$y = \exp(4.477 - 19.474\,x) \times \varepsilon$
SW480	$y = \exp(3.722 - 24.330\,x) \times \varepsilon$

The regression coefficients R-square were 0.8584 and 0.8601, respectively. With these models we obtain the following estimates, and the following confidence intervals, for the subtoxic concentration in each cell line.

cell line	concentration estimate	95% confidence interval
HCT116	0.0183	(0.0147 , 0.0242)
SW480	0.0147	(0.0118, 0.0193)

4 Conclusions

Our exploration of the data sets (see the Appendix) demonstrates the non-homogeneity of the variances of the errors of the corresponding models, and consequently the non-adequateness of the classical least squares method to estimate the parameters. For instance, for the Ethoposide line HCT116, the variances corresponding to each concentration are 2.583, 288.250, 64.333, 376.333 and 82.583. The Levene test, that can be performed with the SPSS statistical package, rejects the equaliy of variances with $p = .029$.

Then a possible option would be to transform the dependent variable by using a suitable stabilizing variance transformation. However the usual transformations (powers, logarithms, etc.) are employed when certain specific patterns are observed between the sampling variances and their respective means. This is not the situation as can be shown in the example above by plotting the variances against their corresponding means.

Another option (those considered in this report) is to use the weighted linear regression models. These models can be implemented by using any of the more usual statistical packages such as SAS or SPSS. The ideal setting is when the weight (the inverse of the variance) for each observation is perfectly known. There are real examples where it happens (see Draper and Smith, 1981) but this is not our case here – for the data sets studied in this report the weights are estimated.

For the linear regression model we have expressed the value of the variable y (number of colonies) corresponding to the subtoxic concentration x as a linear combination of the coefficients of the regression line. The same has happened with the exponential model and the variable $\ln y$. This is the key point that has allowed us to find the confidence intervals for the subtoxic concentrations. That is, in a linear regression model (respectively, in an exponential model), this method can be applied to find confidence intervals for the value of the variable x for which the variable y (respectively, the variable $\ln y$) can be expressed as a linear combination of the coefficients of the regression line.

Appendix

In this Appendix the data used in this work are provided. The methodology used in order to obtain these results has been explained in Section 1. We express also, between parenthesis, the weight of each datum. Recall from the theory of linear models that the weights are given by the inverse of the estimated variances of the three data obtained for each concentration. Observe that some data corresponding to the number of colonies are not integer numbers. The reason is that the number of colonies was counted twice, and we have used their average.

Bleomycin

observation	concentration	HCT116	LoVo	SW480
1	0.0000	59.0 (0.0159)	26.5 (0.1967)	54.5 (0.0612)
2	0.0000	56.0 (0.0159)	29.0 (0.1967)	46.5 (0.0612)
3	0.0000	71.0 (0.0159)	31.0 (0.1967)	51.5 (0.0612)
4	0.0001	64.0 (0.0273)	32.5 (0.0221)	65.5 (0.0263)
5	0.0001	61.5 (0.0273)	23.0 (0.0221)	62.0 (0.0263)
6	0.0001	73.0 (0.0273)	*	53.5 (0.0263)
7	0.0050	58.0 (0.0058)	24.5 (0.1071)	55.0 (0.0569)
8	0.0050	76.5 (0.0058)	18.5 (0.1071)	48.0 (0.0569)
9	0.0050	*	22.5 (0.1071)	47.5 (0.0569)
10	0.0100	79.0 (0.0106)	19.0 (0.0556)	41.0 (0.0065)
11	0.0100	66.0 (0.0106)	25.0 (0.0556)	58.5 (0.0065)
12	0.0100	60.0 (0.0106)	*	*
13	0.0500	11.5 (0.0473)	6.5 (0.2449)	23.5 (0.4285)
14	0.0500	5.0 (0.0473)	3.0 (0.2449)	25.5 (0.4285)
15	0.0500	*	6.5 (0.2449)	26.5 (0.4285)

N-methyl-N-nitrosourea

observation	concentration	LoVo	SW480
1	0	34.5 (0.0317)	73.5 (0.0698)
2	0	27.0 (0.0317)	74.5 (0.0698)
3	0	23.5 (0.0317)	67.5 (0.0698)
4	1	23.0 (0.0323)	53.0 (0.0811)
5	1	34.0 (0.0323)	60.0 (0.0811)
6	1	30.0 (0.0323)	56.0 (0.0811)
7	5	20.5 (0.1791)	114.0 (0.0018)
8	5	19.5 (0.1791)	69.0 (0.0018)
9	5	16.0 (0.1791)	81.0 (0.0018)
10	10	15.5 (0.0968)	91.0 (0.0200)
11	10	21.5 (0.0968)	78.5 (0.0200)
12	10	16.5 (0.0968)	79.0 (0.0200)
13	25	16.5 (0.0424)	99.5 (0.0029)
14	25	26.0 (0.0424)	95.5 (0.0029)
15	25	19.5 (0.0424)	65.5 (0.0029)
16	50	22.0 (0.0141)	45.5 (0.0293)
17	50	20.0 (0.0141)	41.5 (0.0293)
18	50	6.5 (0.0141)	34.0 (0.0293)
19	100	1.0 (3.0000)	0.0 (0.6316)
20	100	0.0 (3.0000)	1.5 (0.6316)
21	100	0.0 (3.0000)	2.5 (0.6316)

Ethoposide

observation	concentration	HCT116	LoVo	SW480
1	0.0000	129.0 (0.3871)	81.0 (0.0012)	140.5 (0.0167)
2	0.0000	129.5 (0.3871)	28.0 (0.0012)	112.5 (0.0167)
3	0.0000	132.0 (0.3871)	35.0 (0.0012)	120.0 (0.0167)
4	0.0006	136.0 (0.0035)	39.0 (0.0033)	104.0 (0.1071)
5	0.0006	132.5 (0.0035)	70.0 (0.0033)	100.0 (0.1071)
6	0.0006	105.0 (0.0035)	40.5 (0.0033)	106.0 (0.1071)
7	0.0060	131.5 (0.0155)	38.5 (0.1081)	86.0 (0.0017)
8	0.0060	122.5 (0.0155)	43.5 (0.1081)	86.5 (0.0017)
9	0.0060	115.5 (0.0155)	44.0 (0.1081)	128.0 (0.0017)
10	0.0300	96.5 (0.0027)	33.0 (0.0116)	65.5 (0.0060)
11	0.0300	131.5 (0.0027)	43.5 (0.0116)	82.0 (0.0060)
12	0.0300	99.5 (0.0027)	25.0 (0.0116)	91.0 (0.0060)
13	0.0600	81.0 (0.0121)	27.5 (0.1165)	75.0 (0.0603)
14	0.0600	68.0 (0.0121)	28.5 (0.1165)	82.5 (0.0603)
15	0.0600	85.5 (0.0121)	33.0 (0.1165)	81.5 (0.0603)

Mitomycin C

observation	concentration	HCT116	LoVo	SW480
1	0.0000	78.5 (0.0204)	18.0 (0.0902)	*
2	0.0000	92.0 (0.0204)	12.5 (0.0902)	*
3	0.0000	88.5 (0.0204)	18.5 (0.0902)	*
4	0.0005	86.0 (0.0227)	20.0 (0.1081)	60.0 (0.0142)
5	0.0005	74.5 (0.0227)	19.5 (0.1081)	59.0 (0.0142)
6	0.0005	74.5 (0.0227)	25.0 (0.1081)	45.0 (0.0142)
7	0.0010	73.0 (0.0553)	12.5 (4.0000)	39.0 (0.0663)
8	0.0010	74.5 (0.0553)	13.0 (4.0000)	46.5 (0.0663)
9	0.0010	81.0 (0.0553)	13.5 (4.0000)	41.0 (0.0663)
10	0.0025	72.0 (0.0769)	15.0 (0.0902)	36.0 (0.0074)
11	0.0025	67.0 (0.0769)	9.0 (0.0902)	50.5 (0.0074)
12	0.0025	74.0 (0.0769)	9.5 (0.0902)	27.5 (0.0074)
13	0.0050	63.0 (0.0526)	16.0 (0.0383)	31.0 (0.0902)
14	0.0050	55.0 (0.0526)	14.5 (0.0383)	29.0 (0.0902)
15	0.0050	56.0 (0.0526)	6.5 (0.0383)	24.5 (0.0902)
16	0.0100	13.0 (0.0356)	8.0 (0.5000)	5.0 (0.0902)
17	0.0100	20.5 (0.0356)	6.0 (0.5000)	4.5 (0.0902)
18	0.0100	*	*	10.5 (0.0902)

Ethidium bromide

observation	conc.	HCT116	LN HCT116	SW480	LN SW480
1	0.000	105.5	4.66 (14.92)	23.5	3.16 (12.29)
2	0.000	85.0	4.44 (14.92)	39.0	3.66 (12.29)
3	0.000	63.0	4.14 (14.92)	38.0	3.64 (12.29)
4	0.001	100.0	4.61 (3.78)	53.0	3.97 (87.81)
5	0.001	49.0	3.89 (3.78)	46.0	3.83 (87.81)
6	0.001	133.0	4.89 (3.78)	43.0	3.76 (87.81)
7	0.010	82.5	4.41 (27.33)	29.0	3.37 (114.39)
8	0.010	102.5	4.63 (27.33)	27.0	3.30 (114.39)
9	0.010	70.0	4.24 (27.33)	32.5	3.48 (114.39)
10	0.050	26.0	3.26 (57.23)	9.0	2.20 (12.03)
11	0.050	33.5	3.51 (57.23)	15.5	2.74 (12.03)
12	0.050	31.5	3.45 (57.23)	10.0	2.30 (12.03)
13	0.100	15.0	2.71 (7.47)	5.0	1.61 (6.47)
14	0.100	13.0	2.56 (7.47)	3.0	1.10 (6.47)
15	0.100	26.0	3.26 (7.47)	6.5	1.87 (6.47)

Acknowledgements

The experimental data shown in this report has obtained thanks to a research grant by the Fundació Marató TV3. We would like to thank Professors Federic Utzet and Pere Puig for their helpful comments and suggestions during the preparation of the report.

References

Boyer, J. C., Umar, A., Risinger, J. I., Lipford, J. R., Kane, M., Ying, S., Barret, J. C., Kolodner, R. D. and Kunkel, T. A. (1995). *Microsatellite instability, mismatch repair deficency, and genetic defects in human cancer cell lines.* Cancer research, 55, 6063-6070.

Draper, N. and Smith, H. (1981). *Applied Regression Analysis* (second edition). New York: John Wiley.

Montgomery, DC. and Peck, EA. (1992). *Introduction to Linear Regression Analysis.* New York: John Wiley.

Shuterland, G. I. and Richards, R. I. (1995). Simple tandem repeats and human genetic disease. Proceedings of the National Academy of Sciences of the United States of America, 92, 3636-3641.

SORT 30 (2) July-December 2006, 205-218

Univariate Parametric Survival Analysis using GS-distributions

A. Sorribas*, J. M. Muiño and M. Rué

Universitat de Lleida

Abstract

The GS-distribution is a family of distributions that provide an accurate representation of any unimodal univariate continuous distribution. In this contribution we explore the utility of this family as a general model in survival analysis. We show that the survival function based on the GS-distribution is able to provide a model for univariate survival data and that appropriate estimates can be obtained. We develop some hypotheses tests that can be used for checking the underlying survival model and for comparing the survival of different groups.

MSC: 62N01, 62N02, 62N03, 62E17, 62P10

Keywords: Univariate distributions, survival analysis, Kaplan-Meier

1 Introduction

Survival analysis refers to the analysis of time-to-event processes when a censoring mechanism prevents the observation of the precise time at what the event occurs in some of the individuals. A typical situation arises in medical studies where the goal is to estimate the time that a given patient will survive after a treatment and to evaluate the different factors that can influence this process. Censoring appears when a patient is lost to follow-up or when the event occurs between two observation times. In all those cases, estimation of the survival function $P(T > t) = S(t)$ can be obtained by constructing the appropriate likelihood. For example, in a right-censoring scheme some events are not

* *Address for correspondence:* Departament de Ciències Mèdiques Bàsiques, Institut de Recerca Biomèdica de Lleida (IRBLLEIDA), Universitat de Lleida, Carrer Montserrat Roig, 2. 25008-Lleida (Spain). tel: +34 973 702 406, fax: +34 973 702 426, e-mail: albert.sorribas@cmb.udl.es

Received: October 2006
Accepted: November 2006

observed due to loss of follow-up. In that case the likelihood can be constructed as (see for instance Hosmer & Lemeshow, 1999; and Klein & Moeschberger, 1997)

$$L \propto \prod_{i \in U} f(x_i) \prod_{i \in R} S(t_i) \quad i = 1, \ldots, n \tag{1}$$

where U indicates the set of uncensored event times and R the set of right censored times. More complicated censoring schemes can be accommodated by considering the contribution of each observation to the likelihood. If a parametric model exists for the survival process, the analysis is straightforward and reduces to estimating the corresponding parameters. However, in most cases such a model is unknown making it necessary to consider non-parametric approaches. For example, the Kaplan-Meier estimator provides a good estimation of $S(t)$ for right-censored data (see for instance Klein & Moeschberger, 1997). Methods for estimating $S(t)$ for interval-censored data and other censoring schemes are still under investigation (Zhao & Sun, 2004).

If we could identify a sufficiently general family of distributions that can be used as a general model for survival analysis, we could use it as an alternative to the non-parametric approach for any censoring scheme. In this work, we introduce the GS-distribution as a family of distributions that can be used for this purpose in univariate unimodal continuous survival processes. We shall briefly recall the GS-distribution family of distributions and then we will introduce its use in survival analysis.

1.1 GS-distribution

The GS-distribution is defined as (Muiño *et al.* (2006)):

$$\frac{dF}{dx} = \alpha F^g (1 - F^k)^\gamma \;\; F(x_0) = F_0 \tag{2}$$

Here, $F = P(X \leq x)$ and α, g, k, γ are real non-negative parameters. F_0 is the value of the cumulative distribution at x_0. Without lack of generality, we can take $F_0 = 0.5$. In such a case, x_0 is the median. We shall refer to this distribution as GSD$[F_0, x_0, \alpha, g, k, \gamma]$. This family of distributions provides a useful model for accurately represent univariant unimodal continuous distributions. It can be shown that some distributions are exactly represented by the GS-distribution. For instance, the logistic case corresponds to $g = \gamma$ and $k = 1$. The uniform, the exponential, and some Beta and **F** distributions are also exact cases. All symmetric distributions correspond to the special cases $g = \gamma$ and $k = 1$ (see Muiño *et al.* (2006)). Other distributions are accurately approximated by a GS-distribution, although they do not correspond to an exact case (see Muiño *et al.* (2006) for a complete discussion and examples). This family generalizes the S-distribution which corresponds to the case $\gamma = 1$ (Voit 1992, 2000; Voit & Schwacke, 1998; Voit & Sorribas, 2000; Voit *et al.* 1995). Other families based on quantiles are also included.

For instance, the case $k = 1$ corresponds to the Q-distribution family (see Turner & Pruitt, 1978; Parzen, 1979; Kamps, 1991; Jones 2002, 2004).

As the GS-distribution is defined as a differential equation, computation of its properties and its use for data modeling requires appropriate algorithms (Voit 1992; Voit & Schwake, 2000; Sorribas *et al.* 2000, 2002; Hernández-Bermejo & Sorribas, 2001; March *et al.* 2002). Due to the flexibility of this family, it can be used as a general model that can appropriately fit observed data. As the GS-distribution can approximate almost any univariate unimodal continuous distribution, fitting a GS-distribution is equivalent to selecting the best distribution that can describe the data. As the GS-distribution is quite general, if necessary we can introduce some constraints that limit the fit. For example, if we want to consider only symmetric distributions, we can fit a GS-distribution with $g = \gamma$ and $k = 1$.

In this work we consider using this family as a parametric family for survival analysis (see also Yu & Voit, 1995). In the next section, we shall introduce this model in the context of survival analysis and show its utility with uncensored data. Then we will move to comparing survival curves using this model and to more complex censoring schemes.

2 Survival model based on GS-distributions

The survival function is related to the distribution function as:

$$S(t) = 1 - F(t) \tag{3}$$

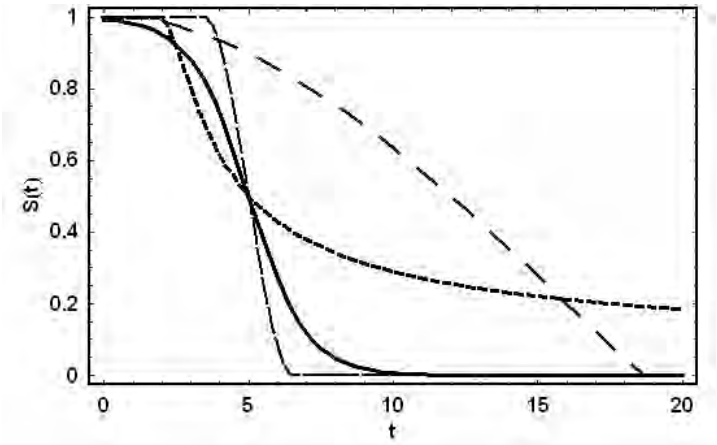

Figure 1: *Survival functions based on GS-distributions: GSD[0.5, 5.0, 1.0, 1.0, 1.0, 1.0] (continuous line), GSD[0.5, 5.0, 1.0, 0.5, 1.0, 3.0] (short dashed line), GSD[0.5, 5.0, 1.0, 0.5, 1.0, 0.5] (medium dashed line), GSD[0.5, 12.0, 0.1, 0.4, 1.0, 0.1] (long dashed line).*

According to this definition, the survival function $S(t)$ based in a GS-distribution is the solution of the differential equation:

$$\frac{dS}{dt} = -\frac{dF}{dt} = -\alpha F^g (1 - F^k)^\gamma \quad S(t_0) = 1 - F(t_0) \tag{4}$$

Some examples of GS-distribution based survival functions are shown in Figure 1.

Using this model, the hazard function can be written as:

$$h(t) = \frac{f(t)}{S(t)} = \alpha F^g (1 - F^k)^\gamma (1 - F)^{-1} \tag{5}$$

The survival function and the corresponding hazard function, for a set of parameters $(F_0, t_0, \alpha, g, k, \gamma)$, will correspond to the solution of the following set of equations:

$$\begin{aligned}
\frac{dh}{dt} &= \frac{F^{2g-1}\left(1-F^k\right)^{2\gamma-1}\alpha^2\left(F\left(1-F^k\right)+(1-F)\left((g+k\gamma)F^k+g\right)\right)}{(1-F)^2} && h(t_0) = f(t_0)/S(t_0) \\
\frac{dS}{dt} &= -\alpha F^g (1 - F^k)^\gamma && S(t_0) = 1 - F(t_0) \\
\frac{dF}{dt} &= \alpha F^g (1 - F^k)^\gamma && F(t_0) = F_0
\end{aligned} \tag{6}$$

The hazard functions corresponding to the survival functions in Figure 1 are shown in Figure 2.

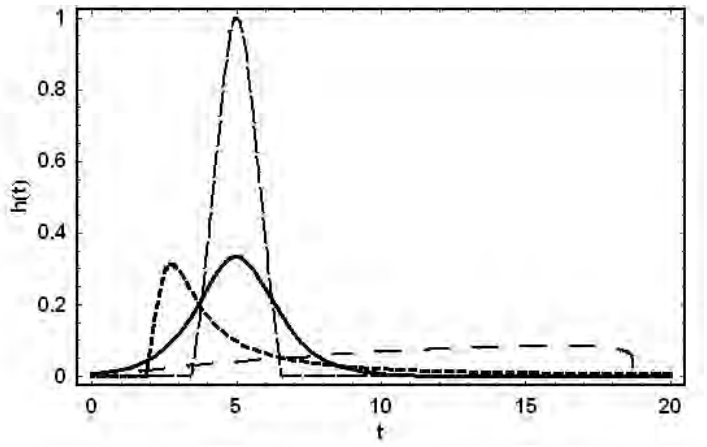

Figure 2: *Hazard functions corresponding to the survival functions in Figure 1. Lines follow the same scheme as in that figure.*

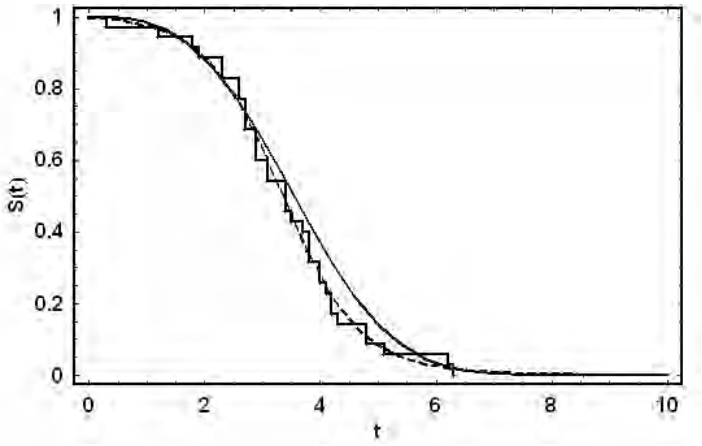

Figure 3: *Fitting a GS-distribution to uncensored survival data. Data is generated from a Weibull model with shape parameter 3 and scale parameter 4. The original model corresponds to the continuous line. The fitted GS-distribution model is indicated with a dashed line. The step function corresponds to the Kaplan-Meier estimator.*

2.1 Fitting the GS-survival function

The structure of the observations is a fundamental issue for parameter estimation in survival analysis. As censoring and truncation impose special conditions, we shall begin with uncensored data to discuss the basics of maximum-likelihood estimation using the GS-distribution survival model. Let us suppose that we have a set of observations $t_1, t_2, \ldots, t_n$ where t_i indicates the survival time for a given individual, i.e. the exact time at which the event of interest is observed. For this situation, the likelihood function is simply:

$$L(0.5, t_0, \alpha, g, k, \gamma) \propto \prod_{i=1}^{n} f(t_i) = \prod_{i=1}^{n} \alpha F(t_i)^g (1 - F(t_i)^k)^\gamma \tag{7}$$

To compute the likelihood for a given set of parameters, we shall follow a numerical procedure. First, we compute the $F(t_i)$ for a given survival time t_i. This requires integration of the GS-distribution equation from t_0 to t_i. Then, once all the $F(t_i)$ values are obtained, we calculate the likelihood using equation (7). Maximum-likelihood estimation will be performed by numerically finding the set of parameters that maximize equation (7). This requires an iterative computation of this equation until a maximum is reached. Although we have implemented this procedure in Mathematica, any alternative program could be used. At this point, we are developing **R** routines for using this model.

As an example, we generate survival times from a Weibull model and fit the data using the procedure discussed above (Figure 3). The corresponding Kaplan-Meier

estimation of the survival curve is also computed for comparison. It can be seen that the model provides an appropriate representation for the data. One advantage is that we obtain a parametric model that can be used for other purposes. For instance, using the quantile function for a GS-distribution (Muiño *et al.*, 2006) we can generate random survival times.

2.2 Using the GS-distribution as a model for survival data generation

The GS-distribution can be used as a model for generating survival data. First, we recall that the quantile x_q that is the solution of the equation $F(x_q) = q$ for a GS-distribution can be computed as (Muiño *et al.* 2006):

$$x_q = x_0 + \frac{B_{t_0^k, q^k}\left(\frac{1-g}{k}, 1-\gamma\right)}{\alpha\, k} \tag{8}$$

where B is the incomplete Beta function. According to this result, we can use this expression for obtaining a sample of values of the GS-distribution, which would correspond to a sample of survival times for the corresponding model. First, we generate a sample of a uniform distribution between 0 and 1. Then, we use those numbers as values of q in (refquantile) to obtain a sample of the required GS-distribution. This may be useful in testing some procedures in survival analysis. As an example, we generate a random sample of survival data from a GSD[0.5, 10, .3, .6, 1, 1] (n = 50) (Figure 4). Censored data can be obtained using a similar procedure and taking into account the corresponding censoring scheme.

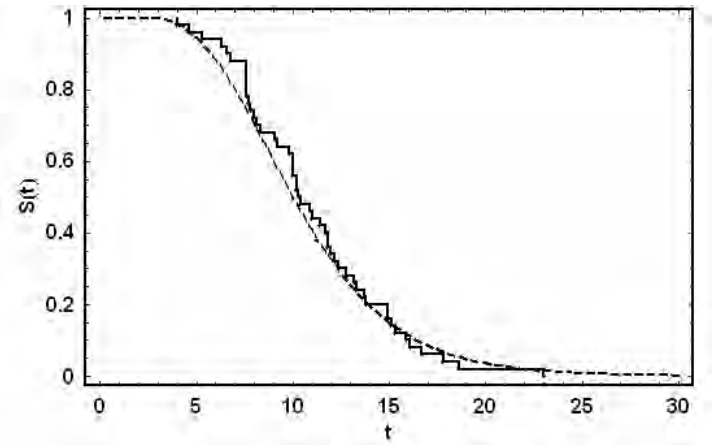

Figure 4: *Uncensored random sample from a survival model. The original model (dashed line) corresponds to a GSD[0.5, 10.0, 0.3, 0.6, 1.0, 1.0]. The step function corresponds to the Kaplan-Meier estimator from the generated data.*

2.3 Fitting S-survival functions on different censoring schemes

As the GS-distribution provides a survival model that is general for the unimodal continuous case, we can obtain the likelihood for any censoring scheme. According to the discussion above, computation of the likelihood requires a numerical procedure in all cases.

2.3.1 Right censoring

Suppose we have uncensored (U) observations and right-censored (R) observations. For the R cases, the contribution to the likelihood, in the general case of non-informative censoring, corresponds to terms $P(T \geq t_i) = S(t_i) = 1 - F(t_i)$. Thus, the likelihood is:

$$L(0.5, t_0, \alpha, g, k, \gamma) \propto \left(\prod_{i \in U} f(t_i)\right)\left(\prod_{i \in R} S(t_i)\right) = \left(\prod_{i \in U} f(t_i)\right)\left(\prod_{i \in R}(1 - F(t_i))\right) \qquad (9)$$

This computation can be made with a slight modification of the procedure for the uncensored case. Here, for the censored observations, we just need to compute $F(t_i)$ as indicated in the previous section. As an example, we generate data from a Weibull model. We produce censored observations and fit a GS-distribution model and obtain the Kaplan-Meier estimator. As we can see in Figure 5 the GS-distribution produces an appropriate estimation also in this case.

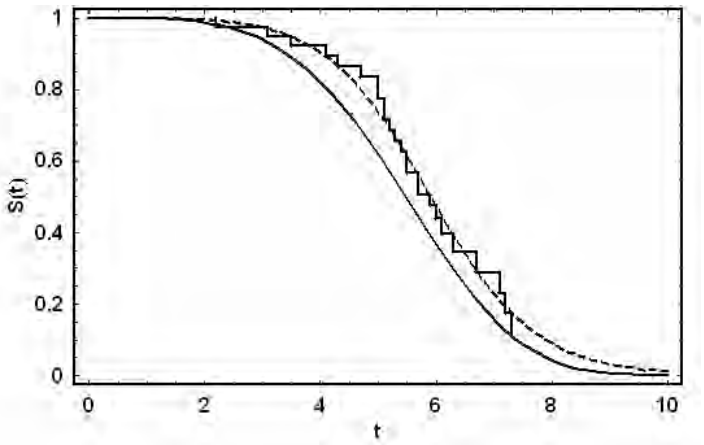

Figure 5: *Fitting a GS-distribution survival model to right-censored data. The original model is indicated by a continuous line. The fitted GS-distribution model is indicated by a dashed line. The step line corresponds to the Kaplan-Meier estimator.*

2.3.2 Interval censoring

A given observation is interval censored if it is known to have happened between two time points L_i and R_i. In that case, the contribution to the likelihood is $P(L_i \leq T \leq R_i) = F(R_i) - F(L_i) = S(L_i) - S(R_i)$. If we indicate by I the set of interval censored observations, the likelihood can be written as:

$$L(0.5, t_0, \alpha, g, k, \gamma) \propto \left(\prod_U f(t_i)\right)\left(\prod_{i \in R} S(t_i)\right)\left(\prod_{i \in I}(S(L_i) - S(R_i))\right) \qquad (10)$$

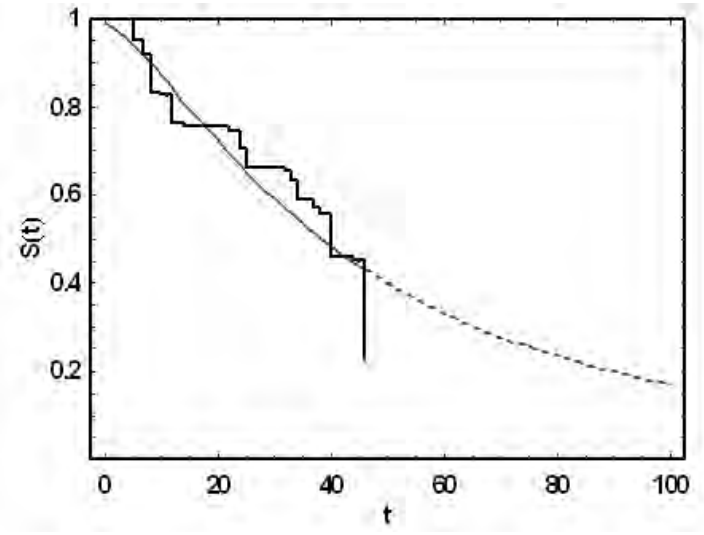

Figure 6: *Fitting a GS-distribution survival model to interval-censored data corresponding to the example 5.2 in Klein and Moeschberger (1997). The fitted GS-distribution model is indicated by a continuous and a dashed line. The step line corresponds to the estimation provided by Klein and Moeschberger (1997). See text for discussion.*

As in previous cases, the required values of $S(t)$ can be computed numerically using the GS-distribution to obtain the likelihood. Maximization of L would provide the best GS-distribution survival model according to the data. Following this procedure, we can accommodate other censoring schemes. As an example of this procedure, we have used a set of data discussed by Klein & Moeschberger (1997, example 5.2). These data refers to breast cancer patients that follow a radiotherapy treatment. The time to cosmetic deterioration (in months) is collected. For many patients, deterioration appears between two visits. For instance, and given patient present cosmetic deterioration in its visit corresponding to 11 months after therapy. In the previous visit at 4 months no deterioration was observed. Thus, for this patient the event has occurred within $(4, 11]$ months. This kind of data corresponds to an interval-censoring scheme. In Figure 6 we show the results of fitting a GS-distribution survival model. For comparison, the

estimation provided by Klein and Moeschberger is also provided. The GS-distribution procedure provides an estimation that allows extrapolating the results to times above the last observed time. Although this prediction must be considered with care, it seems reasonable according to the available data.

3 Hypothesis tests using GS-distribution survival models

3.1 Testing a given survival model

As the GS-distribution can accurately approximate unimodal continuous distributions, we can use it for defining a test that evaluates if the observed data follows a given model. This test can be defined as follows. First, a GS-distribution representation is obtained for the model to test (see Muiño *et al.* (2006) for a discussion on obtaining appropriate GS-distributions for a given statistical distribution). This model will be considered as the model under H_0. Second, we fit a unconstrained GS-distribution to the data. This is considered the estimated survival model under H_1. Then, a likelihood ratio test is applied to evaluate the hypothesis that the reference model is appropriate for the observed data. As an example, consider that we want to test if our data follow a Weibull model with shape parameter 4 and scale parameter 3. This distribution corresponds to a GS-distribution with parameters GSD[0.5, 2.74, 2.66, 0.92, 0.38, 0.7]. In that case, the corresponding hypotheses are:

$$\begin{aligned} H_0 &: S(t) \text{ follows a model GSD[0.5, 2.74, 2.66, 0.92, 0.38, 0.7]} \\ H_1 &: S(t) \text{ follows a model GSD[0.5,x}_0,\alpha,\text{g,k,}\gamma\text{]} \end{aligned}$$

To test these hypotheses, we compute the resulting likelihood $L(\omega)$ under H_0 using a GSD[0.5, 2.74, 2.66, 0.92, 0.38, 0.7] as a model. Then we fit a GSD[0.5,x$_0$,α,g,k,γ] and obtain the likelihood under H_1, i.e. $L(\Omega)$. The likelihood ratio test is:

$$\Lambda = \frac{L(\omega)}{L(\Omega)} \tag{11}$$

The statistic $U = -2ln(\Lambda)$ follows a χ^2 distribution with 5 degrees of freedom. As an example, we generate a random sample of 35 survival times from a Weibull distribution with shape parameter 4 and scale parameter 3. The obtained sample is:

3.9	3.7	3.3	2.2	1.1	2.6	2.3
2.6	2.2	3.7	3.8	2.3	2.4	2.5
2.9	3.1	2.9	2.5	3.8	2.2	1.0
4.3	2.3	2.3	1.6	1.6	2.8	1.4
1.5	3.1	2.3	2.9	3.4	3.3	2.4

The log likelihood corresponding to a GSD([0.5, 2.74, 2.66, 0.92, 0.38, 0.7] is equal to −42.1479. Fitting an unconstrained GSD[0.5,x$_0$,α,g,k,γ] gives a log likelihood of −41.23. With that, $U = -2ln(\Lambda) = 1.84$. With 5 degrees of freedom the significance of this result is 0.643, which leads to not reject H_0. Other hypotheses can be tested following the same procedure.

For example, we can test some specific restrictions on the shape of the survival function. For instance, we can define a test for evaluating if the data follows a symmetric survival model. In that case, if we recall that a symmetric GS-distribution corresponds to the case $g = \gamma$ and $k = 1$, we have

$$\begin{aligned} H_0 &: S(t) \text{ follows a model GSD[0.5,x}_0\text{,}\alpha\text{,g,1,g]} \\ H_1 &: S(t) \text{ follows a model GSD[0.5,x}_0\text{,}\alpha\text{,g,k,}\gamma\text{]} \end{aligned} \tag{12}$$

In that case, the resulting statistic $U = -2ln(\Lambda)$ follows a χ^2 distribution with 2 degrees of freedom.

3.2 Comparing GS-survival curves

Since we have a parametric model for the survival function, we can compare two groups using the likelihood-ratio test. In general, we want to test if the survival function in one group $S^a(t)$ is different than the survival function in another $S^b(t)$. This can be indicated as:

$$\begin{aligned} H_0 : S^a(t) &= S^b(t) \quad \forall t \\ H_1 : S^a(t) &\neq S^b(t) \quad \text{for some } t \end{aligned} \tag{13}$$

Under H_0 the likelihood $L(\omega)$ can be obtained by fitting a GS-distribution model with common parameters for both groups:

$$L(\omega) = L^a_{0.5,t_0,\alpha,g,k,\gamma} L^b_{0.5,t_0,\alpha,g,k,\gamma} \tag{14}$$

Under H_1 the likelihood $L(\Omega)$ can be obtained by fitting a GS-distribution to each of both groups:

$$L(\Omega) = L^a_{0.5,t_{0_a},\alpha_a,g_a,k_a,\gamma_a} L^b_{t_{0_b},\alpha_b,g_b,k_b,\gamma_b} \tag{15}$$

In that case, we estimate 5 parameters under H_0 and 10 parameters under H_1. Accordingly, the statistic $U = -2ln(\Lambda)$ follows a χ^2 distribution with 5 degrees of freedom. This test is an alternative to the log-rank test and can be used with any censoring scheme for which an estimation of the likelihood can be obtained.

It is important to indicate that this test can be used independently of the censoring scheme. In the previous sections we have shown that the likelihood can be numerically obtained for any of such schemes, which provides the required information for hypothesis testing through the likelihood ratio test. Although more research is required to test the performance of this procedure, our preliminary results indicate that hypothesis testing based on GS-distributions may be a practical alternative to existing non-parametric methods.

3.3 An application to clinical data

Progression rate to the acquired immunodeficiency syndrome (AIDS) in HIV-1 infected patients is caused by a combination of environmental and genetic risk factors. The most significant association for host risk factor has been attributed to the C-C chemokine (CCR) receptor 5 (CCR5) gene variant Δ32. Homozygous individuals for a 32bp deletion in the CCR5 coding region (*CCR*5Δ32, allele) do not synthesize CCR5 protein. Those individuals are resistant to infection by HIV-1 macrophage tropic strains that use CCR5 as coreceptor. In addition, a slow progression rate to AIDS has been attributed to heterozygous patients (Huang, 1996). We have studied the role of *CCR*5Δ32 genotype status on disease progression rate to AIDS in a seroprevalent cohort of HIV-1 infected patients belonging to the injection drug use (IDU) risk group.

Given the natural history of HIV infection, HIV-1 seropositive patients could have been infected a few months to several years before their seropositivity was recorded. To reduce the lack of precision on seroconversion date, only patients with a first HIV positive test prior to 31 December 1989 were selected (137 patients). In addition, patients with an age lower than 20 years at the first HIV positive test were also included (42 patients). For patients infected prior to 31 December 1989, seroconversion date was estimated as the mean time between the date at the first HIV positive test and 1 January 1981, which correspond to the earliest year for the first AIDS case reported in Catalonia (Vilaseca, 1982). In addition, for patients with age lower than 20 years old in their first HIV positive test, seroconversion date was estimated as the mean time between the date of the first HIV positive test and the date at age 15 years. According to HIV epidemiological data from the Centro Nacional de Epidemiología in Spain, in the 1981-2004 period only 3 AIDS cases have been reported from IDU individuals with age less than 15 years old (Centro Nacional de Epidemiología, 2005). HIV-1 disease progression was analyzed according to Centers for Disease Control and Prevention (CDC) 1993 criteria (Centers for Disease Control and Prevention, 1993). The resulting data were analyzed using the Kaplan-Meier estimator and no significant differences were found among the genotypes. In Figure 7, we show the estimated survival function for both genotypes obtained by the GS-distribution method (the corresponding Kaplan-Meir estimation is not shown for clarity). The result of the corresponding test was similar to

the Kaplan-Meier method. As this result is obtained by guessing the infection time, one may consider other alternatives.

For instance, instead of taking the mean time as indicated before, we can consider that the time to developing AIDS is greater than t_i, i.e. the time between the first seropositive test and the moment they develop AIDS, and less that $t_i + T_i$, where T_i is the time from first reasonable date for infection, as computed above, to the day of the first seropositivity. Thus, independently of the final outcome of AIDS, all the observations will contribute to the likelihood by a factor of $P(t_i \leq T \leq t_i + T_i)$. This procedure is less restrictive than the method of computing the mean of the interval. In Figure 7 we present the results of estimating the time to AIDS following this strategy. We can see that the estimations obtained by this new strategy is different than the results corresponding for the first method. The procedure based on interval censored data estimates a shorter time to developing AIDS than the results computed from the mean method of estimating the time of infection. In both cases, the differences are not significant between both genotypes. In that case, independently of the method used, we can conclude that among IDU users, the CCR5 genotype is not likely to produce a different rate of progression to clinical AIDS.

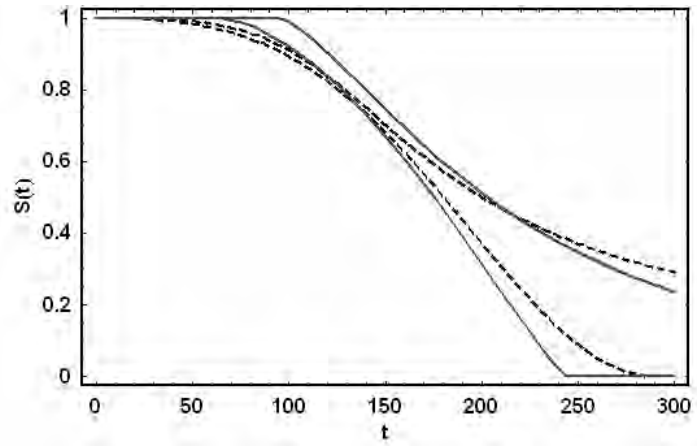

Figure 7: *GS-distributions fitted to data of HIV infected patients. The dashed line indicates the estimated survival functions for the wild type CCR5 gentoptype (n = 151). The continuous line indicates the estimation for the mutant (n = 27) genotype. The steepest curves correspond to the estimation using interval censoring. The other curves correspond to the estimation using the mean method for computing the time of infection. See text for details.*

These results show the interest of further analyzing the influence of the method used for obtaining an estimation of the date of infection. Clearly, considering an interval censored approach may be less subjective as we only establish a reasonable interval in which the event has occurred. The usual method of taken the mean of this interval as the time of infection is somehow more subjective. A careful study of the influence of this choice may shed new light on the problem of progression rate to AIDS. The GS-

distribution approach introduced in this paper would provide a valuable tool for this analysis.

4 Conclusions

We have introduced the GS-distribution as a parametric model for univariate survival analysis. We have shown its utility in simulation examples and we have developed the basis for hypothesis testing using this approach. The use of the GS-distribution as a general model allows for obtaining a parametric representation for survival data independently of the censoring scheme. Once obtained, we can use this model for further exploring the survival process. The tools introduced by Muiño *et al.* (2006) for the GS-distribution can be used for this purpose.

Our results show that the GS-distribution survival function can be used to easily obtain estimates of the survival function in different censoring schemes. Group comparison is straightforward through the likelihood ratio test. Although a systematic study is yet to be performed, our preliminary results indicate that the GS-distribution based method can be an interesting alternative to non-parametric methods. Using this approach, we could even compare two groups when the samples obey different censoring schemes. As a drawback, the extension of this model to cope with covariables is not straightforward. We shall explore practical solutions for this problem in the near future.

References

Centro Nacional de Epidemiología. (2005). *Vigilancia Epidemiológica del SIDA en España. Registro nacional de casos de SIDA*. Available at: http://193.146.50.130/htdocs/sida/sidavih.htm

Centers for Disease Control and Prevention. (1993). From the Centers for Disease Control and prevention. 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *Jama*, 269, 460.

Huang, Y., Paxton, W.A., Wolinsky, S.M., Neumann, A.U., Zhang, L., He, T., Kang, S., Ceradini, D., Jin, Z., Yazdanbakhsh, K., Kunstman, K., Erickson, D., Dragon, E., Landau, N.R., Phair, J., Ho, D.D. and Koup, R.A. (1996). The role of a mutant CCR5 allele in HIV-1 transmission and disease progression. *Nature Medicine*, 2, 1240-1243.

Hosmer, D.W.Jr., Lemeshow, S. (1999). *Applied Survival Analysis*. New York: John Wiley.

Hernández-Bermejo, B. and Sorribas, A. (2001). Analytical quantile solution for the S-distribution, random number generation and statistical data modeling. *Biometrical Journal*, 43, 1017-1025

Jones, M.C., (2002). The complementary beta distribution. *Journal of Statistical Planning and Inference*, 104, 329-337.

Jones, M.C., (2004). Families of distributions arising from distributions of order statistics. *Test*, 13, 1-43.

Kamps, U., (1991). A general recurrence relation for moments of order statistics in a class of probability distributions and characterizations. *Metrika*, 38, 215-225.

Klein, K.P., Moeschberger, M.L., (1997). *Survival Analysis*. New York: Springer.

March, J., Trujillano, J, Tort,M. and Sorribas, A. (2003). Estimating conditional distributions using a method based on S-distributions: Reference percentile curves for body mass index in Spanish Children *Growth, Development, and Aging*, **67**, 59-72.

Muiño, J.M., Voit, E.O. and Sorribas, A. (2006). GS-distributions: A new family of distributions for continuous unimodal variables. *Computational Statistics & Data Analysis*, 50, 2769-2798.

Parzen, E., (1979). Nonparametric statistical data modelling (with comments). *Journal of the American Statistical Association*, 74, 105-131.

Sorribas, A., March, J. and Trujillano, J. (2002). A new parametric mtehod based on S-distributions for computing Receiver Operating Characteristic curves for continuous diagnostic tests. *Statistics in Medicine*, 21, 1215-1235.

Sorribas, A., March, J. and Voit, E. O. (2000). Estimating age-related trends in cross-sectional studies using S-distributions *Statistics in Medicine*, 19, 697-713.

Turner, M.E. and Pruitt, K.M. (1978). A common basis for survival, growth, and autocatalysis *Mathematical Biosciences*, 39, 113-123.

Vilaseca, J., Arnau, J.M., Bacardi, R., Mieras, C., Serrano, A. and Navarro, C. (1982). Kaposi's sarcoma and toxoplasma gondii brain abscess in a Spanish homosexual. *Lancet*, 1, 572.

Voit, E.O. (1992). The S-distribution: A tool for approximation and classification of univariate, unimodal probability distributions. *Biometrical Journal*, 7, 855-878.

Voit, E.O. (2000). A maximum likelihood estimator for shape parameters of S-distributions. *Biometrical Journal*, 42, 471-479.

Voit, E.O. and Schwacke,L.. (1998). Scalability properties of the S-distribution, *Biometrical Journal*, 40, 665-684.

Voit, E.O. and Schawcke, L.H. (2000). Random number generation from right-skewed, symmetric and left-skewed distributions. *Risk Analysis*, 20, 59-71.

Voit, E.O. and Sorribas, A. (2000). Computer modeling of dynamically changing distributions of random variables *Mathematical and Computer Modelling*, 31, 217-225.

Voit, E.O., Balthis, W.L. and Holser, R.A. (1995). Hierarchical Monte-Carlo modeling with S-distributions: concepts and illustrative analysis of mercury contamination in king mackerel. *Environmental International*, 21, 627-635.

Yu, S.Y. and Voit E.O. (1995). A simple, flexible failure model. *Biometrical Journal*, 37, 595-609.

Zhao, Q. and Sun, J. (2004). Generalized log-rank test for mixed interval-censored failure time data. *Statistics in Medicine*, 23, 1621-1629.

Book reviews

MULTIPLE CORRESPONDENCE ANALYSIS AND RELATED METHODS

Edited by Michael Greenacre and Jörg Blasius

Chapman & Hall/CRC, 2006.

This book gathers selected papers of the CARME 2003 conference held at Pompeu Fabra University on the topic of multiple correspondence analysis (MCA), as well as several chapters written specially to make the book self-contained. At present several books exist on correspondence analysis (CA) but none is devoted specifically to MCA, in this sense this book fills a gap in the scientific series of books. Professors Greenacre and Blasius should be thanked for their efforts to publish it in a very accurate and coherent way.

MCA is a more difficult topic to grasp than simple CA, hence it is often applied in a naïve manner, taking the default value of its parameters, without taking advantage of all its possibilities. MCA inherits from its ancestor, simple CA, its roots and its multiple facets. Different fields and hence difference schools have led to CA, ecology, psychometrics, linguistics, ... each one having its own contribution and particular point of view, one focusing on the quantification of categorical variables, another in finding the correlation between categorical variables, other in finding a latent variable among categorical variables, for others it is just a model for categorical data. In my opinion, however, the main strength of MCA is the visual representation of the information as popularized by the French "*Analyse des Données*" school and this has also been the main perspective of this book.

The book is very well written, it is composed of 23 chapters assembled in five parts, the first two being devoted to MCA, whereas the last three are about methods that are related in a general way. This is unavoidable in a book composed of contributions from a conference. In spite of that, the large majority of chapters present the state of the art of the methods in a very clear manner, using good and motivating examples that make enjoyable reading.

From a mathematical point of view MCA is a generalisation to categorical variables of the canonical correlation analysis, and this is the starting point of the book, but it goes too short in this path and moves to the geometric presentation as a particular case of PCA, which is the classic way of presenting CA. In this approach MCA consists of

the PCA of the triplet (**Y**,**M**,**N**), where **Y** represents a convenient matrix of profiles, **M** a chi-square metric and **N** is the diagonal matrix of relative weights. Then, it is easy to produce the desired visualisations and by modifying the metric to enlarge the possibilities of MCA. However, under this approach the introduction of the chi-square metric is rather arbitrary and so are the different possible visualisations.

In addition, it shows some drawbacks concerning the unexpected high importance of rare modalities and the problematic interpretation of the explained variance. All of this makes MCA a little messy and this gives to the book all its value.

CA is not a particular case of PCA indeed. CA deals with the analysis of two way tables and not with the analysis of data matrices, as PCA does, as John Gower points out very cleverly in Chapter 3. This is why CA needs to be presented as a double PCA, one for the row profiles and the other for the column profiles. However, the criticism of the chi-square metric has no justification; the more correct way of presenting CA is like a canonical correlation analysis of two categorical variables. Under this formalism, the chi-square metric emerges naturally as the categorical counterpart of the Mahalonobis metric for continuous data, well accepted to measure distances with multinormal data. Of course, it does not imply that other metrics can be used as in the case of continuous data, leading to Tucker analysis, taking the identity as metric for the row and column spaces, redundancy analysis, SIMPLS, (Tenenhaus, 1998). Nevertheless, the central role of the chi-square metric is clear, like the Mahalanobis one for continuous data.

Concerning the different possibilities for the representation of points, we know the infinity of different biplots available but only two have meaningful value, the row preserving or the column preserving biplot (the asymmetric visualisations), nevertheless the most used visualisation for CA and MCA is the symmetric one, which is not a biplot. In fact all these visualisations stem again from the CCA. In CCA, two representations are equally possible, one in the space of the first variable and the other in the space of the second variable, which correspond to the two asymmetric biplots. These two representations are not unique. It is possible to perform a representation in the direct sum space of the two variables, this representation holds the property (interesting) of maximizing the variance represented, yielding up to the so-called symmetric representation (Saporta, 1990).

All these representations have clear interpretation rules that give the user deep insight to pass from data to knowledge. This is what happens when reading the introduction to MCA with the ISSP (International Social Survey Program) data as an example in Chapters 1 and 2, starting from the simple case of a simple table, then going on to the stacking of tables and finally the analysis of all two-way tables, which is the situation of MCA, showing very clearly the interest and the differences between MCA and the CA of stacked tables.

Another important point for interpretative purposes is the adjustment of the inertia explained for each dimension, proposed by Greenacre in Chapter 2 – its simplicity merits it being routinely computed in the standard packages of MCA. Alternatively JCA (joint correspondence analysis) can be used, which by means of an iterative process annihilates the inertia of the block diagonal tables, like the method of principal factor analysis.

An endless debate is the link between modelling and visualisation. Chapter 3 makes bridges between both approaches, by bilinear models modelling the residuals with respect to some baseline model or by the visual approach using the SVD of the same residuals. Also here it is important to notice the difference between dealing with data matrices (two dimensions and two modes) and two-way tables (two dimensions but one mode) and the formal similarities between continuous and categorical variables, which leads John Gower to say that MCA is more related to PCA than to CA, and this is true. Saporta (1990) already presented MCA as a non-linear generalisation of PCA (PCA of an expansion of the variables space by splines of order 0). Also it is clever to point out that PCA and MCA yield up an approximation of matrix of residuals **Y**, and not the correlation matrix **R** (which is the case for factor analysis). But the difference is very slightly, since both PCA and MCA are used to define latent variables from the observed ones (this is the explicit goal of homogeneity analysis and it is one of the foundations of the PLS path modelling community). The topics raised in Chapter 3 have attracted the attention of many researchers; Escofier (1984) presented the MCA respect to any model with the same margins; within the SPAD community biplots were usual from 1980 even though they were not called biplots. In Chapter 21 models are enlarged by Kroonenberg and Anderson to cover three-way interactions with additive terms and more sophisticated models with multiplicative terms but with lack of interpretability in this case. In Chapter 22 Groenen and Koning present a biplot for the visualisation of the interaction of an ANOVA model and in Chapter 23 Vicente-Villardon, Galindo and Blazquez visualize a logistic response via biplots.

Accepting that MCA is a non-linear analysis of data, it makes sense to constrain the ordering of the categories (for instance the Likert scale) to present the results in a meaningful way for the user. Chapter 4 presents it in a very general way, leading to non-linear PCA (NLPCA), also known as categorical PCA (CatPCA). In fact it is arguable whether a MCA would be preferable to CatPCA, as Nishisato says in Chapter 6, it depends on the application but in general it would be better to reveal all the non-linearities present in the data rather than to filter them. In fact CatPCA is a good alternative to PCA when analyzing ordinal data. Very often ordinal data is analyzed by PCA as if the variables were continuous, assuming equally sequenced scales for them. Then CatPCA and its comparison with standard PCA allow assessing whether the respondents have understood the questions and what subjective scales they have applied, as is shown in Chapter 20 by Blasius and Thiessen.

In fact, MCA is one of the most successful statistical techniques to analyze survey data, because MCA fits well the requisites for survey data: it can be applied to large data sets, very often collected with mild probabilistic assumptions, it reveals the salient parts of the information by looking at the multivariate distribution defined from a homogeneous group of variables (respect to the concept they measure, for instance opinions about a group of questions) which is called active, extracting their common denominators (the significant axes) and relating them to the external (supplementary) information, which very often forms a set of structuring factors (sex, age, level of studies, region, ...) leading to what Alain Morineau has called the "themascope" approach (Aluja and Morineau, 1999). This is the heart of the French "*Analyse des Données*" which is explained by Henry Rouanet in Chapter 5. This approach can be enhanced by incorporating inferential aspects like the confidence ellipses around the centroids of the categories of the structuring factors to assess its significance.

Ludovic Lebart goes deeply into this argument in Chapter 7, introducing validation tools of the revealed patterns. MCA gains all its value when applied to large data sets, where it is possible to rely on the central limit theorem and to compute t-test values for all centroids in every significant dimension. This allows linking the revealed pattern with the background information of individuals in a substantive way. In addition we can visualize the uncertainty inherent in the position of these centroids to compare pairs of them, by bootstrapping, to obtain confidence ellipses or convex hulls for supplementary category points. Another possibility for stabilizing the results is by regularized MCA (presented by Takane in Chapter 11), by means of a change of the metric of the column space in a similar way as in ridge regression; this leads to more stable estimates of the factorial axes and hence more reliable confidence ellipses, at a price of the greater complexity of the method.

Another problem when analyzing survey data is that of missing values, Figure 8.1 presents a typical display of its effect. Although it is arguable whether or not to eliminate these "don't know" points in the displays, since they do reveal some information about the individuals who had chosen this option, especially if we have socio-demographic supplementary points in the display, it is nevertheless interesting to focus the attention only on certain categories, for instance the expressed opinions or the most extreme ones, to get insight into the understanding of data. In Chapter 8 Greenacre and Pardo propose subset MCA as a clever and useful solution to this problem; it consists of performing MCA in a selection of categories with the same global margins to not lose the additive properties of each subset MCA. This topic was also addressed by Bénali and Escofier (1987), also it has been incorporated in the procedure COREM of SPAD. This problem of missing data is also treated in Chapter 12 by Matschinger and Angermeyer as a particular case of canonical correlation analysis but the rationale is more complicated than that of subset MCA.

We have stated that MCA is a correlational technique, however, what is the correlation between categorical variables is not an easy question to answer. The practice of just taking the correlation with the first solution of the MCA is obviously a partial solution. Nishishato in Chapter 6 proposes to measure the correlation between two variables in the space spanned by one of them, leading once again to the analysis of stacked tables. This measure turns out to be related to Cramer's V coefficient.

Another usage of MCA is to measure a latent variable, with maximal (squared) correlation with every question. Item response theory (IRT) serves to define scales measuring a latent unidimensional trait from a set of items. There are two main models for IRT, the dominance model, which implies a monotonic utility of the latent variable and the proximity model, implying a unimodal pattern, where the utility depends on the distance of one individual respect to the mode point. The Guttman model and the Rasch model are cases of the former type, whereas the unfolding model is an example of the latter. Chapters 9 and 10 (by Warrens and Heiser and van Schuur and Blasius respectively) deal with how MCA can be helpful to ascertain which type of data, dominance or proximity, we have at hand. The arch effect and the horseshoe are basic displays for these kinds of data. Here it is worth pointing out what van Schuur and Blasius state about the contradiction regarding the need of having simple questions in questionnaires and the difficulty of asking questions necessary to obtain unfolding data.

Sometimes we want to analyze a concatenation of tables coming from different studies, (like different countries, years, ...). This analysis will include an inertia effect between tables due to the different centroids of each table. Also it is required to give the same importance to all tables. Bécue and Escofier in Chapter 13 and Amaya and Goitisolo in Chapter 14 present similar methodologies to solve this problem (MFACT and SA respectively). First, each table is centred with respect to its own centroid to eliminate the between inertia (intra-analysis) and second the influence of each table is equalized respect its corresponding first dimension, that is, we operate a change of the metric in the space of columns (but other methods for balancing are possible). The two approaches differ in how they define the weight for the rows, which coming from different surveys, will be different in general and hence there exists no natural weight for each row. These kinds of approaches lean on the geometric facilities of the PCA. In Chapter 15 Abascal, Lautre and Landaluce propose a variant of the previous MFA method to treat mixed tables formed by categorical variables and continuous variables together. The obtained results are more difficult to interpret but this can be an alternative to CatPCA when we know *a priori* which variables have a non-linear effect.

The discrimination ability is another facet of CA. Very often we have to predict a categorical variable from a set of categorical predictors. CA provides a simple solution consisting of stacking the tables crossing the response variable with all the predictors (or interactions between the predictors). It stems from the method's very origins (Fisher, 1940) and other variants have been proposed such as PLS discrimination (Tenenhaus,

1998) or nonsymmetric CA (Lauro and d'Ambra, 1984). In Chapter 16 Saporta and Niang present the Disqual methodology, consisting of a MCA of the predictors followed by a linear discriminant analysis using the significant factors – in this sense Disqual acts like a principal component regression on categorical data. An important result is the reduction of the Vapnik-Cervonenkis dimension due to the reduction of dimensionality entailed by the use of MCA. Conjoint analysis can enter within the same framework of CA of the response variable (expressing a preference) crossed by the set of attributes of products, because very often conjoint analysis uses least-squares regression instead of monotonic regression, as shown by Torres-Lacomba in Chapter 19.

In Chapter 17, Bougeard, Hanafi, Noçairi and Qannari propose a unified method to obtain factors balancing the discrimination power of CA of the stacked table crossing the response variable and the predictors and the self explanation of factors issued from the MCA of the predictors. Then, successive solutions are obtained by deflation, in a similar way to PLS discriminant analysis.

Changing the metric of rows it is possible to highlight other features rather than the dispersion of individuals, in particular if we consider a metric downweighting the high distances to the global centroid, it is possible then to perform a robust MCA, to detect outliers, or choosing a metric to download the high inter-pair distances among individuals, it is possible to obtain a very flexible technique to reveal the natural clusters of the individuals in the display (in Chapter 18 by Caussinus and Ruiz-Gazen). A similar technique just retaining the distances lower to a given threshold has been proposed by Lebart (1995).

All these chapters emphasize the great versatility of MCA and all its possibilities; it can serve to analyze real problems in a large variety of different fields, social surveys, psychometry, marketing, ... and also to further develop the methodology to produce research papers, as shown in Figure 1.1. However, I would stress the need for having clear interpretation rules for a method to be useful. There is no a intelligent method, what is intelligent is the use of the method. This book serves the purpose of making the method useful.

Tomàs Aluja-Banet

Universitat Politècnica de Catalunya

References

Aluja, T. and Morineau, A. (1999). *Aprender de los Datos: el Análisis de Componentes Principales*. Barcelona: EUB.

Bénali, H. and Escofier, B. (1987). Stabilité de l'Analyse Factorielle des Correspondances Multiples en cas de données manquantes et de modalités à faibles effectifs. *Revue de Statistique Appliquée*, 35, 41-51.

Escofier, B. (1984). Analyse factorielle en référence à un modèle : application à l'analyse d'un tableau d'échanges. *Revue de Statistique Appliquée*, 32, 25-36.

Lauro, C. and d'Ambra, L. (1984). L'analyse non-symétrique des correspondances. In *Data Analysis and Informatics III*, E. Diday ed., 433-446. Amsterdam: North Holland.

Lebart, L., Morineau, A., and Piron, N. (1995). *Statistique Exploratoire Multidimensionelles*. Paris: Dunod.

Saporta, G. (1990). *Probabilités, Analyse des Données et Statistique*. Paris: Technip, (2006, 2nd. edition).

Tenenhaus, M. (1998). *La Régression PLS*. Paris: Technip.

CORRESPONDENCE ANALYSIS AND DATA CODING WITH JAVA AND R

Fionn Murtagh

Chapman & Hall/CRC, 2005.

MULTIDIMENSIONAL NONLINEAR DATA ANALYSIS

Shizuhiko Nishisato

Chapman & Hall/CRC, 2006.

Apart from the book on Multiple Correspondence Analysis, reviewed above, Chapman & Hall/CRC have published two other books recently that fall directly in the same area of Statistics, by Fionn Murtagh and Shizuhiko Nishisato respectively. While treating methods which rely on the same mathematical theory, these two books could not be more different, hence this comparative review.

Fionn Murtagh was a doctoral student of Jean-Paul Benzécri, and his book is imbued with Benzécri's teachings, ideas and philosophy. In my view this is the first English-language text that brings across the flavour of Benzécri's contributions to data analysis. Almost all statisticians would think of correspondence analysis (CA) as just another multivariate technique in their toolbox, which gives interesting visualizations of specific sets of categorical data. For Benzécri, however, CA is a core method of multivariate analysis, which can expose structure in any data that is suitably transformed, or *coded* – the method is really central to his thinking, as it was to the sociologist Pierre Bourdieu who popularized it amongst social scientists as a way to define social space. After 33 years of working in this field since my own doctoral studies with Benzécri in Paris, my opinion is that Benzécri is closer to being right than wrong: I often say to students that if they were stranded on a desert island with only one computer program, then it should be CA, because they could answer most of the questions they might have about a data set. The method provides a framework for investigating discrete and continuous structure, associations and relationships between variables.

Murtagh's book is packed with examples of the way data can be coded to feed into the CA algorithm, which leads to visualizations of the structure in the coded data.

Literally, what goes in comes out – this could be the motto of the book. After an historical introduction (Chapter 1), very much from the "Benzécrean" perspective, the mathematics of CA and cluster analysis is given. Here there is an unfortunate excursion (in my opinion), but just for 10 pages, into the over-complicated tensor notation used by Benzécri in his books, incomprehensible to most readers. This notation was something I had to learn to endure personally in my own studies in Paris and, in retrospect, I cannot understand why mathematical elegance should shroud what are basically simple mathematical concepts, much more easily understood using pragmatic matrix-vector notation. Anyway, fortunately for the readers, the rest of the book uses straightforward scalar notation and on some rare ocasions a matrix and a vector can be spotted (this is in strange contrast to the R program code throughout the book, which is necessarily constructed on data matrices and row and column vectors).

Chapter 3 is the core of the book, treating the various types of data coding, for example disjunctive coding (i.e., dummy variables), fuzzy coding (a less strict form of disjunctive coding, used for continuous-scale variables to conserve more information) and doubling (creating two variables for each original one, representing two extreme poles).

Chapter 4 includes five case studies, all of which involve data originally on continous scales – this demonstrates that this book is quite different from the the usual publications on CA.

Chapter 5 is exclusively devoted to longitudinal and textual analysis, which is fitting since all the early development of CA by Benzécri was in a linguistic context. As Murtagh says (page 161): "the way in which textual and document analysis is carried out in the correspondence analysis approach is quite different from other contemporary approaches". Linguists might not all agree, but at least here they will find this approach excellently presented and illustrated here.

One of the main features of the book is the support for practical application provided in the form of R code, which is also available on the author's website `www.correspondances.info`. In summary, this book is highly recommended as a text explaining Benzécri's ideas and giving many examples of the application of CA in non-standard situations. On the negative side, related work by other researchers is not even referred to, for example the work of Nishisato (e.g., Nishisato 1994), the Leiden group (e.g., Gifi 1990) and the two edited books by Greenacre & Blasius (1993) and Blasius & Greenacre (1998). The non-citing of Gifi (1990) is probably the most serious, since the work of Jan de Leeuw and co-workers is just as innovative in using the optimal scaling ideas inherent in CA as the basis for generalizing multivariate methods to categorical data, which fits very closely the general Benzécrean philosophy.

Shizuhiko Nishisato's book *Multidimensional Nonlinear Data Analysis* is so different from Murtagh's that, apart from the appearance of some maps of data with respect to component axes, one might think that this was a completely different subject. The first

sentence, however, tells us that “multidimensional nonlinear data analysis” (MUNDA) is a family of methods for quantifying categorical data” and that “this procedure covers such methods as correspondence analysis, dual scaling, homogeneity analysis, quantification theory, optimal scaling and the method of reciprocal averages.” This book looks like a second edition of the book on dual scaling by Nishisato (1994) but now dual scaling is apparently subsumed in MUNDA. Later on page 51 it is stated that “correspondence analysis (is) one of the many names of MUNDA”. This rather confusing classification serves to confirm that, apart from a few academic nuances, all the methods mentioned above are really the same, and differ mainly for historical-cultural reasons.

Putting this aspect aside, Nishisato does an excellent job of putting the historical record straight, right up to the present time (see Chapter 3-Historical Overview, 17 pages long). In reading the rest of the book I was interested to see that Nishisato has now recognized all the geometric concepts inherent in the correspondence analysis approach, although his description of chi-square distances will leave readers baffled as to what it really is. I could find no definition of this distance except for an erroneous set of formulas on page 79 which expressed it in terms of the solution “weights” (or coordinates) rather than on the original data.

Subsequent chapters treat different data types and how the method deals with them. Chapter 6 deals with the analysis of “incidence data”, alias simple correspondence analysis. Chapter 7 deals with the analysis of “multiple choice data”, alias multiple correspondence analysis. Chapter 8 deals with “sorting data”, which is analyzed just like “multiple choice data”.

Chapter 9 is on “forced classification of incidence data”, an idea which is inherent in the method known as canonical correspondence analysis (CCA). The “forcing” is the same as the “constraining” or “restricting” of the solution to be linearly related to an external set of continuous or dummy variables, which splits the space into constrained and unconstrained parts (Nishisato calls the unconstrained dimensions the “conditional components”). This methodology is routinely used by ecologists and published in R software such as the vegan package by Jari Oksanen. In the introduction Nishisato apologizes for not treating CCA in his book, which is a pity since CCA includes forced classification if the constraining variables are categorical.

Chapters 10 to 12 are on the analysis of “dominance data”, i.e. paired comparisons, rank-order data and successive categories (or ratings) data. The analysis of these types of data can be performed equivalently using the doubling coding prior to CA. This fact is not referred to explicitly by Nishisato, but he does say (page 76) that “researchers in correspondence analysis...have recently derived formulations of correspondence analysis for dominance data as well, thus diminishing the initial differences between dual scaling and correspondence analysis, except for some details.”

The main problem with Nishisato's book is that there is no reference to computing and how to perform the methods that he describes. Theory is explained but not the algorithms for finding solutions, neither is software referred to or commented on. The main benefit of this book is the extensive reference list and Nishisato's comprehensive treatment of the history of this area of multivariate analysis.

Michael Greenacre

Universitat Pompeu Fabra

References

Blasius, J. and Greenacre, M. J. (1998). *Visualization of Categorical Data*. Academic Press, San Diego.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Wiley, Chichester.

Greenacre, M. J. and Blasius, J. (1993). *Correspondence Analysis in the Social Sciences*. Academic Press, London.

Nishisato, S. (1994). *Elements of Dual Scaling: An Introduction to Practical Data Analysis*. Lawrence Erlbaum, New Jersey.

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.es) 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_ε.

In any case, upon request the journal secretary will provide authors with LaTeX 2_ε 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

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.es

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 2006 (volume 30)

Annual subscription rates:

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

Price for individual issues (current and back issues):

– Spain: €9/issue (VAT included)
– Other countries: €11/issue (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

Quatre modalitats de subscripció al DOGC

(Diari Oficial de la Generalitat de Catalunya)

Generalitat de Catalunya

Imprès, edició diària

Base de dades, actualització diària

DVD, edició semestral

A la carta, servei diari personalitzat

A més, per als subscriptors de l'edició impresa i del DVD, tramesa gratuïta d'un CD-ROM trimestral que conté les pàgines en format PDF (DOGC en imatges)

L'Administració més a prop

EADOP • Informació i subscripcions • Rocafort, 120 - Calàbria, 147 • 08015 Barcelona
Tel. 93.292.54.17 • Fax 93.292.54.18 • subsdogc@gencat.net • www.gencat.net/eadop