ISSN: 1696-2281
SORT 31 (1) January-June (2007)

# 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 31 (1), January-June 2007**

**Formerly Qüestiió**

SORT 31 (1) January-June 2007, 3-44

# Objective Bayesian point and region estimation in location-scale models

José M. Bernardo
*Universitat de València, Spain*

**Abstract**

Point and region estimation may both be described as specific *decision problems*. In point estimation, the action space is the set of possible values of the quantity on interest; in region estimation, the action space is the set of its possible credible regions. Foundations dictate that the solution to these decision problems must depend on both the utility function and the prior distribution. Estimators intended for general use should surely be invariant under one-to-one transformations, and this requires the use of an invariant loss function; moreover, an objective solution requires the use of a prior which does not introduce subjective elements. The combined use of an invariant information-theory based loss function, the *intrinsic discrepancy*, and an objective prior, the *reference prior*, produces a general solution to both point and region estimation problems. In this paper, estimation of the two parameters of univariate location-scale models is considered in detail from this point of view, with special attention to the normal model. The solutions found are compared with a range of conventional solutions.

*MSC:* Primary: 62F15, 62C10; secondary: 62F10, 62F15, 62B10

*Keywords:* Confidence Intervals, Credible Regions, Decision Theory, Intrinsic Discrepancy, Intrinsic Loss, Location-Scale Models, Noninformative Prior, Reference Analysis, Region Estimation, Point Estimation.

## 1 Introduction

Point and region estimation of the parameters of location-scale models have a long, fascinating history which is far from settled. Indeed, the list of contributors to the simpler examples of this class of problems, estimation of the normal mean and estimation of the normal variance, reads like a *Who's Who* in 20th century statistics.

*Address for correspondence:* José M. Bernardo is Professor of Statistics at the Universitat de València. Departamento de Estadística e I. O., Facultad de Matemàticas, 46100-Burjassot, Valencia, Spain.
jose.m.bernardo@uv.es, www.uv.es/bernardo
Received: March 2006

In this paper, an objective Bayesian decision-theoretic solution to both point and region estimation of the parameters of location-scale models is presented, with special attention devoted to the normal model. In marked contrast with most approaches, the solutions found are *invariant* under one-to-one reparametrization.

### 1.1 Notation

Probability distributions are described through their probability density functions, and no notational distinction is made between a random quantity and the particular values that it may take. Bold italic roman fonts are used for observable random vectors (typically data) and bold italic greek fonts for unobservable random vectors (typically parameters); lower case is used for variables and upper case calligraphic for their dominion sets. The standard mathematical convention of referring to functions, say $f_x(\cdot)$ and $g_x(\cdot)$ of $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, respectively by $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ is often used. Thus, the conditional probability density of observable data $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ given $\boldsymbol{\omega}$ is represented by either $p_x(\cdot \,|\, \boldsymbol{\omega})$ or $p(\boldsymbol{x} \,|\, \boldsymbol{\omega})$, with $p(\boldsymbol{x} \,|\, \boldsymbol{\omega}) \geq 0$, $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, and $\int_{\boldsymbol{\mathcal{X}}} p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, d\boldsymbol{x} = 1$, and the posterior density of a non-observable parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ given data $\boldsymbol{x}$ is represented by either $\pi_{\boldsymbol{\theta}}(\cdot \,|\, \boldsymbol{x})$ or $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$, with $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \geq 0$ and $\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta} = 1$. Density functions of specific distributions are denoted by appropriate names. In particular, if $x$ has a normal distribution with mean $\mu$ and standard deviation $\sigma$, its probability density function will be denoted $\mathrm{N}(x \,|\, \mu, \sigma)$, and if $\lambda$ has a gamma distribution with parameters $\alpha$ and $\beta$, its probability density function will be denoted $\mathrm{Ga}(\lambda \,|\, \alpha, \beta)$, with $\mathrm{E}[\lambda] = \alpha/\beta$, and $\mathrm{Var}[\lambda] = \alpha/\beta^2$.

It is assumed that available data $\boldsymbol{x}$ consist of one observation from the family $\mathcal{F} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\omega} \in \Omega\}$ of probability distributions for $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, and that one is interested in point and region estimation of some function $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \boldsymbol{\Theta}$ of the unknown parameter vector $\boldsymbol{\omega}$. Often, but not necessarily, data consist of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of some simpler model $\{q(x \,|\, \boldsymbol{\omega}),\, x \in \mathcal{X},\, \boldsymbol{\omega} \in \Omega\}$, in which case, $\boldsymbol{\mathcal{X}} = \mathcal{X}^n$ and the likelihood function is $p(\boldsymbol{x} \,|\, \boldsymbol{\omega}) = \prod_{j=1}^{n} q(x_j \,|\, \boldsymbol{\omega})$. Without loss of generality, the original parametric family $\mathcal{F}$ may be written as

$$\mathcal{F} \equiv \{p_x(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\theta} \in \boldsymbol{\Theta},\, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\} \tag{1}$$

in terms of the vector of interest $\boldsymbol{\theta}$, and a vector $\boldsymbol{\lambda}$ of nuisance parameters. A *point estimator* of $\boldsymbol{\theta}$ is some function of the data $\tilde{\boldsymbol{\theta}}(\boldsymbol{x}) \in \boldsymbol{\Theta}$ such that, for each possible set of observed data $\boldsymbol{x}$, $\tilde{\boldsymbol{\theta}}(\boldsymbol{x})$ could be regarded as an appropriate proxy for the actual, unknown value of $\boldsymbol{\theta}$. A *p-credible region* of $\boldsymbol{\theta}$ is some subset $C_p(\boldsymbol{x}, \boldsymbol{\Theta})$ of $\boldsymbol{\Theta}$ whose posterior probability is $p$. Within this framework, attention in this paper focuses on problems where data consist of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from a *location-scale* model

$m(x\,|\,\mu, \sigma, f)$, of the form

$$m(x\,|\,\mu, \sigma, f) = \sigma^{-1} f\{\sigma^{-1}(x-\mu)\}, \quad x \in \Re, \quad \mu \in \Re, \quad \sigma > 0, \tag{2}$$

where $f(.)$ is some probability density in $\Re$, so that $f(y) \geq 0$, $\int_{\Re} f(y)\, dy = 1$. Interest lies in either the location parameter $\mu$, the scale parameter $\sigma$, or some one-to-one function of these, and the likelihood function is

$$p(\boldsymbol{x}\,|\,\mu, \sigma) = \prod_{j=1}^{n} m(x_j\,|\,\mu, \sigma, f) = \sigma^{-n} \prod_{j=1}^{n} f\{\sigma^{-1}(x_j-\mu)\}. \tag{3}$$

Standard notation is used for the sample mean and the sample variance, respectively denoted by $\overline{x} = \sum_{j=1}^{n} x_j/n$ and $s^2 = \sum_{j=1}^{n}(x_j - \overline{x})^2/n$. Many conventional point estimators of the variance of location-scale models are members of the family of *affine invariant estimators*,

$$\tilde{\sigma}^2_{\nu} = \frac{ns^2}{\nu} = \frac{1}{\nu}\sum_{j=1}^{n}(x_j - \overline{x})^2, \quad \nu > 0\,. \tag{4}$$

In particular, with normal data, the MLE of the variance $\sigma^2$ is $s^2 = \tilde{\sigma}^2_n$, and the unbiased estimator is $\tilde{\sigma}^2_{n-1}$. More sophisticated estimators may sometimes be defined in terms of affine estimators; for instance, Stein (1964) and Brown (1968) estimators of the normal variance may respectively be written as

$$\tilde{\sigma}^2_{stein} = \min\left\{\tilde{\sigma}^2_{n+1},\, \tilde{\sigma}^2_{(n+2)/(1+z^2)}\right\}, \quad \tilde{\sigma}^2_{brown} = \min\left\{\tilde{\sigma}^2_{n-1},\, \tilde{\sigma}^2_{n/(1+z^2)}\right\},$$

where $z = \overline{x}/s$ is the standardized sample mean.

### 1.2 Contents

Section 2 provides a short review of intrinsic estimation, our approach to both point and region estimation. An information-theory based invariant loss function, the *intrinsic discrepancy*, is proposed as a reasonable general alternative to the conventional (non-invariant) quadratic loss. As is usually the case in modern literature, point estimation is described as a decision problem where the action space is the set of possible values for the quantity of interest; an intrinsic point estimator is then defined as the Bayes estimator which corresponds to the intrinsic loss and the appropriate reference prior. This provides a general *invariant* objective Bayes point estimator. Less conventionally, region estimation is also described as a decision problem where, for each $p$, the action space is the set of possible $p$-credible regions for the quantity of interest; a $p$-credible intrinsic region estimator is then defined as the lowest posterior loss $p$-credible region with respect to the intrinsic loss and the appropriate reference prior. This provides a

general *invariant* objective Bayes region estimator which always contains the intrinsic point estimator.

In Section 3 location-scale models are analyzed from this point of view. In particular, intrinsic point estimators and intrinsic region estimators are derived for the mean of a normal model, the variance of a normal model, and the scale parameter of a Cauchy model.

## 2 Intrinsic Estimation

### 2.1 The intrinsic discrepancy loss function

Point estimation of some parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is customarily described as a decision problem where the action space is the set $\mathcal{A} = \{\tilde{\boldsymbol{\theta}};\, \tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}\}$ of possible values of the vector of interest. Foundations dictate (see e.g., Bernardo and Smith, 1994, Ch. 2 and references therein) that to solve this decision problem it is necessary to specify a *loss function* $\ell\{\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\}$, such that $\ell\{\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\} \geq 0$ and $\ell\{\boldsymbol{\theta}, \boldsymbol{\theta}\} = 0$, which describes, as a function of $\boldsymbol{\theta}$, the loss suffered from using $\tilde{\boldsymbol{\theta}}$ as a proxy for the unknown value of $\boldsymbol{\theta}$. The loss function is context specific, and should be chosen in terms of the anticipated uses of the estimate; however, a number of conventional loss functions have been suggested for those situations where no particular uses are envisaged, as in scientific inference. The simplest of these conventional loss functions (which typically ignore the presence nuisance parameters) is the ubiquitous *quadratic loss*, $\ell\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^t(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$; the corresponding Bayes estimator, if this exists, is the *posterior mean*, $E[\boldsymbol{\theta} \,|\, \boldsymbol{x}]$. Another common conventional loss function is the *zero-one loss*, defined as $\ell\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = 1$, if $\tilde{\boldsymbol{\theta}}$ does not belong to a $\epsilon$-radius neighbourhood of $\boldsymbol{\theta}$, and zero otherwise; as $\epsilon \to 0$, the corresponding Bayes estimator converges to the posterior mode, $\text{Mo}[\boldsymbol{\theta} \,|\, \boldsymbol{x}]$. For details, see, e.g., Bernardo and Smith (1994, p. 257).

**Example 1** *(Normal variance)* With the usual objective prior $\pi(\mu, \sigma) = \sigma^{-1}$, the (marginal) reference posterior density of $\sigma$ is the square root inverted gamma

$$\pi(\sigma \,|\, \boldsymbol{x}) = \pi(\sigma \,|\, s, n) = \frac{n^{(n-1)/2}\, s^{n-1}}{2^{(n-3)/2}\, \Gamma[(n-1)/2]}\; \sigma^{-n}\, e^{-\frac{1}{2} n s^2/\sigma^2}, \quad n \geq 2. \tag{5}$$

The quadratic loss in terms of the variance, $\ell\{\tilde{\sigma}^2, \sigma^2\} = (\tilde{\sigma}^2 - \sigma^2)^2$, leads to $\text{E}[\sigma^2 \,|\, \boldsymbol{x}] = \tilde{\sigma}^2_{n-3}$ (which obviously requires $n \geq 3$). Similarly, the quadratic loss in terms of the standard deviation, $\ell\{\tilde{\sigma}, \sigma\} = (\tilde{\sigma} - \sigma)^2$, yields

$$\text{E}[\sigma \,|\, \boldsymbol{x}] = \sqrt{\frac{n}{2}}\, \frac{\Gamma[(n-2)/2]}{\Gamma[(n-1)/2]}\, s, \quad n \geq 3. \tag{6}$$

For moderate $n$ values, the Stirling approximation of the Gamma functions in (6) produces $\mathrm{E}[\sigma \mid \boldsymbol{x}]^2 \approx \tilde{\sigma}^2_{n-5/2}$. Notice that, using the conventional quadratic loss function, the Bayes estimate of $\sigma^2$ is *not* the square of the Bayes estimate of $\sigma$. This lack of invariance is not an special feature of the quadratic loss; on the contrary, this is the case of most conventional loss functions. For instance, the use of the slightly more sophisticated standardized quadratic loss function on the variance,

$$\ell_{stq}(\tilde{\sigma}^2, \sigma^2) = [(\tilde{\sigma}^2/\sigma^2) - 1]^2 \tag{7}$$

yields (if $n \geq 2$, for $\pi(\sigma \mid \boldsymbol{x})$ to be proper)

$$\arg\min_{\tilde{\sigma}^2>0} \int_0^\infty [(\tilde{\sigma}^2/\sigma^2) - 1]^2 \, \pi(\sigma \mid \boldsymbol{x}) \, d\sigma = \frac{n\, s^2}{n+1} = \tilde{\sigma}^2_{n+1}, \tag{8}$$

which is also the minimum risk equivariant estimator (MRIE) of $\sigma^2$ under this loss, while the standardized quadratic loss function it terms of the standard deviation, $\ell(\tilde{\sigma}, \sigma) = [(\tilde{\sigma}/\sigma) - 1]^2$ yields (again, if $n \geq 2$)

$$\arg\min_{\tilde{\sigma}^2>0} \int_0^\infty [(\tilde{\sigma}/\sigma) - 1]^2 \, \pi(\sigma \mid \boldsymbol{x}) \, d\sigma = \sqrt{\frac{n}{2}} \, \frac{\Gamma[n/2]}{\Gamma[(n+1)/2]} \, s, \tag{9}$$

which is different from (6), and whose square is *not* (8). Similarly, for the zero-one loss in terms of $\sigma^2$, the Bayes estimator is the mode of the posterior distribution of $\sigma^2$, $\pi(\sigma^2 \mid \boldsymbol{x}) = \pi(\sigma \mid \boldsymbol{x})/(2\sigma)$, which is $\mathrm{Mo}(\sigma^2 \mid \boldsymbol{x}) = \tilde{\sigma}^2_{n+1}$, the same as (8), while the Bayes estimator for the zero-one loss in terms of $\sigma$ is $\mathrm{Mo}(\sigma \mid \boldsymbol{x}) = s$, the MLE of $\sigma$, whose square is obviously not the same as (8). For further information on alternative point estimators of the normal variance, see Brewster and Zidek (1974) and Rukhin (1987).

As Example 1 dramatically illustrates, conventional loss functions are typically *not* invariant under reparametrization. As a consequence, the Bayes estimator $\boldsymbol{\phi}^*$ of a one-to-one transformation $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ of the original parameter $\boldsymbol{\theta}$ is not necessarily $\boldsymbol{\phi}(\boldsymbol{\theta}^*)$ and thus, for each loss function, one may produce as many *different* estimators of the same quantity as alternative parametrizations one is prepared to consider, a less than satisfactory situation. Indeed, scientific applications *require* this type of invariance. It would certainly be hard to argue that the best estimate of, say the age of the universe is $\theta^*$ but that the best estimate of the logarithm of that age is *not* $\log(\theta^*)$. Invariant loss functions are required to guarantee invariant estimators.

With no nuisance parameters, *intrinsic loss functions* (Robert, 1996), of the general form $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \ell\{p_x(. \mid \tilde{\boldsymbol{\theta}}), \, p_x(. \mid \boldsymbol{\theta})\}$ shift attention from the discrepancy between the estimate $\tilde{\boldsymbol{\theta}}$ and the true value $\boldsymbol{\theta}$, to the more relevant discrepancy between the statistical *models* they label, and they are always invariant under one-to-one reparametrization. The *intrinsic discrepancy*, introduced by Bernardo and Rueda (2002), is a particular intrinsic loss with specially attractive properties.

**Definition 1** ***(Intrinsic Discrepancy)*** *The* intrinsic discrepancy *between two elements* $p_x(\cdot\,|\,\boldsymbol{\omega}_1)$ *and* $p_x(\cdot\,|\,\boldsymbol{\omega}_2)$ *of the parametric family of distributions* $\mathcal{F} = \{p_x(\cdot\,|\,\boldsymbol{\omega}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}(\boldsymbol{\omega}),\ \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$*, is*

$$\begin{aligned}\delta_x(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2) &= \delta\{p_x(.\,|\,\boldsymbol{\omega}_1),\ p_x(.\,|\,\boldsymbol{\omega}_2)\} = \min\{\kappa_x(\boldsymbol{\omega}_1\,|\,\boldsymbol{\omega}_2),\ \kappa_x(\boldsymbol{\omega}_2\,|\,\boldsymbol{\omega}_1)\},\\ \kappa_x(\boldsymbol{\omega}_j\,|\,\boldsymbol{\omega}_i) &= \int_{\boldsymbol{\mathcal{X}}(\boldsymbol{\omega}_i)} p_x(\boldsymbol{x}\,|\,\boldsymbol{\omega}_i) \log \frac{p_x(\boldsymbol{x}\,|\,\boldsymbol{\omega}_i)}{p_x(\boldsymbol{x}\,|\,\boldsymbol{\omega}_j)}\, d\boldsymbol{x},\end{aligned}$$

*The intrinsic discrepancy* $\delta_x\{\mathcal{F}_1, \mathcal{F}_2\}$ *between two subsets* $\mathcal{F}_1$ *and* $\mathcal{F}_2$ *of* $\mathcal{F}$ *is the minimum intrinsic discrepancy between its elements,*

$$\delta_x(\mathcal{F}_1, \mathcal{F}_2) = \min_{\boldsymbol{\omega}_1 \in \mathcal{F}_1,\ \boldsymbol{\omega}_2 \in \mathcal{F}_2} \delta\{p_x(.\,|\,\boldsymbol{\omega}_1),\ p_x(.\,|\,\boldsymbol{\omega}_2)\}$$

Thus, the intrinsic discrepancy $\delta(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ between two parameter values $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ is the minimum Kullback-Leibler directed logarithmic divergence (Kullback and Leibler, 1951) between the distributions $p_x(.\,|\,\boldsymbol{\omega}_1)$ and $p_x(.\,|\,\boldsymbol{\omega}_2)$ which they label. Notice that this is obviously independent of the particular parametrization chosen to describe the distributions. The intrinsic discrepancy is a divergence measure in the class $\mathcal{F}$; indeed, (i) it is symmetric, (ii) it is non-negative and (iii) it is zero if, and only if, $p_x(\boldsymbol{x}\,|\,\boldsymbol{\omega}_1) = p_x(\boldsymbol{x}\,|\,\boldsymbol{\omega}_2)$ almost everywhere. Notice that in Definition 1 the possible dependence of the sampling space $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{X}}(\boldsymbol{\omega})$ on the parameter value $\boldsymbol{\omega}$ is explicitly allowed, so that the intrinsic discrepancy may be used with non-regular models where the support $\boldsymbol{\mathcal{X}}(\boldsymbol{\omega}_1)$ of, say, $p_x(.\,|\,\boldsymbol{\omega}_1)$ may be strictly smaller than the support $\boldsymbol{\mathcal{X}}(\boldsymbol{\omega}_2)$ of $p_x(.\,|\,\boldsymbol{\omega}_2)$.

The intrinsic discrepancy is also invariant under one-to-one transformations of the random vector $\boldsymbol{x}$. Moreover, directed logarithmic divergences are *additive* with respect to conditionally independent observations. Consequently, if $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is a random sample from, say $q_x(\cdot\,|\,\boldsymbol{\omega})$ so that the probability model is $p(\boldsymbol{x}\,|\,\boldsymbol{\omega}) = \prod_{i=1}^{n} q(x_j\,|\,\boldsymbol{\omega})$, then the intrinsic discrepancy $\delta_x\{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2\}$ between $p_x(\cdot\,|\,\boldsymbol{\omega}_1)$ and $p_x(\cdot\,|\,\boldsymbol{\omega}_2)$ is simply $n\,\delta_x\{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2\}$, that is, $n$ times the intrinsic discrepancy between $q_x(\cdot\,|\,\boldsymbol{\omega}_1)$ and $q_x(\cdot\,|\,\boldsymbol{\omega}_2)$.

In the context of point estimation, the intrinsic discrepancy leads naturally to the (invariant) intrinsic discrepancy loss $\delta_x\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ defined as the intrinsic discrepancy between the assumed model $p_x(\cdot\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})$ and its closest approximation within the set $\{p_x(\cdot\,|\,\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}),\ \tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}\}$ of all models with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

**Definition 2** ***(Intrinsic discrepancy loss)*** *Consider the family of probability distributions* $\mathcal{F} = \{p_x(\cdot\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda},\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}(\boldsymbol{\omega}, \boldsymbol{\lambda})\}$*. The* intrinsic discrepancy loss *from using* $\tilde{\boldsymbol{\theta}}$ *as a proxy for* $\boldsymbol{\theta}$ *is*

$$\delta_x\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \inf_{\tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}} \delta_x\{(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}), (\boldsymbol{\theta}, \boldsymbol{\lambda})\},$$

*the intrinsic discrepancy between* $p_x(\cdot\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})$ *and the set* $\{p_x(\cdot\,|\,\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}),\ \tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}\}$.

Notice that the value of $\delta_x\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ does *not* depend on the particular parametrization chosen to describe the problem. Indeed, for any one-to-one reparametrizations $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ and $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\lambda})$,

$$\delta_x\{\tilde{\boldsymbol{\phi}}, (\boldsymbol{\phi}, \boldsymbol{\psi})\} = \delta_x\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} \tag{10}$$

so that, as one should surely require, the loss suffered from using $\tilde{\boldsymbol{\phi}} = \boldsymbol{\phi}(\tilde{\boldsymbol{\theta}})$ as a proxy for $\boldsymbol{\phi}(\boldsymbol{\theta})$ is precisely the same as the loss suffered from using $\tilde{\boldsymbol{\theta}}$ as a proxy for $\boldsymbol{\theta}$, and this is true for any parametrization of the nuisance parameter vector.

Under frequently met regularity conditions, the two minimizations required in Definition 2 may be interchanged. This makes analytical derivation of the intrinsic loss considerably simpler.

**Theorem 1** ***(Computation of the intrinsic discrepancy loss)*** *Let $\mathcal{F}$ be a parametric family of probability distributions*

$$\mathcal{F} = \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda},\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}(\boldsymbol{\theta}, \boldsymbol{\lambda})\},$$

*with convex support $\boldsymbol{\mathcal{X}}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ for all $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$. Then,*

$$\begin{aligned} \delta_x\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} &= \inf_{\tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}} \min\left\{\kappa_x\{\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}\},\ \kappa_x\{\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}\}\right\} \\ &= \min\left\{\inf_{\tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}} \kappa_x\{\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}\},\ \inf_{\tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}} \kappa_x\{\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}\}\right\} \end{aligned}$$

*Proof.* This follows from the fact that, if $\boldsymbol{\mathcal{X}}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is a convex set, then the two directed logarithmic divergences involved in the definition are convex functionals. For details, see Juárez (2004). □

**Example 2** *(Normal variance, continued)* Consider $\boldsymbol{x} = \{x_1, \ldots, x_n\}$. The directed logarithmic divergence of $\prod_{i=1}^n \mathrm{N}(x_j \,|\, \tilde{\mu}, \tilde{\sigma})$ from $\prod_{i=1}^n \mathrm{N}(x_j \,|\, \mu, \sigma)$ is

$$\begin{aligned} \kappa_x\{\tilde{\mu}, \tilde{\sigma} \,|\, \mu, \sigma\} &= n \int_{-\infty}^{\infty} \mathrm{N}(x \,|\, \mu, \sigma) \log\left[\frac{\mathrm{N}(x \,|\, \mu, \sigma)}{\mathrm{N}(x \,|\, \tilde{\mu}, \tilde{\sigma})}\right] dx \\ &= \frac{n}{2}\left[\frac{\sigma^2}{\tilde{\sigma}^2} - 1 - \log\frac{\sigma^2}{\tilde{\sigma}^2} + \frac{(\mu - \tilde{\mu})^2}{\tilde{\sigma}^2}\right]. \end{aligned} \tag{11}$$

This is minimized when $\tilde{\mu} = \mu$, to yield

$$\inf_{\tilde{\mu} \in \mathbb{R}} \kappa_x\{\tilde{\mu}, \tilde{\sigma} \,|\, \mu, \sigma\} = \frac{n}{2}\, g\left(\frac{\sigma^2}{\tilde{\sigma}^2}\right) = \frac{n}{2}\, g\,(\phi)\,,$$

where $\phi = \tilde{\sigma}^2/\sigma^2$, and $g(.)$ is the *linlog* function defined by

$$g(t) = (t-1) - \log t, \quad t > 0. \tag{12}$$

Notice that $g(t) \geq 0$ and $g(t) = 0$ if (and only if) $t = 1$. This follows from the fact that $g(t)$ is the absolute distance between $\log t$ and its tangent at $t = 1$.

Exchanging the roles of $(\tilde{\mu}, \tilde{\sigma})$ and $(\mu, \sigma)$, it is similarly found that $\kappa_x\{\mu, \sigma \,|\, \tilde{\mu}, \tilde{\sigma}\}$ is also minimized when $\tilde{\mu} = \mu$, to yield

$$\inf_{\tilde{\mu} \in \mathbb{R}} \kappa_x\{\mu, \sigma \,|\, \tilde{\mu}, \tilde{\sigma}\} = \frac{n}{2}\, g\left(\frac{\tilde{\sigma}^2}{\sigma^2}\right) = \frac{n}{2}\, g\left(\frac{1}{\phi}\right) \cdot$$

Moreover, $g(t) < g(1/t)$ if, and only if, $t < 1$ and hence, using Theorem 1,

$$\delta_x\{\tilde{\sigma}, (\mu, \sigma)\} \;=\; \delta_x\{\phi\} = \begin{cases} \frac{n}{2}\, g(\phi) & \text{if} \quad \phi < 1, \\ \frac{n}{2}\, g(1/\phi) & \text{if} \quad \phi \geq 1, \end{cases} \qquad \phi = \frac{\tilde{\sigma}^2}{\sigma^2} \cdot \tag{13}$$

Thus, for fixed $n$, the intrinsic discrepancy loss $\delta_x\{\tilde{\sigma}, (\mu, \sigma)\}$ only depends on the ratio $\phi = \tilde{\sigma}^2/\sigma^2$. The intrinsic discrepancy loss is closely related to the (also invariant) *entropy* loss,

$$\ell_{ent}\{\tilde{\sigma}, \sigma\} = \ell_{ent}\{\phi\} = \int_{-\infty}^{\infty} \mathrm{N}(x \,|\, \mu, \sigma) \log\left[\frac{\mathrm{N}(x \,|\, \mu, \sigma)}{\mathrm{N}(x \,|\, \mu, \tilde{\sigma})}\right] dx = \frac{1}{2}\, g(\phi), \tag{14}$$

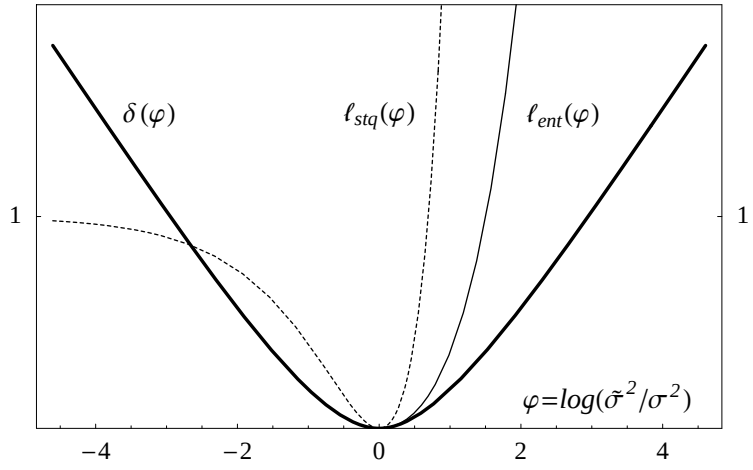


***Figure 1:*** *Intrinsic discrepancy loss (solid line), entropy loss (continuous line), and standardized quadratic loss (dotted line) for point estimation of the normal variance, as a function of* $\psi = \log(\tilde{\sigma}^2/\sigma^2)$.

which Brown (1990) attributes to Stein. Except for the proportionality constant $n$ (which does not affect estimation), the entropy loss (14) is the same as the intrinsic discrepancy loss (13) whenever $\tilde{\sigma} < \sigma$. Indeed, the intrinsic discrepancy loss may be seen as a symmetrized version of the entropy loss.

Notice that, for all values of the ratio $\phi = \tilde{\sigma}^2/\sigma^2$, $\delta_x\{\phi\} = \delta_x\{1/\phi\}$; hence, the intrinsic loss *equally* penalizes overestimation and underestimation. In sharp contrast, both the entropy loss and the often recommended standardized quadratic loss function, which is also a function of the ratio $\phi$,

$$\ell_{stq}(\tilde{\sigma}^2, \sigma^2) = [(\tilde{\sigma}^2/\sigma^2) - 1]^2 = (\phi - 1)^2,$$

clearly underpenalize small estimators, thus yielding estimators of the variance which are too small. This is illustrated in Figure 1, where the functions $\delta_x(\phi)$ (for $n = 1$), $\ell_{ent}\{\phi\}$, and $\ell_{stq}(\phi)$ are all represented as a function of $\varphi = \log \phi$. More conventional loss functions, as the usual quadratic loss,

$$\ell_{quad}(\tilde{\sigma}^2, \sigma^2) = [\tilde{\sigma}^2 - \sigma^2]^2 = \sigma^4(\phi - 1)^2,$$

are not even invariant with respect to affine transformations. All this led Stein (1964, p. 156) to write "I find it hard to take the problem of estimating $\sigma^2$ with quadratic loss very seriously".

### 2.2 Reference posterior expectation

Given data $\boldsymbol{x}$ generated by $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\lambda})$, a situation with no prior information about the value of $\boldsymbol{\theta}$ is formally described by the *reference prior* $\pi(\boldsymbol{\lambda} \mid \boldsymbol{\theta})\,\pi(\boldsymbol{\theta})$ which corresponds to the model $p_x(\cdot \mid \boldsymbol{\theta}, \boldsymbol{\lambda})$ when $\boldsymbol{\theta}$ is the quantity of interest (Bernardo, 1979; Berger and Bernardo, 1992; Bernardo, 2005a). In this case, all available information about $\boldsymbol{\theta}$ is encapsulated it its (marginal) reference posterior distribution, $\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) = \int_{\Lambda} \pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{x})\, d\boldsymbol{\lambda}$ where, by Bayes theorem, $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})\,\pi(\boldsymbol{\lambda} \mid \boldsymbol{\theta})\,\pi(\boldsymbol{\theta})$. If numerical summaries of the information encapsulated in $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ are further required in the form of either point or region estimators of $\boldsymbol{\theta}$ under some specified loss function $\ell\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, then the reference posterior expected loss

$$l(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{x}) = \int_{\Theta} \int_{\Omega} \ell\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, \pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{x})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}$$

(from using $\tilde{\boldsymbol{\theta}}$ as a proxy for $\boldsymbol{\theta}$) has to be evaluated. In view of the arguments given above, attention will focus on the intrinsic discrepancy reference expected loss, or *intrinsic expected loss*, for short.

**Definition 3** ***(Intrinsic expected loss)*** *Consider the parametric family of probability distributions*

$$\mathcal{F} = \{p_x(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}(\boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda}, )\}. \tag{15}$$

*The* intrinsic expected loss *from using $\tilde{\boldsymbol{\theta}}$ given data $\boldsymbol{x}$, denoted $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x})$, is the posterior expectation of the intrinsic discrepancy loss, $\delta_x\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ with respect to the joint reference posterior, $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x})$,*

$$d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}) = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Omega}} \delta_x\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, \pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}, \tag{16}$$

*where $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})\, \pi(\boldsymbol{\lambda} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$, and $\pi(\boldsymbol{\lambda} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$ is the (joint) reference prior when $\boldsymbol{\theta}$ is the quantity of interest.*

The function $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x})$ measures the posterior expected loss from using $\tilde{\boldsymbol{\theta}}$ as a proxy for the unknown value of $\boldsymbol{\theta}$, in terms of the expected intrinsic discrepancy between the assumed model, $p_x(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$, and the class

$$\mathcal{F}_{\tilde{\boldsymbol{\theta}}} = \{p(\boldsymbol{x} \,|\, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}),\ \tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda},\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}(\boldsymbol{\omega}, \boldsymbol{\lambda}) \tag{17}$$

of those models in $\mathcal{F}$ for which $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

The intrinsic expected loss provides an objective measure of the *compatibility* of the value $\tilde{\boldsymbol{\theta}}$ with the observed data $\boldsymbol{x}$, with a nice interpretation in terms of likelihood ratios. Indeed, it immediately follows from Definitions 1, 2 and 3, that $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x})$ is the posterior expectation of the minimum expected log-likelihood ratio between the true model and the closest model for which $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$. For instance, if $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}) = \log 100$, then data $\boldsymbol{x}$ are expected to be about 100 times more likely under the true (unknown) model than under any model within this family with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

**Example 3** *(Normal variance, continued)* In Example 2 the intrinsic discrepancy loss from using $\tilde{\sigma}$ as a proxy for $\sigma$ was seen to be a function $\delta_x(\phi)$ of the ratio $\phi = \tilde{\sigma}^2/\sigma^2$ (Equation 13). Changing variables in (5), the reference posterior of $\phi$ is

$$\pi(\phi \,|\, \boldsymbol{x}) = \mathrm{Ga}(\phi \,\Big|\, \frac{n-1}{2}\,,\, \frac{n\,s^2}{2\tilde{\sigma}^2})\,. \tag{18}$$

Hence, the intrinsic expected loss from using $\tilde{\sigma}$ as a proxy for $\sigma$ is

$$d(\tilde{\sigma} \,|\, \boldsymbol{x}) = d(\tilde{\sigma} \,|\, s^2, n) = \int_0^\infty \delta_x(\phi)\, \mathrm{Ga}(\phi \,|\, \frac{n-1}{2}\,,\, \frac{n\,s^2}{2\tilde{\sigma}^2})\, d\phi \tag{19}$$

which may easily be computed by one-dimensional numerical integration. Good analytical approximations will however be provided is Section 3. As a numerical

illustration, a random sample of size $n = 12$ was simulated from a normal distribution with $\mu = 5$ and $\sigma = 2$, yielding $\overline{x} = 4.214$ and $s = 2.071$. The corresponding intrinsic expected loss $d(\tilde{\sigma} \mid \boldsymbol{x})$, represented in the lower panel of Figure 2, is locally convex around a unique minimum.

### 2.3 Intrinsic Point and Region Estimation

*Bayes estimates* are, by definition, those which minimize the expected posterior loss. The *intrinsic estimate* is the Bayes estimate which corresponds to the intrinsic discrepancy loss and the reference posterior distribution, i.e., that value $\tilde{\boldsymbol{\theta}}_{int}(\boldsymbol{x}) \in \boldsymbol{\Theta}$ which minimizes the intrinsic expected loss. Formally,

**Definition 4** ***(Intrinsic point estimator)*** *Consider again the parametric family of probability distributions $\mathcal{F}$ defined by* (15). *An* intrinsic estimator *of $\boldsymbol{\theta}$ is a value*

$$\tilde{\boldsymbol{\theta}}_{int}(\boldsymbol{x}) = \min_{\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}} d\{\tilde{\boldsymbol{\theta}} \mid \boldsymbol{x}\},$$

*which minimizes the intrinsic discrepancy reference posterior loss* (16).

Under general regularity conditions, the intrinsic expected loss $d\{\tilde{\boldsymbol{\theta}} \mid \boldsymbol{x}\}$ is locally convex near its absolute minimum and, therefore, the intrinsic estimate is typically exists and it is unique. Moreover, since both the intrinsic loss function and the reference prior are invariant under one-to-one reparametrization, the intrinsic estimator $\tilde{\boldsymbol{\psi}}_{int}(\boldsymbol{x})$ of any one-to-one function $\boldsymbol{\psi}(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ will simply be $\tilde{\boldsymbol{\psi}}_{int} = \boldsymbol{\psi}(\tilde{\boldsymbol{\theta}}_{int})$. For more details on intrinsic estimation, see Bernardo and Juárez (2003).

Bayesian region estimation is typically based on posterior credible regions, i.e., sets of $\boldsymbol{\theta}$ values with pre-specified posterior probabilities. However, for any fixed $p$, there are typically infinitely many $p$-credible regions. In most cases, these are chosen to be either highest posterior density (HPD) regions, or probability centred regions.

It is well known that the ubiquitous *highest posterior density* (HPD) regions are *not* consistent under reparametrization. Thus, if $\boldsymbol{\phi}\{\boldsymbol{\theta}\}$ is a one-to-one function of $\boldsymbol{\theta}$, the image of a HPD $p$-credible region of $\boldsymbol{\theta}$ will *not* generally be HPD for $\phi$. Thus, if $C_p$ is a HPD $p$-credible set estimate for, say, the perihelion of the Earth, $\log(C_p)$ will *not* be a HPD $p$-credible set estimate for its logarithm. This suggests that highest posterior density may not be a good criterion for set estimation in scientific inference.

In *one-dimensional* problems, one may define *probability centred* credible intervals, and these are invariant under reparametrization. Indeed, the probability centred $p$-credible interval of a real-valued quantity of interest $\theta$ is defined by the $(1 - p)/2$ and $(1 + p)/2$ quantiles of its posterior distribution $\pi(\theta \mid \boldsymbol{x})$, and this is invariant under one-to-one reparametrizations, since all quantiles are invariant. However, the notion cannot be uniquely extended to multidimensional problems and, even in one-dimensional

problems, their use may less than satisfactory as, for instance, in those situations where the posterior density is monotonously decreasing within its support.

Whenever a loss structure has been established, foundations dictate that values with smaller expected loss are to be preferred. Thus, for any loss function $\ell\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\omega})\}$ is it natural to define $p$-credible *lowest posterior loss* (LDL) region estimators (Bernardo, 2005b) as those $p$-credible regions which contain $\tilde{\boldsymbol{\theta}}$ values whose expected loss $l(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x})$ (Eq. 15), is smaller than that of any $\tilde{\boldsymbol{\theta}}$ values outside the region.

In particular, if the loss function is quadratic, so that

$$\ell\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^t (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}),$$

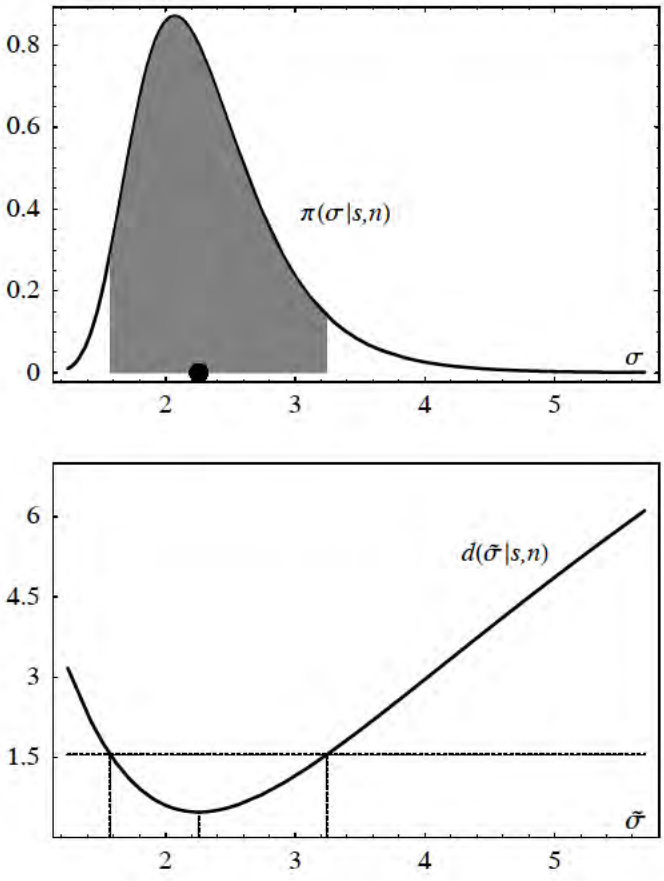


***Figure 2:*** *Reference posterior density of the standard deviation $\sigma$ of a normal distribution (upper panel), and intrinsic expected loss from using $\tilde{\sigma}$ as a proxy for $\sigma$ (lower panel), given a random sample $\boldsymbol{x}$ of size $n = 12$ with standard deviation $s = 2.071$. The intrinsic estimate (solid dot) is $\tilde{\sigma}_{int} = 2.256$; the 0.90 intrinsic credible region (shaded region) is $C^{int}_{0.90} = (1.575, 3.243)$.*

the expected loss is

$$\begin{aligned} l(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{x}) &= \int_{\boldsymbol{\Theta}} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^t (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\, d\boldsymbol{\theta} \\ &= (\tilde{\boldsymbol{\theta}} - \mathrm{E}[\boldsymbol{\theta} \mid \boldsymbol{x}])^t (\tilde{\boldsymbol{\theta}} - \mathrm{E}[\boldsymbol{\theta} \mid \boldsymbol{x}]) + \mathrm{Var}[\boldsymbol{\theta} \mid \boldsymbol{x}]. \end{aligned}$$

Hence, with quadratic loss, the lowest posterior loss $p$-credible region consists of those $\tilde{\boldsymbol{\theta}}$ values with the smallest Euclidean distance to the posterior mean $\mathrm{E}[\boldsymbol{\theta} \mid \boldsymbol{x}]$. Notice that these LDL $p$-credible regions are *not* invariant under reparametrization.

To obtain LDL invariant region estimators the loss function used must be invariant under one-to-one reparametrization. The arguments mentioned above suggest the use of the intrinsic discrepancy loss. The $p$-credible *intrinsic region estimator* is the lowest posterior loss $p$-credible region which corresponds to the intrinsic discrepancy loss.

**Definition 5** ***(Intrinsic region estimator)*** *Consider once more the parametric family of probability distributions $\mathcal{F}$ defined by* (15). *An* intrinsic $p$-credible region *for $\boldsymbol{\theta}$ is a subset $C_p^{int} = C_p^{int}(\boldsymbol{x}, \boldsymbol{\Theta})$ of $\boldsymbol{\Theta}$ such that*

$$\int_{C_p^{int}} \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\, d\boldsymbol{\theta} = p, \quad \forall \tilde{\boldsymbol{\theta}}_i \in C_p^{int},\ \forall \tilde{\boldsymbol{\theta}}_j \notin C_p^{int},\ d\{\tilde{\boldsymbol{\theta}}_i \mid \boldsymbol{x}\} < d\{\tilde{\boldsymbol{\theta}}_j \mid \boldsymbol{x}\},$$

*where, again, $d\{\tilde{\boldsymbol{\theta}} \mid \boldsymbol{x}\}$ is the intrinsic expected loss* (16).

Intrinsic credible regions are typically unique and, since they are based in the invariant intrinsic discrepancy loss, they are consistent under one-to-one reparametrization. Thus, if $\boldsymbol{\psi}(\boldsymbol{\theta})$ is a one-to-one function of $\boldsymbol{\theta}$, the image $C_p^{int}(\boldsymbol{x}, \boldsymbol{\Psi}) = \boldsymbol{\psi}\{C_p^{int}(\boldsymbol{x}, \boldsymbol{\Theta})\}$ of an intrinsic $p$-credible region for $\boldsymbol{\theta}$ is an intrinsic $p$-credible region for $\boldsymbol{\phi}$. For more details on intrinsic region estimation, see Bernardo (2005b).

**Example 4** *(Normal variance, continued)* Numerical minimization of the intrinsic expected loss (19) in Example 3 immediately yields the intrinsic estimator of of the standard deviation $\sigma$. This is

$$\sigma^*(\boldsymbol{x}) = \sigma^*(n, s) = \arg\min_{\tilde{\sigma}>0}\ d(\tilde{\sigma} \mid \boldsymbol{x}) = 2.256, \tag{20}$$

and it is marked with a solid dot in the top panel of Figure 2. Since intrinsic estimation is invariant, the intrinsic estimates of $\sigma^2$ or $\log \sigma$ are respectively $(\sigma^*)^2$ and $\log(\sigma^*)$.

Moreover, the intrinsic $p$-credible interval for $\sigma$ is given by $C_p^{int} = (\sigma_0, \sigma_1)$, where $\{\sigma_0, \sigma_1\}$ is the unique solution to the equations system

$$\begin{cases} d(\sigma_0 \mid \boldsymbol{x}) = d(\sigma_1 \mid \boldsymbol{x}) \\ \int_{\sigma_0}^{\sigma_1} \pi(\sigma \mid \boldsymbol{x})\, d\sigma = p \end{cases}$$

For instance, with $p = 0.90$ this yields $C^{int}_{0.90} = (1.575, 3.243)$, the shaded region in the top panel of Figure 2. Since intrinsic region estimation is also invariant under reparametrization, the intrinsic $p$-credible intervals for $\sigma^2$ or $\log\sigma$ will respectively be $(C^{int}_p)^2$ and $\log(C^{int}_p)$.

## 3 Intrinsic estimation in location-scale models

This section analyses intrinsic point and region estimation of the parameters $\mu$ and $\sigma$ (or arbitrary one-to-one functions of these) of variation-independent location-scale models.

### *3.1 Reference analysis of location-scale models*

The likelihood function $p(\boldsymbol{x}\,|\,\mu, \sigma, f)$ which corresponds to a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from a location-scale model $m(x\,|\,\mu, \sigma, f)$ of the form (2), is given by Equation (3). This will typically have a unique maximum, denoted $(\hat{\mu}, \hat{\sigma})$, which not always has a simple analytical expression.

Under appropriate regularity conditions (see, e.g., Bernardo and Smith, 1994, Sec. 5.3 and references therein) the joint posterior distribution of $\mu$ and $\sigma$ will be asymptotically normal with mean $(\hat{\mu}, \hat{\sigma})$ and covariance matrix

$$V(\hat{\mu}, \hat{\sigma}, n) = n^{-1}\, F^{-1}(\hat{\mu}, \hat{\sigma}) = (\tilde{\sigma}^2/n)\, A^{-1}(f) \tag{21}$$

where $F(\mu, \sigma)$ is Fisher's information matrix which, in location-scale models, is always of the form $F(\mu, \sigma) = \sigma^{-2} A(f)$, where $A(f)$ is a $2 \times 2$ matrix which depends on the probability density $f(\cdot)$, but not on $\mu$ or $\sigma$. As a consequence, if the parameter of interest is either $\mu$ (or a one-to-one function of $\mu$) or $\sigma$ (or a one-to-one function of $\sigma$) with independent variation, then (Fernández and Steel, 1999, Th. 1), under regularity conditions sufficient to guarantee posterior asymptotic normality, and variation independence of $\mu$ and $\sigma$, the joint reference prior is independent of the function $f(\cdot)$ and, in terms of $\mu$ and $\sigma$, is given by

$$\pi(\mu)\,\pi(\sigma\,|\,\mu) = \pi(\sigma)\,\pi(\mu\,|\,\sigma) = \sigma^{-1}. \tag{22}$$

Using Bayes theorem, the corresponding joint reference posterior is

$$\pi(\mu, \sigma\,|\,\boldsymbol{x}) = \sigma^{-(n+1)} \prod_{j=1}^{n} f\{\sigma^{-1}(x_j - \mu)\}, \tag{23}$$

which is typically proper for all $n \geq 2$. In particular, it is proper for all $n \geq 2$ whenever

$f(.)$ is either a standard normal or a scale mixture of standard normals, what includes Student models.

In the normal case, with $f(x) = \mathrm{N}(x \,|\, 0, 1)$, the joint posterior (23) becomes

$$\pi(\mu, \sigma \,|\, \overline{x}, s, n) = \mathrm{N}(\mu \,|\, \overline{x}, \sigma/\sqrt{n})\, \pi(\sigma \,|\, s, n),$$

where $\pi(\sigma \,|\, s, n)$ is the square root inverted gamma given by (5). The corresponding marginal reference posterior of the precision $\lambda = \sigma^{-2}$ is found to be $\pi(\lambda \,|\, \boldsymbol{x}) = \mathrm{Ga}(\lambda \,|\, (n-1)/2, (ns^2)/2)$ and, thus,

$$\mathrm{E}[\lambda \,|\, \boldsymbol{x}] = \frac{n-1}{n\, s^2}\,, \quad \mathrm{Var}[\lambda \,|\, \boldsymbol{x}] = \frac{2(n-1)}{n^2\, s^4}\,. \tag{24}$$

The marginal reference posterior of $\mu$ is the Student distribution

$$\pi(\mu \,|\, \boldsymbol{x}) = \mathrm{St}\Big(\mu \,\Big|\, \overline{x},\ \frac{s}{\sqrt{n-1}}\,,\, n\Big) \propto \Big(1 + \frac{(\mu - \overline{x})^2}{s^2}\Big)^{-n/2}. \tag{25}$$

For details see, for instance, Bernardo and Smith (1994, Sec. 5.4).

### *3.2 Intrinsic estimation of the normal mean*

As stated in Example 2 (Eq. 11), the directed divergence $\kappa\{\mu_j, \sigma_j \,|\, \mu_i, \sigma_i\}$, of $\mathrm{N}(x \,|\, \mu_j, \sigma_j)$ from $\mathrm{N}(x \,|\, \mu_i, \sigma_i)$, is

$$\kappa\{\mu_j, \sigma_j \,|\, \mu_i, \sigma_i\} = \frac{1}{2}\Big\{\frac{\sigma_i^2}{\sigma_j^2} - 1 - \log\frac{\sigma_i^2}{\sigma_j^2} + \frac{(\mu_i - \mu_j)^2}{\sigma_j^2}\Big\}.$$

As a function of $\tilde{\sigma}$, the directed divergence $\kappa\{\tilde{\mu}, \tilde{\sigma} \,|\, \mu, \sigma\}$ is minimized when $\tilde{\sigma}^2$ takes the value $\tilde{\sigma}^2_{min} = (\mu - \mu_i)^2 + \sigma^2$, and substitution yields

$$\kappa\{\tilde{\mu}, \tilde{\sigma}_{min} \,|\, \mu, \sigma\} = \frac{1}{2}\log\Big[1 + \frac{(\mu - \tilde{\mu})^2}{\sigma^2}\Big].$$

Similarly, the directed divergence $\kappa\{\mu, \sigma \,|\, \tilde{\mu}, \tilde{\sigma}\}$ is minimized, as a function of $\tilde{\sigma}$, when $\tilde{\sigma} = \sigma$, and substitution now yields

$$\kappa\{\mu, \sigma \,|\, \tilde{\mu}, \sigma\} = \frac{1}{2}\,\frac{(\mu - \tilde{\mu})^2}{\sigma^2}\,.$$

Hence, making use of Theorem 1 and the fact that, for all $x > 0$, $\log(1 + x) \leq x$, the intrinsic discrepancy loss $\delta\{\tilde{\mu}, (\mu, \sigma)\}$ from using $\tilde{\mu}$ as a proxy for $\mu$ with a normal sample of size $n$ is

$$\delta\{\tilde{\mu}, (\mu, \sigma)\} = \delta\{\theta^2\} = \frac{n}{2} \log[1 + \frac{\theta^2}{n}], \quad \theta = \theta(\tilde{\mu}, \mu, \sigma) = \frac{\mu - \tilde{\mu}}{\sigma/\sqrt{n}}, \tag{26}$$

which only depends on the number $\theta$ of standard deviations which separate $\tilde{\mu}$ from $\mu$. Figure 3 represents the intrinsic loss function (26), as a function of $\theta$, for several values of $n$.

As one might expect, $\delta\{\tilde{\mu}, (\mu, \sigma)\}$ increases with $|\theta|$. The dependence is essentially quadratic in a neighbourhood of zero, but shows a very reasonable concavity in regions where $|\theta|$ is large.

Using Definition 3, the intrinsic discrepancy reference expected loss $d(\tilde{\mu} \mid \boldsymbol{x})$ may be written in terms of the reference posterior of $\theta$; indeed,

$$\begin{aligned} d(\tilde{\mu} \mid \boldsymbol{x}) &= \int_0^\infty \int_{-\infty}^\infty \delta\{\tilde{\mu}, (\mu, \sigma)\}\, \pi(\mu, \sigma \mid \boldsymbol{x})\, d\mu\, d\sigma \\ &= \int_0^\infty \frac{n}{2} \log\left[1 + \frac{\theta^2}{n}\right] \pi(\theta \mid \boldsymbol{x})\, d\theta. \end{aligned} \tag{27}$$

But $\theta = (\tilde{\mu} - \mu)/(\sigma/\sqrt{n})$ may be written as $a + \beta$ where, as a function of $\mu$ and $\sigma$, $\beta = (\mu - \overline{x})/(\sigma/\sqrt{n})$ has a standard normal reference posterior, and $a$ is the constant $a = (\overline{x} - \tilde{\mu})/(\sigma/\sqrt{n})$. Hence, the conditional posterior distribution of $\theta^2$ given $\sigma$ is noncentral $\chi^2$ with one degree of freedom and non centrality parameter $a^2$,

$$\pi(\theta^2 \mid \boldsymbol{x}, \sigma) = \chi^2(\theta^2 \mid 1, a^2), \quad a^2 = n\, \frac{(\overline{x} - \tilde{\mu})^2}{\sigma^2}.$$

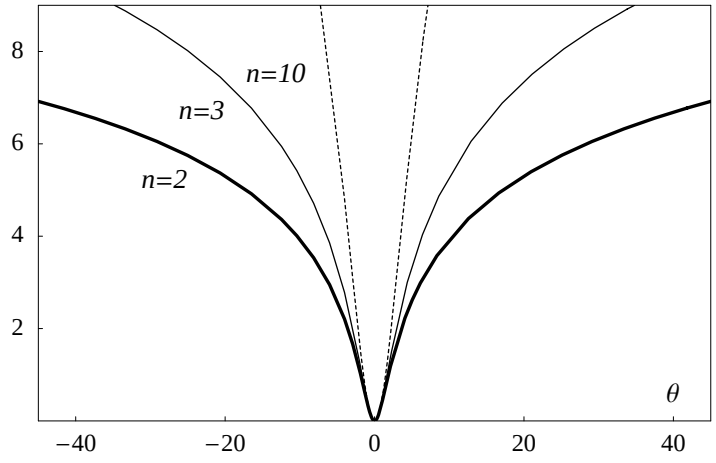


***Figure 3:*** *Intrinsic discrepancy loss for estimation of the normal mean as a function of the number* $\theta = (\tilde{\mu} - \mu)/(\sigma/\sqrt{n})$ *of standard deviations which separates* $\tilde{\mu}$ *from* $\mu$*, for* $n = 2$*,* $n = 3$*, and* $n = 10$*.*

It follows that the intrinsic expected loss $d(\tilde{\mu} \mid \boldsymbol{x})$ only depends on $\tilde{\mu}$ through $(\overline{x} - \tilde{\mu})^2$, and increases with $(\overline{x} - \tilde{\mu})^2$; therefore, the intrinsic estimator of $\mu$ is

$$\tilde{\mu}_{int}(\boldsymbol{x}) = \arg\min_{\tilde{\mu} \in \Re} \; d(\tilde{\mu} \mid \boldsymbol{x}) = \arg\min_{\tilde{\mu} \in \Re} \; (\overline{x} - \tilde{\mu})^2 = \overline{x}.$$

Moreover, $d(\tilde{\mu} \mid \boldsymbol{x})$ is symmetric around $\overline{x}$ and, hence, all intrinsic credible regions must be centered at $\overline{x}$. In view of (25), this implies that the intrinsic the $p$-credible regions are just the usual Student-$t$ HPD $p$-credible intervals

$$C_p^{int}(\boldsymbol{x}) = \left\{ \tilde{\mu};\; \overline{x} - q_{p,n}\, s/\sqrt{n-1} \le \tilde{\mu} \le \overline{x} + q_{p,n}\, s/\sqrt{n-1} \right\}, \tag{28}$$

where $q_{p,n}$ is the $(p+1)/2$ quantile of a standard Student-$t$ with $n-1$ degrees of freedom.

It immediately follows from (28) that $C_p^{int}$ consist of the set of $\tilde{\mu}$ values such that $(\overline{x} - \tilde{\mu})/(s/\sqrt{n-1})$ belongs to a probability $p$ centred interval of a standard Student-$t$ with $n-1$ degrees of freedom. But, as a function of the data $\boldsymbol{x}$, the sampling distribution of

$$t(\boldsymbol{x}) = (\overline{x} - \mu)/(s/\sqrt{n-1}) \tag{29}$$

is also a standard Student-$t$ with $n-1$ degrees of freedom. Hence, for all sample sizes, the *expected coverage under sampling* of the $p$-credible intervals (28) is *exactly* $p$, and the intrinsic credible regions are exact frequentist confidence intervals.

A simple asymptotic approximation to $d(\tilde{\mu} \mid \boldsymbol{x})$, which provides a direct measure in a log-likelihood ratio scale of the expected loss associated to the use of $\tilde{\mu}$, may easily be obtained. Indeed, a variation of the delta method shows that, under appropriate regularity conditions, the expectation of some function $y = g(x)$ of a random quantity $x$ with mean $\mu_x$ and variance $\sigma_x^2$ may be approximated by

$$E[g(x)] \approx g\left[\mu_x + \frac{\sigma_x^2}{2} \frac{g''(\mu_x)}{g'(\mu_x)}\right]. \tag{30}$$

On the other hand, the conditional posterior mean of $\theta^2$ is $1 + a^2$, and its conditional posterior variance is $2 + 4a^2$; but $\mathrm{E}[\sigma^{-2} \mid \boldsymbol{x}] = \mathrm{E}[\lambda \mid \boldsymbol{x}] = (n-1)/(ns^2)$ (Eq. 24) and hence, the unconditional posterior mean and variance of $\theta^2(\tilde{\mu})$ are, respectively,

$$\mathrm{E}[\theta^2 \mid \boldsymbol{x}] = 1 + t^2, \quad \mathrm{Var}[\theta^2 \mid \boldsymbol{x}] = 2 + 4\,t^2,$$

both functions of the conventional $t$ statistic (29). Using these in (30) to approximate the posterior expectation of $\log(1 + \theta^2/n)$ required in (27) yields

$$d(\tilde{\mu} \mid \boldsymbol{x}) \approx \frac{n}{2} \log\left[1 + \frac{1}{n} \frac{n(1+t^2) + t^4}{n + t^2 + 1}\right]. \tag{31}$$

Progressively cruder, but simpler approximations are

$$d(\tilde{\mu} \,|\, \boldsymbol{x}) \approx \frac{n}{2} \log\left[1 + \frac{1}{n}\left(1 + t^2\right)\right] \approx \frac{1}{2}\left(1 + t^2\right) . \tag{32}$$

Thus, for large $n$, the intrinsic expected loss $d(\tilde{\mu} \,|\, \boldsymbol{x})$ is essentially quadratic in the number $t = (\overline{x} - \tilde{\mu})/(s/\sqrt{n-1})$ of standard deviations which separate $\overline{x}$ from $\tilde{\mu}$. Summarizing, we have thus established

**Theorem 2** ***(Intrinsic estimation of the Normal mean)*** *Let $\boldsymbol{x}$ be a random sample of size n from $\mathrm{N}(x \,|\, \mu, \sigma)$, with mean and variance $\overline{x}$ and $s^2$, and let $t = \sqrt{n-1}\,(\overline{x} - \tilde{\mu})/s)$ be the conventional t statistic.*

*(i) The intrinsic point estimator of $\mu$ is $\tilde{\mu}_{int}(\boldsymbol{x}) = \overline{x}$.*

*(ii) The unique p-credible intrinsic region for $\mu$ is the probability centred interval*

$$C_p^{int}(\boldsymbol{x}) = \overline{x} \;\pm\; q_{p,n}\; s/\sqrt{n-1}\;,$$

*where $q_{p,n}$ is the $(p+1)/2$ quantile of a standard Student-t distribution with $n-1$ degrees of freedom. For all sample sizes, the frequentist coverage of $C_p^{int}(\boldsymbol{x})$ is exactly p.*

*(iii) The expected intrinsic loss associated to the use of $\tilde{\mu}$ as a proxy for $\mu$ is*

$$d(\tilde{\mu} \,|\, \boldsymbol{x}) \approx \frac{n}{2} \log\left[1 + \frac{1}{n}\,\frac{n(t^2+1) + t^4}{n + t^2 + 1}\right] \approx \frac{n}{2} \log\left[1 + \frac{1}{n}\left(1 + t^2\right)\right] .$$

As a numerical illustration, a random sample of size $n = 25$ was generated from a standard normal, yielding $\overline{x} = -0.162$ and $s = 0.840$. The intrinsic estimator is $\mu^* = \overline{x} = -0.162$ and the 0.99-intrinsic credible region is the interval $[-0.642, 0.318]$. The exact value of the expected intrinsic loss $d(1/3 \,|\, \boldsymbol{x})$, computed from (27) by numerical integration, is 3.768, while (31) and the two approximations in (32) respectively yield 3.781, 3.970 and 4.673. Hence, the observed data may be expected to be about $\exp(3.768) \approx 43$ times more likely under the true value of $\mu$ that under the closest normal model with $\mu = 1/3$, suggesting that the value $\mu = 1/3$ is hardly compatible with the observed data.

### *3.3 Intrinsic estimation of the normal variance*

It has already been established (Example 2, Eq. 13) that the intrinsic discrepancy loss from using $\tilde{\sigma}^2$ as a proxy for $\sigma^2$ is

$$\delta_x\{\tilde{\sigma}^2, (\mu, \sigma)\} \;=\; \delta_x\{\phi\} = \begin{cases} \dfrac{n}{2}\, g(\phi) & \text{if} \quad \phi < 1, \\ \dfrac{n}{2}\, g(1/\phi) & \text{if} \quad \phi \geq 1, \end{cases} \tag{33}$$

where $g(t) = (t-1) - \log t$, and $\phi = \tilde{\sigma}^2/\sigma^2$, and that this is also the intrinsic loss $\delta_x\{\tilde{\psi}, (\mu, \sigma)\}$ from using $\tilde{\psi}$ a a proxy for $\psi$ for any one-to-one function $\psi(\sigma^2)$ of $\sigma^2$. Moreover, the reference posterior of $\phi$ is the gamma distribution of Eq. 18. Hence, the intrinsic estimator of $\sigma^2$ is

$$\tilde{\sigma}^2_{int}(\boldsymbol{x}) \;=\; \arg\min_{\tilde{\sigma}^2>0} \int_0^\infty \delta_x(\phi)\, \mathrm{Ga}\Big(\phi \,\Big|\, \frac{n-1}{2}\,,\; \frac{n\,s^2}{2\tilde{\sigma}^2}\Big)\, d\phi,$$

where $\delta_x(\phi)$ is given by (33). Moreover, it immediately follows from (18) that, as a function of $\sigma$, the reference posterior distribution of $\tau = n\,s^2/\sigma^2$ is

$$\pi(\tau \,|\, \boldsymbol{x}) = \pi(\tau \,|\, n) = \chi^2(\tau \,|\, n-1) \tag{34}$$

a central $\chi^2$ with $n-1$ degrees of freedom; but $\phi = c\,\tau/n$, with $c = \tilde{\sigma}^2/s^2$ and, therefore, the expected posterior loss from using $\tilde{\sigma}$ may further be written as

$$d(\tilde{\sigma}^2 \,|\, s^2, n) = d(c \,|\, n) = \int_0^\infty \delta\Big(\frac{c\,\tau}{n}\Big)\, \chi^2(\tau \,|\, n-1)\, d\tau, \qquad c = \tilde{\sigma}^2/s^2. \tag{35}$$

Thus, the intrinsic estimator of the normal variance is an affine equivariant estimator of the form

$$\sigma^2_{int}(s, n) = c^*_n\, s^2, \qquad c^*_n > 0, \tag{36}$$

where $c^*_n$ is the value of $c$ which minimizes $d(c \,|\, n)$ in (35). The exact value of $c^*_n$ may be numerically found by one-dimensional numerical integration, followed by numerical optimization. The first row of Table 1 displays the exact values of $c^*_n$ for several sample sizes. However, good analytical approximations for $c^*_n$ may be obtained.

We first consider a general approximation method. Let $\omega$ be a particular parametrization of the problem, and consider a (variance stabilizing) *reference reparametrization*

***Table 1:*** *Exact and alternative approximate values for the intrinsic point estimator of the normal variance $\sigma^2_{int} = c^*_n s^2$.*

| $n$ | 2 | 3 | 4 | 5 | 10 | 20 | 50 |
|---|---|---|---|---|---|---|---|
| $c^*_n$ | 4.982 | 2.347 | 1.803 | 1.569 | 1.231 | 1.106 | 1.041 |
| $\Big(\frac{n}{n-1}\Big)^2$ | 4.000 | 2.250 | 1.778 | 1.563 | 1.235 | 1.108 | 1.041 |
| $\frac{n}{n-1}\, e^{1/(n-1)}$ | 5.437 | 2.473 | 1.861 | 1.605 | 1.242 | 1.110 | 1.041 |
| $\frac{n}{n-2}$ | — | 3.000 | 2.000 | 1.667 | 1.250 | 1.111 | 1.042 |

$\phi(\omega)$ defined as one with a uniform reference prior. This is given by any solution to the differential equation $\phi'(\omega) = \pi(\omega)$, where $\pi(\omega)$ is the marginal reference prior for $\omega$. Under regularity conditions, the sampling distribution of $\phi(\hat{\omega})$, where $\hat{\omega} = \hat{\omega}(\boldsymbol{x})$ is the MLE of $\omega$, and the reference posterior of $\phi(\omega)$, are both asymptotically normal. Using these approximations, the intrinsic expected loss from using $\tilde{\omega}$ is found to be (Bernardo, 2005b, Theo. 4.1)

$$d(\tilde{\omega} \,|\, \boldsymbol{x}) \approx \frac{n}{2}\Big\{\sigma_\phi^2 + [\mu_\phi - \phi(\tilde{\omega})\,]^2\Big\}. \tag{37}$$

where $\mu_\phi$ and $\sigma_\phi^2$ are respectively the posterior mean and posterior variance of $\phi = \phi(\omega)$. This is minimized by $\phi(\tilde{\omega}) = \mu_\phi = \mathrm{E}[\phi \,|\, \boldsymbol{x}]$. Hence, in terms of any reference parametrization $\phi$, the intrinsic point estimate is approximately the posterior mean $\mu_\phi$ and, by invariance, the intrinsic estimator of any one-to-one function, $\psi = \psi(\phi)$ is approximately given by $\tilde{\psi}_{int} = \psi(\mu_\phi)$. Thus,

$$\tilde{\phi}_{int}(\boldsymbol{x}) \approx \mu_\phi = \mathrm{E}[\phi \,|\, \boldsymbol{x}], \quad \tilde{\omega}_{int}(\boldsymbol{x}) \approx \phi^{-1}\{\mu_\phi\}. \tag{38}$$

Under regularity conditions (see, e.g., Schervish, 1995, Sec. 7.1.3) the delta method may be used to obtain simple approximations to the posterior moments of $\phi$ is terms of those of $\omega$, namely

$$\mu_\phi \approx \phi\{\mu_\omega\} + \sigma_\omega^2\, \phi''\{\mu_\omega\}/2\,, \tag{39}$$

$$\sigma_\phi^2 \approx \sigma_\omega^2\, [\,\phi'\{\mu_\omega\}]^2. \tag{40}$$

Substitution into (38) and (37) respectively provide useful approximations to the intrinsic point estimator of $\omega$, and to the expected loss from using $\tilde{\omega}$ as a proxy for $\omega$.

In the particular case of the normal variance, it is convenient to start from the parametrization in terms of the precision $\lambda = \sigma^{-2}$, whose posterior moments have simple expressions. Since reference priors are consistent under reparametrization, the reference prior for $\lambda$ is $\pi(\lambda) = \pi(\sigma)|\partial\sigma/\partial\lambda| \propto \lambda^{-1}$ and, therefore, a reference parametrization is

$$\phi = \phi(\lambda) = \int^{\lambda} \pi(\lambda)\, d\lambda = \int^{\lambda} \lambda^{-1}\, d\lambda = \log \lambda.$$

Notice that the reference prior of $\phi = \log\lambda$ is indeed uniform, as it is the case for the logarithm of any other power of $\sigma$. Using (39) and (40) with the first posterior moments of $\lambda$, given in (24), yields

$$\tilde{\phi}_{int}(\boldsymbol{x}) \approx \mu_\phi = \mathrm{E}[\log\lambda \,|\, \boldsymbol{x}] \approx \log\Big(\frac{n-1}{n\,s^2}\Big) - \frac{1}{n-1}\,, \tag{41}$$

$$\sigma_\phi^2 = \mathrm{Var}[\log\lambda \,|\, \boldsymbol{x}] \approx \frac{2}{n-1}\,\cdot \tag{42}$$

By invariance, (41) directly provides an approximation to the intrinsic estimator of the variance. This has the form of a modified version of the conventional unbiased estimator $\tilde{\sigma}^2_{n-1}$; indeed, since $\sigma^2 = e^{-\phi}$,

$$\tilde{\sigma}^2_{int}(\boldsymbol{x}) = e^{-\tilde{\phi}_{int}(\boldsymbol{x})} \approx \frac{n\, s^2}{n-1}\; e^{\frac{1}{n-1}} = \tilde{\sigma}^2_{n-1}\, e^{\frac{1}{n-1}},$$

which, as shown in the third row of Table 1, provides good approximations, even for small values of $n$.

A better analytical approximation to the intrinsic estimator of the normal variance may be obtained making use of the particular features of this example. This is done by separately minimizing the expected value of each of the two functions which enter the definition of the intrinsic discrepancy loss $\delta_x\{\phi\}$, and using the arithmetic mean of the corresponding results.

Indeed, the delta method may be used to approximate both $\mathrm{E}[g(c\tau/n)]$ and $\mathrm{E}[g(n/(c\tau))]$ in terms $\mathrm{E}[\tau\,|\,n] = n-1$ and $\mathrm{Var}[\tau\,|\,n] = 2(n-1)$. The approximation to $\mathrm{E}[g(c\tau/n)]$ is minimized by $\hat{c}^*_{1n} = n/(n-1)$, while the approximation to $\mathrm{E}[g(n/(c\tau))]$ is minimized by by $\hat{c}^*_{2n} = n(n+1)/(n-1)^2$. As one would expect, their average,

$$\hat{c}^*_n = \frac{\hat{c}^*_{1n} + \hat{c}^*_{2n}}{2} = \Big(\frac{n}{n-1}\Big)^2 = \frac{n}{n-(2-n^{-1})}\,, \tag{43}$$

provides a good approximation to the value $c^*_n$ which minimizes (35). As shown in the second row of Table 1, the approximation remains good even for small values of $n$.
Combination of (36) and (43) establishes that, for all but very small $n$ values,

$$\tilde{\sigma}^2_{int}(\boldsymbol{x}) = \tilde{\sigma}^2_{int}(s,n) \approx \Big(\frac{n}{n-1}\Big)^2\, s^2. \tag{44}$$

In view of the second expression for $\hat{c}^*_n$ in (43), a cruder approximation is given by $\tilde{\sigma}^2_{int} \approx \tilde{\sigma}^2_{n-2}$. This is larger than the MLE $\hat{\sigma}^2 = s^2$ (which divides by $n$ the sum of squares), and also larger than the conventional unbiased estimate of the variance $\tilde{\sigma}^2_{n-1}$ (which divides by $n-1$). Notice that numerical differences between intrinsic and conventional estimators may be large for small values of $n$. In particular, with only two observations $\{x_1, x_2\}$, the intrinsic estimator of the variance is $\sigma^2_{int}(2, s^2) \approx 5\, s^2 = 5(x_1 - x_2)^2/4$; this is 2.5 times larger than the unbiased estimator, $(x_1 - x_2)^2/2$ in this case, which (with good reason) is generally considered to be too small.

As shown by (35), the expected intrinsic loss $d(\tilde{\sigma}^2\,|\,n, s^2)$ of any affine equivariant estimator of the variance $\tilde{\sigma}^2 = k_n s^2$, is actually independent of $s^2$ and only depends on the sample size $n$. Moreover, it is easily verified that the expected intrinsic loss $d(\tilde{\sigma}^2\,|\,n, s^2)$ is precisely equal to the *frequentist risk* associated to the intrinsic discrepancy loss,

$$r(\tilde{\sigma}^2_i\,|\,n, \sigma^2) = \int_0^\infty \delta\{\tilde{\sigma}^2_i(s^2), \sigma^2\}\, p(s^2\,|\,n, \sigma^2)\, ds^2.$$

Thus, under intrinsic discrepancy loss, the intrinsic estimator $\tilde{\sigma}^2_{int}$ dominates all affine equivariant estimators. For details, see Bernardo (2006).

Region estimation is now considered. As described in Example 4, the intrinsic $p$-credible region for $\sigma$ is the unique solution $C_p^{int} = \{\sigma_0, \sigma_1\}$ to the equations system

$$\Big\{ d(\sigma_0^2 \,|\, \boldsymbol{x}) = d(\sigma_1^2 \,|\, \boldsymbol{x}), \quad \int_{\sigma_0}^{\sigma_1} \pi(\sigma \,|\, \boldsymbol{x})\, d\sigma = p \Big\}.$$

Using (34), this may equivalently be written in terms of $\tau = n\, s^2/\sigma^2$ as

$$\Big\{ d(\sigma_0^2 \,|\, \boldsymbol{x}) = d(\sigma_1^2 \,|\, \boldsymbol{x}), \quad \int_{ns^2/\sigma_1}^{ns^2/\sigma_0^2} \chi(\tau \,|\, n-1)\, d\tau = p \Big\}. \tag{45}$$

Thus, the unique $p$-credible intrinsic region for $\sigma^2$ is the interval

$$C_p^{int}(\boldsymbol{x}) = \Big\{ \frac{n\, s^2}{\chi^2_{n-1}(1-\alpha)} \,,\, \frac{n\, s^2}{\chi^2_{n-1}(1-p-\alpha)} \Big\} \tag{46}$$

where $\chi^2_{n-1}(q)$ is the $q$ quantile of a $\chi^2_{n-1}$ distribution, and $\alpha$ is the solution to the equation

$$d\Big( \frac{n\, s^2}{\chi^2_{n-1}(1-\alpha)} \,|\, \boldsymbol{x} \Big) = d\Big( \frac{n\, s^2}{\chi^2_{n-1}(1-p-\alpha)} \,|\, \boldsymbol{x} \Big). \tag{47}$$

By invariance, this provides the intrinsic $p$-credible region of any one-to-one function of $\sigma^2$.

As a function of the data $\boldsymbol{x}$, the sampling distribution of $n\, s^2/\sigma^2$ is also a $\chi^2$ with $n-1$ degrees of freedom. Hence, for all sample sizes, the expected coverage under sampling of the $p$-credible intervals (46) is *exactly* $p$.

Using (35) to evaluate expected losses, the exact solution to equation (47) may easily be obtained by numerical methods. However, good analytical approximations are may be obtained.

Working again in terms of the reference parametrization for this problem, $\phi = \log \lambda = -2 \log \sigma$, and using (37), (39) and (40), the expected loss from using $\tilde{\phi}$ as a proxy for $\phi$ is approximately

$$d(\tilde{\phi} \,|\, \boldsymbol{x}) \approx \frac{n}{2} \left[ \frac{2}{n-1} + (\, \tilde{\phi}_{int} - \tilde{\phi}\, )^2 \right]. \tag{48}$$

But this is symmetric around $\tilde{\phi}_{int} = \log(\tilde{\lambda}_{int}) = -\log(\sigma^2_{int})$ and therefore, to keep those $\tilde{\phi}$ points with smaller expected loss, any intrinsic credible region for $\phi = \log \lambda$ must be (approximately) centered at $\tilde{\phi}_{int}$. Thus, using (42) and (44) this will be of the form

$$C_p^{int}(\boldsymbol{x},\Phi) \approx \tilde{\phi}_{int} \pm \alpha_{pn}\sigma_\phi \approx \log\Big[\Big(\frac{n-1}{n}\Big)^2 \frac{1}{s^2}\Big] \pm \gamma_{pn}\sqrt{\frac{2}{n-1}} \tag{49}$$

where $\gamma_{pn}$ is the solution to the equation

$$\int_{\tilde{\phi}_{int}-\gamma_{pn}\,\sigma_\phi}^{\tilde{\phi}_{int}+\gamma_{pn}\,\sigma_\phi} \pi(\phi\,|\,\boldsymbol{x})\,d\phi = p,$$

or, equivalently since

$$\begin{aligned} \tau = n\,s^2\lambda &= n\,s^2 e^{\phi}, \\ ns^2\,e^{\tilde{\phi}_{int}\pm\gamma_{pn}\,\sigma_\phi} &= \frac{(n-1)^2}{n}\,\exp\Big[\,\pm\gamma_{pn}\sqrt{\frac{2}{n-1}}\,\Big], \\ \pi(\tau\,|\,\boldsymbol{x}) &= \chi^2(\tau\,|\,n-1), \end{aligned}$$

$\gamma_{pn}$ is the unique solution to the equation

$$F_{n-1}\Big\{\frac{(n-1)^2}{n}\,e^{+\gamma_{pn}\sqrt{\frac{2}{n-1}}}\Big\} - F_{n-1}\Big\{\frac{(n-1)^2}{n}\,e^{-\gamma_{pn}\sqrt{\frac{2}{n-1}}}\Big\} = p, \tag{50}$$

where $F_\nu$ is the cumulative distribution function of a $\chi^2_\nu$ distribution.

A numerical solution to (50) is immediately found with standard statistical software. However, a simple analytical approximation may be derived using the fact that the reference posterior distribution of $\phi = \log\lambda$ becomes approximately normal (at a faster rate that any other simple function of $\sigma$) as the sample size $n$ increases. Using this approximation and (49), the $p$-credible intrinsic region for $\phi$ is approximated by the interval

$$C_p^{int}(\boldsymbol{x},\Phi) \approx \log\Big[\Big(\frac{n-1}{n}\Big)^2 \frac{1}{s^2}\Big] \pm q_p\sqrt{\frac{2}{n-1}} \tag{51}$$

where $q_p$ is the $(p+1)/2$ quantile of a standard normal distribution. By invariance, the $p$-credible intrinsic region for the variance $\sigma^2 = e^{-\phi}$ will be approximated by

$$\Big\{\, s^2\Big(\frac{n}{n-1}\Big)^2\, e^{-\gamma_{np}\sqrt{\frac{2}{n-1}}},\;\; s^2\Big(\frac{n}{n-1}\Big)^2\, e^{+\gamma_{np}\sqrt{\frac{2}{n-1}}}\,\Big\} \tag{52}$$

where $\gamma_{np}$ is the solution to (50) which, as $n$ increases, converges to $q_p$, the $(p+1)/2$ quantile of an standard normal distribution.

Summarizing, we have thus established

**Theorem 3** ***(Intrinsic estimation of the normal variance)*** *Let* $\boldsymbol{x}$ *be a random sample of size n from* $\mathrm{N}(x\,|\,\mu,\sigma)$, *with variance* $s^2$.

*(i) The intrinsic point estimator $\tilde{\sigma}^2_{int}(\boldsymbol{x})$ of $\sigma^2$ is*

$$\begin{aligned}
\tilde{\sigma}^2_{int}(\boldsymbol{x}) &= \arg\min_{\tilde{\sigma}>0} d(\tilde{\sigma}^2 \,|\, \boldsymbol{x}),\\
d(\tilde{\sigma}^2 \,|\, \boldsymbol{x}) &= \frac{n}{2}\int_0^\infty \delta\left(\frac{\tilde{\sigma}^2\,\tau}{n\,s^2}\right)\chi^2(\tau \,|\, n-1)\,d\tau,\\
\delta\{\theta\} &= \min\{g(\theta), g(1/\theta)\}, \quad g(\theta) = (\theta-1) - \log\theta,\\
\tilde{\sigma}^2_{int}(\boldsymbol{x}) &\approx \left(\frac{n}{n-1}\right)^2 s^2.
\end{aligned}$$

*The intrinsic point estimator $\tilde{\sigma}_{int}(\boldsymbol{x})$ is the Bayes estimator with respect to the intrinsic discrepancy loss. Besides, it has smaller frequentist risk with respect to this loss than any other affine equivariant estimator.*

*(ii) The unique p-credible intrinsic region $C^{int}_p(\boldsymbol{x})$ for $\sigma^2$ is the interval*

$$C^{int}_p(\boldsymbol{x}) = \{a(\alpha,\boldsymbol{x}),\ b(\alpha,p,\boldsymbol{x})\} = \left\{\frac{n\,s^2}{\chi^2_{n-1}(1-\alpha)}\ ,\ \frac{n\,s^2}{\chi^2_{n-1}(1-p-\alpha)}\right\},$$

*where $\chi^2_{n-1}(q)$ is the q quantile of a $\chi^2_{n-1}$ distribution, and $\alpha$ is the solution to the equation $d\{a(\alpha,\boldsymbol{x}) \,|\, \boldsymbol{x})\} = d\{b(\alpha,p,\boldsymbol{x}) \,|\, \boldsymbol{x})\}$. For all sample sizes, the frequentist coverage of $C^{int}_p(\boldsymbol{x})$ is exactly p. Moreover,*

$$C^{int}_p(\boldsymbol{x}) \approx s^2\left(\frac{n}{n-1}\right)^2\left\{e^{-\gamma_{np}\sqrt{\frac{2}{n-1}}},\ e^{+\gamma_{np}\sqrt{\frac{2}{n-1}}}\right\}$$

*where $\gamma_{np}$ is the solution to the equation*

$$F_{n-1}\left\{\frac{(n-1)^2}{n}\,e^{+\gamma_{np}\sqrt{\frac{2}{n-1}}}\right\} - F_{n-1}\left\{\frac{(n-1)^2}{n}\,e^{-\gamma_{np}\sqrt{\frac{2}{n-1}}}\right\} = p.$$

*and $F_\nu$ is the cumulative distribution function of a $\chi^2_\nu$ distribution. As n increases, $\gamma_{np}$ converges to the $(p+1)/2$ normal quantile.*

*(iii) The expected intrinsic loss associated to the use of $\tilde{\sigma}^2$ is*

$$d(\tilde{\sigma}^2 \,|\, s^2, n) \approx \frac{n}{2}\left[\frac{2}{n-1} + \left(\log\frac{1}{\sigma^2_{int}} - \log\frac{1}{\tilde{\sigma}^2}\right)^2\right],$$

*with $\tilde{\sigma}^2_{int}(\boldsymbol{x}) \approx n^2\,s^2/(n-1)^2$.*

For the numerical illustration considered in Example 4 (where the sample size was only $n = 12$), the approximation (44) to the intrinsic estimate of $\sigma^2$ yields $\tilde{\sigma}^2_{int} \approx 5.104$. The approximation (52) to the intrinsic 0.90-credible region yields $(2.480, 11.507)$ using the exact solution $\gamma_{np} = 1.693$ to equation (50), and $(2.531, 11.272)$ using the

corresponding normal approximation $\gamma_{np} \approx q_{0.90} = 1.645$. These approximations may be compared eith the exact values $\tilde{\sigma}^2_{int} = 5.090$ and $C^{int}_{0.90} = (2.481, 11.717)$ numerically found in Example 4.

### 3.4 Intrinsic estimation of the Cauchy scale parameter

We finally consider an example where no analytical expressions are possible. With the use of increasingly complex statistical models, this is fast becoming the norm, rather than the exception, in statistical practice.

Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from a Cauchy distribution $\mathrm{Ca}(x \,|\, 0, \sigma)$, centered at zero with unknown scale parameter $\sigma$, so that the likelihood function is

$$p(\boldsymbol{x} \,|\, \sigma) = \prod_{i=1}^{n} \mathrm{Ca}(x_j \,|\, 0, \sigma) \propto \sigma^{-n} \prod_{i=1}^{n} \left(1 + \frac{x_j^2}{\sigma^2}\right)^{-1}.$$

The Cauchy distribution does not belong to the exponential family and, therefore, there is no sufficient statistic of finite dimension. There is no analytical expression for the MLE $\hat{\sigma}$ of the unknown parameter. Fisher information function is $n/(2\sigma^2)$ and, therefore, the posterior distribution of $\sigma$ will be asymptotically normal, $\mathrm{N}(\sigma \,|\, \hat{\sigma},\ \sqrt{2}\,\hat{\sigma}/\sqrt{n})$.

Since this is a scale model, the reference prior is $\pi(\sigma) = \sigma^{-1}$ and, using Bayes theorem, the reference posterior is

$$\pi(\sigma \,|\, \boldsymbol{x}) = \frac{\sigma^{-1} p(\boldsymbol{x} \,|\, \sigma)}{\int_0^\infty \sigma^{-1} p(\boldsymbol{x} \,|\, \sigma)\, d\sigma}\,, \tag{53}$$

which may easily be numerically computed. It may be verified that, provided the data $\boldsymbol{x}$ contain at least two different observations, $\pi(\sigma \,|\, \boldsymbol{x})$ has a gamma-like shape with a unique mode.

Figure 4 represents the reference posteriors of $\sigma$ which correspond to a set of 25 random samples of size $n = 12$, which were all generated from a Cauchy distribution $\mathrm{Ca}(x \,|\, 0, 2)$. This may be seen as a graphical representation of the *sampling distribution of* $\pi_\sigma(\cdot \,|\, \boldsymbol{x})$, the reference posterior of $\sigma$, given $\sigma = 2$ and $n = 12$. Notice that, although all these posteriors contain indeed the true value $\sigma = 2$ from which the samples have been simulated (marked in the figure with a solid dot), the variability is very large.

The logarithmic divergence $\kappa\{\sigma_2 \,|\, \sigma_1\}$ of $\mathrm{Ca}(x \,|\, 0, \sigma_2)$ from $\mathrm{Ca}(x \,|\, 0, \sigma_1)$ is

$$\int_{-\infty}^{\infty} \mathrm{Ca}(x \,|\, 0, \sigma_1) \log \frac{\mathrm{Ca}(x \,|\, 0, \sigma_1)}{\mathrm{Ca}(x \,|\, 0, \sigma_2)}\, dx = \log \frac{1}{4\,\sigma_1\,\sigma_2} + 2 \log(\sigma_1 + \sigma_2).$$

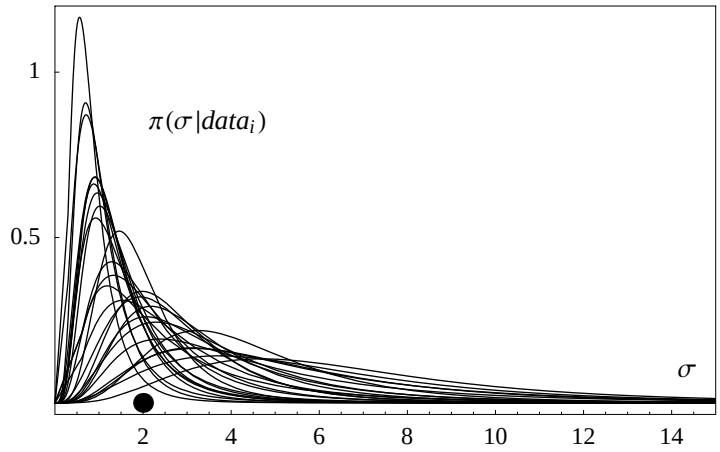


***Figure 4:*** *Reference posterior distributions of $\sigma$ for a set of 25 random samples of size $n = 5$ generated from a $\mathrm{Ca}(x\,|\,0, 2)$ distribution.*

Since, in this case, $\kappa\{\sigma_2\,|\,\sigma_1\} = \kappa\{\sigma_1\,|\,\sigma_2\}$, the intrinsic discrepancy loss from using $\tilde{\sigma}$ as a proxy fo $\sigma$ is (Def. 2)

$$\delta\{\tilde{\sigma}, \sigma\} = \log \frac{1}{4\,\tilde{\sigma}\,\sigma} + 2\log(\tilde{\sigma} + \sigma) = \log \frac{1}{4\,\sqrt{\phi}} + \log(1 + \sqrt{\phi}), \tag{54}$$

where $\phi = \tilde{\sigma}^2/\sigma^2$. Thus, as in the normal case (Eq. 33), the intrinsic discrepancy loss only depends on the variance ratio $\phi = \tilde{\sigma}^2/\sigma^2$.

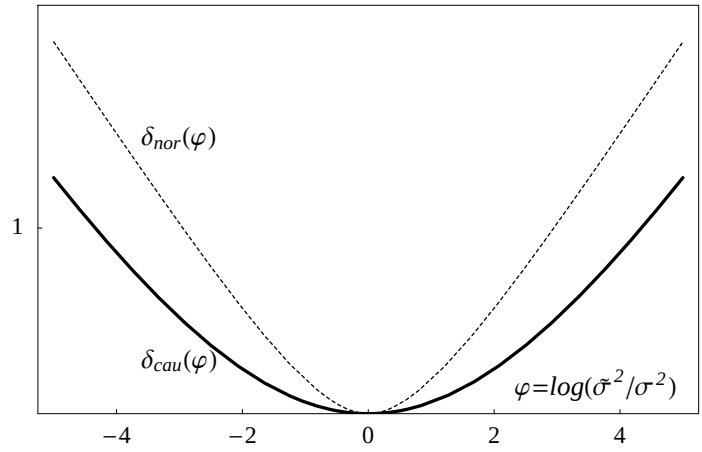


***Figure 5:*** *Intrinsic discrepancy loss from using $\tilde{\sigma}$ as a proxy for $\sigma$ as a function of $\psi = \log(\tilde{\sigma}^2/\sigma^2)$ for Cauchy (solid line) and normal (dotted line) distributions.*

Figure 5 provides a direct comparison between the intrinsic discrepancy loss for the scale parameter in the Cauchy and in the normal case. As one might expect, for any

given value of the ratio $\phi = \tilde{\sigma}^2/\sigma^2$, the intrinsic loss is smaller in the Cauchy case than it is in the normal case.

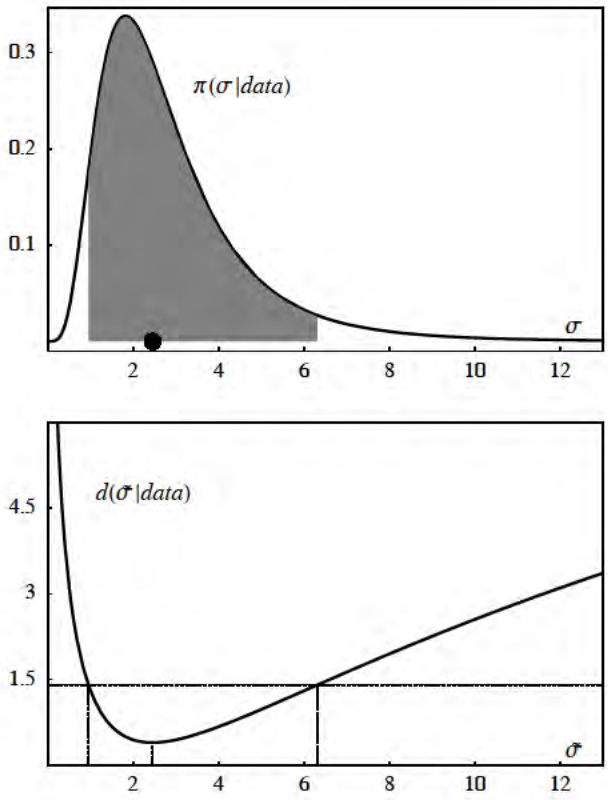


***Figure 6:*** *Reference posterior density (upper panel) and intrinsic expected loss (lower panel) for the scale parameter $\sigma$ of a Cauchy* $\mathrm{Ca}(x\,|\,0,\sigma)$ *distribution, given* $\boldsymbol{x} = \{-1.78, -0.75, -2.44, -3.30, 8.48\}$. *The intrinsic estimator is* $\tilde{\sigma}_{int} = 2.452$ *(solid dot) and the intrinsic* 0.90*-credible region is* $C^{int}_{0.90} = (0.952, 6.314)$ *(shaded region).*

The intrinsic expected loss from using $\tilde{\sigma}$ is the reference posterior expectation of the intrinsic discrepancy loss,

$$d(\tilde{\sigma}\,|\,\boldsymbol{x}) = n \int_0^\infty \delta\{\tilde{\sigma}, \sigma\}\, \pi(\sigma\,|\,\boldsymbol{x})\, d\sigma,$$

where $\delta\{\tilde{\sigma}, \sigma\}$ and $\pi(\sigma\,|\,\boldsymbol{x})$ are respectively given by (54) and (53), and may easily be computed by numerical integration. The intrinsic estimator of $\sigma$ is

$$\tilde{\sigma}_{int}(\boldsymbol{x}) = \arg\inf_{\tilde{\sigma}>0} d(\tilde{\sigma}\,|\,\boldsymbol{x})$$

and the $p$-credible intrinsic region is the solution $C_p^{int}(\boldsymbol{x}) = (\sigma_0, \sigma_1)$ to the equations system

$$\left\{ d(\sigma_0 \,|\, \boldsymbol{x}) = d(\sigma_1 \,|\, \boldsymbol{x}), \quad \int_{\sigma_0}^{\sigma_1} \pi(\sigma \,|\, \boldsymbol{x})\, d\sigma = p \right\}.$$

As a numerical illustration, a random sample of size $n = 5$ was generated from a Cauchy $\text{Ca}(x \,|\, 0, 2)$, yielding $\boldsymbol{x} = \{-1.78, -0.75, -2.44, -3.30, 8.48\}$. The results from the reference analysis of this data set are encapsulated in Figure 6. The reference posterior distribution $\pi(\sigma \,|\, \boldsymbol{x})$ is represented in the upper panel, and the expected intrinsic loss $d(\tilde{\sigma} \,|\, \boldsymbol{x})$ from using $\tilde{\sigma}$ as a proxy for $\sigma$ is represented in the lower panel. The intrinsic estimator, represented by a solid dot, is $\tilde{\sigma}_{int} = 2.452$, and the intrinsic 0.90-credible interval, represented by a shaded region, is $C_{0.90}^{int} = (0.952, 6.314)$.

Neither exact Bayesian credible regions nor exact frequentist confidence intervals may be analytically obtained in this problem. The frequentist coverage of the intrinsic credible regions was however analyzed by simulation. A set of $5,000$ random samples of size $n = 5$ were generated from a Cauchy $\text{Ca}(x \,|\, 0, 2)$, and their corresponding intrinsic 0.90-credible intervals were computed; it was then found that the proportion of those intervals which contained the true value $\sigma = 2$ was 0.905. With $25,000$ random samples this proportion was 0.902. This suggests that (as in the normal case) the expected frequentist coverage of reference $p$-credible regions, the limit of this algorithm as the number of generated random samples increases, is exactly $p$. To further explore this suggestion, a set of $10,000$ random samples of size $n$ were generated from a Cauchy distribution $\text{Ca}(x \,|\, , 0, \sigma)$ for each of several combinations $\{n, \sigma\}$ of sample size $n$ and true value $\sigma$ of the scale parameter and the corresponding intrinsic $p$-credible regions where computed for $p = 0.90$ and $p = 0.95$. Table 2 describes the proportion of these regions which actually contained the value of $\sigma$ from which the samples had been generated.

***Table 2:*** *Proportions of intrinsic $p$-credible intervals which contained the true value of $\sigma$ among $10,000$ random samples generated from each of several combinations of sample size $n$ and true value of $\sigma$.*

| | $p = 0.90$ | | | $p = 0.95$ | | |
|---|---|---|---|---|---|---|
| | $n$ | | | $n$ | | |
| $\sigma$ | 2 | 12 | 30 | 2 | 12 | 30 |
| 0.5 | 0.9002 | 0.9044 | 0.8999 | 0.9490 | 0.9491 | 0.9507 |
| 2.0 | 0.8971 | 0.8971 | 0.9003 | 0.9467 | 0.9517 | 0.9490 |
| 4.0 | 0.9006 | 0.8960 | 0.8990 | 0.9484 | 0.9497 | 0.9507 |

Examination of this table provides strong statistical evidence that the frequentist coverage of reference $p$-credible regions is indeed exactly equal to $p$ for all sample sizes. Indeed, treating each simulation as a Bernoulli trial, the reference posterior

distribution of the frequentist coverage $\theta_{ij}$ which corresponds to the $(i, j)$ cell is approximately normal with mean observed proportion $p_{ij}$ quoted in the table, and standard deviation $(0.90 * 0.10/10000)^{1/2} = 0.0030$ for the 0.90-credible intervals, and $(0.95 * 0.05/10000)^{1/2} = 0.0022$ for the 0.95-credible intervals. This makes the respective nominal values 0.90 and 0.95 clearly compatible with the observed results. Notice that this is *not* an asymptotic analysis, as in probability matching theory (Datta and Sweeting, 2005), for it even applies to the smallest possible samples, those with $n = 2$.

## References


Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 35-60 (with discussion).

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical*, 41, 113-147 (with discussion). Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.) Brookfield, VT: Edward Elgar, 1995, 229-263.

Bernardo, J. M. (2005a). Reference analysis. *Handbook of Statistics* 25 (D. K. Dey and C. R. Rao, eds.). Amsterdam: Elsevier, 17-90.

Bernardo, J. M. (2005b). Intrinsic credible regions: An objective Bayesian approach to interval estimation. *Test*, 14, 317-384 (with discussion).

Bernardo, J. M. (2006). Intrinsic point estimation of the normal variance. *Bayesian Statistics and its Applications* (S. K. Upadhyay, U. Singh and D. K. Dey, eds.) New Delhi: Anamaya, 110-121.

Bernardo, J. M. and Juárez, M. (2003). Intrinsic estimation. *Bayesian Statistics*, 7, 465-476.

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

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley. (2nd edition in preparation).

Brown, L. (1968). Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters. *The Annals of Mathematical Statistics* ,39, 24-48.

Brewster, J. F. and Zidek, J. V. (1974). Improving on equivariant estimators. *Annals of Statistics*, 2, 21-38.

Brown, L. (1990). Comment on Maata and Casella (1990).

Datta, G. S. and Sweeting, T. J. (2005). Probability matching priors. *Handbook of Statistics*, 25 (D. K. Dey and C. R. Rao, eds.). Amsterdam: Elsevier, 91-114.

Fernández, C. and Steel, M. F. J. (1999). Reference priors for the general location-scale model. *Statistics and Probability Letters*, 43, 377-384.

Juárez, M. A. (2004). *Métodos Bayesianos Objetivos de Estimación y Contraste de Hipótesis*. Ph.D. Thesis, Universidad de Valencia, Spain.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 79-86.

Maata, J. M. and Casella, G. (1990). Developments in decision-theoretic variance estimation. *Statistical Science*, 5,90-120, (with discussion).

Robert, C. P. (1996). Intrinsic loss functions. *Theory and Decision*, 40, 192-214.

Rukhin, A. L. (1987). How much better are better estimators of a normal variance. *Journal of The American Statistical Association*, 82, 925-928.

Schervish, M. J. (1995). *Theory of Statistics*. Berlin: Springer

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Annals of the Institute of Statistical Mathematics*, 16, 155-160.

# Discussion of “Objective bayesian point and region estimation in location-scale models” by José M. Bernardo

**Miguel A. Gómez Villegas**

Departamento de Estadística e Investigación Operativa

Universidad Complutense de Madrid

Let me begin by congratulating Professor Bernardo for his excellent job in objective Bayesian analysis. This paper, and the closely related Bernardo (2005), present a unified theory of estimation by point and credible regions based on information ideas he has used previously to define reference priors. The idea originates from the study of both problems as decision problems, where the loss function is the "intrinsic discrepancy" inspired in the Kullblack-Leibler divergence, and defined as the minimum of $k_x\{\widetilde{\theta}, \widetilde{\lambda}|\theta, \lambda\}$ and $k_x\{\theta, \lambda|\widetilde{\theta}, \widetilde{\lambda}\}$ where

$$k_x\{\widetilde{\theta}, \widetilde{\lambda}|\theta, \lambda\} = \int_{\chi(\theta,\lambda)} \pi(x|\theta, \lambda)\ ln\frac{\pi(x|\theta, \lambda)}{\pi(x|\widetilde{\theta}, \widetilde{\lambda})}dx$$

An intrinsic point estimator is then defined as the Bayes estimator which corresponds to the intrinsic loss and the appropriate reference prior. A $p$-credible intrinsic region estimator is defined as the lowest posterior loss $p$-credible with respect to the intrinsic loss and the appropriate reference prior.

A first question is: do we need to employ

$$\int_{C_p^{int}} \pi(\theta|x)d\theta \geq p$$

with the inequality instead of equality to allow the discrete case?

Second, it would be useful to have a better understanding of the proposed approach to applying these ideas to the exponential distribution family instead of location-scale models; this is a family of distributions greater than the other.

Professor Bernardo claims that in one-dimensional problems, one may define probability centred credible intervals, and these are invariant under reparametrization. Will it not be necessary to suppose that the transformation is monotonic?

Third, on a more philosophical basis, I think that invariance is a compelling argument for point estimations and for credible regions. Indeed both point estimations and credible regions are two answers to the same question: how we can eliminate the uncertainty

about $\theta$. Bernardo's approach permits one to obtain invariance under reparametrization in both problems.

Fourth, the chosen examples show the coherence between frequentist inference and Bayesian inference. When intrinsic credible regions that require minimal subjective inputs are employed, exact frequentist confidence regions are obtained, at least in the normal mean and variance. This fact is similar to the one obtained by this discussant in Gómez-Villegas and González-Pérez (2005) and references therein. I wonder if Professor Bernardo has any idea about the essential reasons behind the matching properties between intrinsic credible regions and confidence regions in these cases?

Fifth, adopting this approach to credible set construction, I see problems in computations, the posterior intrinsic loss integrated over a large dimensional space. From the point of view of applications, a simple asymptotic approximation to normality should be necessary.

In closing, I would like to thank the editor of the journal for giving me the opportunity to discuss this paper.

## References

Bernardo, J. M. (2005). Intrinsic credible regions: an objective Bayesian approach to interval estimation. *Test*, 14, 317-384.

Gómez-Villegas, M. A. and González-Pérez, B. (2005). Bayesian analysis of contingency tables. *Communications in Statistics-Theory and Methods*, 34, 1743-1754.

**Dennis V. Lindley**

ThomBayes@aol.com

Two concepts are basic to the ideas of this excellent paper: *objectivity* and the concept of estimation as a *decision* problem. In the author's skilful hands, these lead to reference priors and intrinsic loss functions, and hence, by minimizing expected loss, to estimates which are often superior to the conventional ones. It can be said with some confidence that we have here a solution to the problem Harold Jeffreys first posed around 1939 of providing an objective, coherent method for scientific inference. The development employs several subtle ideas, and considerable mathematical complexity, but one feature that struck me is that the final results are usually fairly simple and look right. An example of this is provided by the loss functions in Figure 3, which have the reasonable convexity property around the true value but, unlike quadratic loss, exhibit sensible concavity at more discrepant values. I would have preferred the loss to have been bounded but, with normal distributions and their thin tails, this scarcely matters. To be bounded may be more important with the fat tails of the Cauchy in Figure 6, in order to avoid paradoxes of the St. Petersburg type. A related point is that although the mathematics can be formidable, at least in the view of some applied statisticians, once it has been done the practitioner can easily use the results in the confidence that the machinery used to produce them is sound. It is comparable to driving a car, without knowing how it was made, but having confidence in the manufacturer.

Granted the basic concepts, this is an important paper, but was Jeffreys right to search for objectivity, and was Fisher wrong in dismissing decision concepts from inference? I think Jeffreys was wrong and Fisher was right. At the risk of repeating what I have said before, it seems to me that inference and decision-making are distinct and both are subjective. In other words, the two basic concepts, that provide the foundations of this paper, are suspect.

Consider first the fixed likelihood upon which all the arguments in the paper rest. Is it really objective? There are a few cases where substantial evidence for normality exists, but often the normal, or another member of the exponential family, is used merely for mathematical simplicity. With the increased computing power available today, statisticians are less constrained and can use other distributions that appear more realistic, thereby introducing subjectivity. There are some popular data sets that have been repeatedly analysed using different likelihoods. Where is the objectivity there? It is interesting that Bernardo uses one symbol, $p$, for probabilities of data but another,

$\pi$, for probabilities of parameters. In reality both $p$ and $\pi$ reflect beliefs about data and parameters respectively, obey the same rules and do not deserve separate treatments.

Inference concerns parameters. (It is more practical to make inference about future data but I do not explore that trail here.) What are these parameters, the $\theta$ and $\lambda$ of the paper? If our statistical analyses are to be of use in data analysis, $\theta$ at least ought to relate to something in the real world. Bernardo has only one sentence about this, referring to $\theta$ as the age of the earth. Putting aside intelligent designers, reputable scientists differ in their views of the age. In other words, their ideas are subjective, so that before relevant data about the age are considered, their different views need to be included. Another relevant fact is that information about the age of the earth does not come from data with normal, or any other objective, likelihood. More conspicuous examples of subjectivity are apparent witch clinical trials, where the different views of drug companies and official bodies are consulted before the trial. This became clear recently when a trial went horribly wrong and experts claimed the probabilities used were, in their opinion, unsound. So often today, $\theta$ is regarded as nothing more than a symbol, whereas, to be of value, it has to refer to reality and hence influenced by opinions about that reality. These opinions should be incorporated into the analysis, not ignored and replaced by a reference prior, especially when this is improper.

All of us have, at some time, expressed an opinion about something without having any intention of basing any action upon that opinion. In statistics, this opinion-forming is inference and means we infer the value of the real $\theta$. In the Bayesian paradigm this is done by means of your probability distribution of $\theta$, given the data and the original information about $\theta$. Whilst it is true that any inference has to be capable of being used as a basis for action, for otherwise what use is it, it is not true that inference has to have immediate actions in mind. In particular, inference does not require a loss function, and certainly not a loss function that ignores reality. In Bayesian terms, there is only one inference, the posterior distribution and, although it may be advantageous to summarize its main features, such approximations scarcely need elaborate techniques, except in the case of many parameters.

Inference from data consists in modelling that data in the form of a likelihood depending on parameters, supplying your opinion of the parameters prior to the data, and combining likelihood and prior by Bayes theorem. Finally the nuisance aspect of the parameters is removed by integration. When several people are involved there may be disagreements over likelihood or prior. These may be removed by discussion but, if this fails, the calculations may be repeated under different subjective opinions and the posteriors compared. That science is objective is a myth. Apparent objectivity in science only arises when the data are extensive.

This paper explores a field that, in my view, is not in the broad stream of statistics. This is not to deny it great merit, for we now know what that field contains, material of real merit from which all can learn.

**Mark J. Schervish**

Carnegie Mellon University, USA

I admire Professor Bernardo for his steadfastness and resolution in staying the course of research into reference priors and other so-called objective Bayesian methods. Despite repeated attacks dating back to the discussion of Bernardo (1979) he has continually risen to the challenge of making these methods palatable to practitioners and theoreticians alike. I will not here rehearse all of the criticisms or the support for his work in this area. I refer the interested reader to the various discussions of the papers listed in the reference list to Professor Bernardo's paper. I will mention just a few problems that I have with the methods as well as what I like about them.

To begin with a positive note, I like the idea of having a transformation-equivariant estimation procedure for non-decision-theoretic inference. When one is faced with a decision problem in which a specific loss function is relevant, then one does not care whether one's inference satisfies an ad hoc criterion such as transformation equivariance. On the other hand, when one merely wishes to report an estimate of some quantity, especially the parameter of a statistical model which most likely is a figment of one's imagination (model) anyway, then it becomes difficult to explain why the estimate of an equivalent parameter is not the equivalent estimate. Indeed, I believe that the intrinsic discrepancy loss satisfies a slightly stronger invariance than is stated in (10). I believe that one could apply a one-to-one reparameterization of the form $\phi = \phi(\theta)$ and $\psi = \psi(\lambda, \theta)$ and still achieve (10). Of course, a completely general reparameterization would change the meaning of the parameter of interest, and yet the desire for an equivariant estimate would remain.

One of the serious concerns with reference priors is their violation of the likelihood principle. The reference priors are different for binomial sampling and negative binomial sampling so that even if the observed data could have come from either sampling scheme, the posterior would depend on the sampling plan. If one were to observe a binomial sample and use the reference prior, and later observe a negative binomial sample, one would get a different inference than if one were to observe the same two samples in the other order. As mentioned earlier, various discussants have described other concerns with the methods advocated in the manuscript, and I will let the reader find them in their original forms. I will add only one other concern that I have, and that is with the use of the description of these methods as "objective". I suppose that, so long as one agrees with all of the reasons put forth for why such methods should be

used, then one will use the methods and they become objective in that sense. But any set of methods could be called objective on those grounds. One of the main strengths of Bayesian methodology is that it forces users to be explicit about the assumptions that they are making. People who think that they are using objective methods are simply borrowing a collection of subjective assumptions and ignoring the fact that choices were made by someone else arriving at those assumptions. When you lay your assumptions out for all to see, you are in a position to evaluate the sensitivity of your inferences to the assumptions. If you hide behind a cloak of objectivity, you may produce the same answer that others produce, but you have lost the ability to see what is the effect of the subjective choices that were made.

## Rejoinder

I am extremely grateful to the three discussants by their thoughtful comments. I will answer them individually.

*Gómez-Villegas*. If the parameter of interest $\boldsymbol{\theta}$ is discrete, then we would certainly need to work with regions $C$ such that $\int_C \pi(\theta \mid \boldsymbol{\theta})\, d\boldsymbol{\theta} \geq p$ since, in that special case, not all credible probabilities $p$ would be attainable. However, point and region estimation are usually done with *continuous* parameter spaces, and this is indeed the case in the location and scale models considered in this paper. In that situation, the equality may always be obtained.

The ideas discussed in the paper may certainly be applied to models in the (generalized) exponential family and it is likely that this would lead to some rather general results. I did not have time and space to do this here, but it is certainly a research line well worth exploring.

As Professor Gómez-Villegas points out, the invariance arguments invoked only refer to monotonic, one-to-one transformations of the parameter. Even though not always explicitly stated, we were indeed always assuming this to be the case.

I believe that the *exact* numerical coincidence between objective credible regions and frequentist confidence interval is the exception, not the rule; when it happens, it is the consequence of the existence of pivotal quantities, so that the reference distribution of the pivot (considered as a function of the parameter) is precisely the same as its sampling distribution (considered as a function of the data). In particular, this coincidence cannot exist if data are discrete, as in the case of binomial or Poisson data. Beyond the particular situations where pivots exist, one may only expect an asymptotic approximation: objective credible regions are typically *approximate* confidence intervals, the approximation improving with the sample size.

Routine application of the methods described in this paper will certainly require either available software producing the exact results (not difficult to write in the standard examples which constitute the vast majority of applications) and/or appropriate analytical approximations. The latter may easily be obtained, as in the examples contained in the paper, by using the normal approximation with the parametrization induced by the appropriate variance-stabilizing transformation, and then making use of the invariance properties of the procedures.

*Lindley*. I am really proud that Professor Lindley may believe that the procedures described provide an objective coherent method for scientific inference in the sense demanded by Jeffreys, and I am very grateful for that comment.

It would certainly be better from a foundations viewpoint if the expected loss were bounded, but information measures with continuous parameters are are not bounded (one needs infinite amount of information to know precisely a real number) and yet have all kind of attractive properties.

To repeat in print the basics of an argument that Professor Lindley and I have often had in private conversations,

(i) I believe, with Jeffreys, that Fisher was wrong in dismissing decision concepts in inference. If, by some reason, you must choose an estimate, then (whether you like it or not) you have a well posed decision problem where the action space is the set of parameter values; then foundations dictate that (to act rationally) you *must* use a loss function. For instance, in one continuous parameter problems, the median may well be an estimate with good robustness properties, but the fact remains that this would be a good estimate if (*and only if*) your loss function is well approximated by a linear, symmetric loss function.

(ii) I applaud the use of subjective priors when the problem is simple and small enough for the required probability assessments to be feasible (which is *not* frequent). But, even in this case, there is no reason while other people should necessarily accept a subjective prior which goes beyond clearly stated assumptions and verifiable (possibly historical) data. There is a clear need for some commonly accepted minimum set of conclusions to be solely derived from assumptions and data, and this is precisely what reference posteriors provide. As their name indicate, they are proposed as a *reference*, to be compared with subjective posteriors when these are available. This is part of a necessary exercise in sensitivity analysis, by making explicit which parts of the conclusions depend on a particular subjective prior, and which parts are implied by the model assumed and the data obtained.

As Professor Lindley points out, although inferential statements are typically used as a basis for action, there are many situations were inferences are to be drawn without any specific action in mind. This is precisely why we suggest the use of an the information-based loss function. If a particular action is in mind, one should certainly use a context dependent loss function which appropriately describes the decision problem analyzed. It no particular decision problem is in mind, one is bound to use some conventional loss function. We have argued that conventional loss functions (such as the ubiquitous quadratic loss) are often unsatisfactory. Instead, for "pure inference" problems one should try to minimize the information loss due to the use of an estimate of the unknown parameter value; and this, I believe, is appropriately captured by the intrinsic discrepancy loss.

*Schervish*. I am very glad to read that Professor Schervish appreciates the importance of invariant procedures. In teaching, I often start my lectures by stating that any inferential

procedure which is not invariant under monotonic transformations of the parameter is suspect, and go on to provide a set of examples of those as "counterexamples" to common statistical procedures.

I agree with Professor Schervish on the importance of the likelihood principle, but I believe that the principle is actually compatible with a sensible use of reference distributions. Indeed, a reference posterior encapsulates, by definition, the (minimal) inferential statements you could proclaim about the parameter of a model *if* your prior was that maximizing the information that data generated from that *particular* model could possibly provide. If you change the model (even if the new model induces a proportional likelihood function), you change the reference prior. Thus, different reference posteriors corresponding to different sampling schemes with Bernoulli observations provide a collection of *conditional* answers (one for each sampling scheme one is willing to consider), which may all be part of the sensitivity analysis to changes in the prior mentioned above.

Objectivity is indeed an emotionally charged word, and it should be explicitly qualified whenever it is used. No statistical analysis is seriously objective, if only because the choice of both the experiment design and the model used have typically very strong subjective inputs. However, the frequentist paradigm is sold as "objective" just because its conclusions are only conditional on the model assumed and the data obtained, and this objectivity illusion has historically helped frequentist to keep a large share of the statistics market. I claim for the procedures described in this paper the right to use "objective" in precisely the same sense: these are procedures which are only conditional on the assumed model and the observed data. The use of the word "objective" in this precise, limited sense may benefit, I believe, the propagation of the Bayesian paradigm. For a recent discussion of this and related issues see Berger (2006) and ensuing discussion.

I fully agree with Professor Schervish on the paramount importance of clearly presenting the assumptions needed for an inferential statement. In the case of reference posteriors this should typically read as a *conditional* statement of the form: "*If* available data $\boldsymbol{x}$ had been generated by model $\mathcal{M} \equiv \{p_x(\cdot \mid \boldsymbol{\omega}),\ \boldsymbol{\omega} \in \Omega\}$ *and* prior information about $\boldsymbol{\theta}(\boldsymbol{\omega})$ were minimal with respect to the information about $\boldsymbol{\theta}(\boldsymbol{\omega})$ that repeated sampling from $\mathcal{M}$ could possibly provide *then*, the marginal reference posterior $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ encapsulates what could be said about the value of $\boldsymbol{\theta}$, solely on the basis of that information".

## References

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1, 385-402 and 457-464 (with discussion).

SORT 31 (1) January-June 2007, 45-54

# Goodness of fit tests for the skew-Laplace distribution

Pedro Puig[1] and Michael A. Stephens[2]

[1] *Departament de Matemàtiques, Universitat Autònoma de Barcelona,*
[2]*Department of Statistics & Actuarial Science, Simon Fraser University, B.C. Canada*

**Abstract**

The skew-Laplace distribution is frequently used to fit the logarithm of particle sizes and it is also used in Economics, Engineering, Finance and Biology. We show the Anderson-Darling and Cramér-von Mises goodness of fit tests for this distribution.

*MSC:* 62G10, 62G30

*Keywords:* Maximum likelihood estimator; non-regular models; Anderson-Darling statistic; Cramér-von Mises statistic

## 1 Introduction

The three densities most commonly proposed to describe the logarithm of particle sizes are the normal, the hyperbolic and the skew-Laplace. Examples showing the use of these three distributions in this context can be found in Fieller et al. (1992). Julià and Vives-Rego (2005) uses the skew-Laplace distribution to analyze bacterial sizes in axenic cultures. In this paper we summarize the main properties of the skew-Laplace distribution and two useful goodness of fit tests are also presented.

The following argument is employed to justify the use of the normal distribution in particle size analysis:

*Address for correspondence:* P. Puig, Universitat Autònoma de Barcelona, Departament de Matemàtiques, Facultat de Ciències Edifici C, 08193 Cerdanyola del Vallés (Barcelona), Spain.
Received: November 2006
Accepted: February 2007

Suppose that a particle of initial size $X_0$ is repeatedly diminished by breaking off random proportions. The size in the step $j$ is a random proportion of the size in the step $j-1$, that is, $X_j = \varepsilon_j X_{j-1}$, where $\varepsilon_j$ are random variates taking values between 0 and 1. Then the size in the step $n$ is

$$X_n = X_0 \prod_{j=1}^{n} \varepsilon_j.$$

If the variates $\varepsilon_j$ are iid, for large $n$ the distribution of $\log(X_n)$ can be approximated by the normal as a consequence of the central limit theorem.

Consider now that each observation $\log(x_i)$ follows a normal distribution with a different mean $\mu_i$ and variance $\sigma_i^2$. It is reasonable to assume that $\mu_i = f(\sigma_i^2)$, where $f()$ is a suitable monotonous function. This corresponds to the generally accepted idea that large scale observations have a wide dispersion. The simplest selection is a linear relationship, $f(\sigma_i^2) = a + b\sigma_i^2$.

We can also suppose that $\sigma_i^2$ are random variates following a suitable distribution defined on the positive reals. If $\sigma_i^2$ follows a distribution whose density function is $p(x;\gamma,\delta) = \frac{e^{-\gamma x - \delta/x}}{C(\gamma,\delta)}$, the resulting model is the hyperbolic distribution (Barndorff-Nielsen,1977). Correspondingly, if $\sigma_i^2$ follows an exponential distribution then the skew-Laplace distribution is obtained.

However the skew-Laplace distribution can arise as the difference of two exponentials as will be seen in Section 2 and in the example in section 4.1. Several properties, generalizations and applications of the skew-Laplace distribution have been reported in Kotz et al. (2001).

Sometimes the maximum likelihood estimators (MLE) of the parameters of this distribution have been calculated by maximizing directly its likelihood function. This can give numerical problems when iterative methods are used. A proper derivation of the MLE was done in Hinkley and Revankar (1977). These authors worked in the context of log-skew-Laplace models (see also Kotz et al., 2001). In Section 3 we study the maximum likelihood estimation and present a simple proof of the result of Hinkley and Revankar (1977).

In Section 4, we show the Anderson-Darling and Cramér-von Mises goodness of fit tests.

## 2 The skew-Laplace distribution

The skew-Laplace (SKL) or skew-double exponential distribution has a density function, defined over all the reals, of the form

$$p(x;\alpha,\beta,\mu) = \begin{cases} \exp(\frac{(x-\mu)}{\alpha})/(\alpha+\beta) & x \le \mu \\ \exp(\frac{(\mu-x)}{\beta})/(\alpha+\beta) & x > \mu \end{cases} \tag{1}$$

where $\alpha, \beta > 0$ and $\mu$ can be any real number. When $\alpha$ or $\beta$ tends to 0 then the two-parameter exponential or negative-exponential distribution is obtained. When $\alpha = \beta$ it corresponds to the classical Laplace distribution. A skew-Laplace distribution with parameters $\mu$, $\alpha$ and $\beta$ will be referred as SKL($\mu, \alpha, \beta$).

The distribution function, is

$$F(x;\alpha,\beta,\mu) = \begin{cases} \alpha \exp(\frac{(x-\mu)}{\alpha})/(\alpha+\beta) & x \le \mu \\ 1 - \beta \exp(\frac{(\mu-x)}{\beta})/(\alpha+\beta) & x > \mu \end{cases} \tag{2}$$

The profile of the log-density is formed by two lines of slopes $1/\alpha$ and $-1/\beta$ intersecting in $x = \mu$, the location parameter and its mode. Therefore this distribution can be easily detected empirically by plotting a log-histogram.

### 2.1 Moments and properties

Many of the properties described in this section can be found in Kotz et al. (2001). Given a random variable $X$, SKL distributed, it is easy to compute its moment generating function $\Phi(t) = E(\exp(tX))$ giving

$$\Phi(t) = \frac{\exp(\mu t)}{(1+\alpha t)(1-\beta t)} \tag{3}$$

From (3), the cumulant generating function has a very simple form, $K(t) = \log(\Phi(t)) = \mu t - \log(1+\alpha t) - \log(1-\beta t)$, and consequently the mean is $E(X) = \mu + \beta - \alpha$, the variance is $V(X) = \alpha^2 + \beta^2$ and for $i > 2$ the cumulants are $k_i = (i-1)!(\beta^i + (-1)^i \alpha^i)$.

The coeficients of skewness and kurtosis are as follows:

$$\sqrt{\beta_1} = \frac{k_3}{k_2^{3/2}} = \frac{2(\beta^3 - \alpha^3)}{(\alpha^2+\beta^2)^{3/2}}, \quad \beta_2 = 3 + \frac{k_4}{k_2^2} = 3 + \frac{6(\beta^4+\alpha^4)}{(\alpha^2+\beta^2)^2}$$

They can be expressed in terms of $\theta = \beta/\alpha$, giving

$$\sqrt{\beta_1} = \frac{2(\theta^3-1)}{(\theta^2+1)^{3/2}}, \quad \beta_2 = 3 + \frac{6(\theta^4+1)}{(\theta^2+1)^2} \tag{4}$$

As $\theta$ varies in $(0,\infty)$, $\sqrt{\beta_1} \in (-2,2)$ and $\beta_2 \in [6,9)$. From (4) it is evident that $\sqrt{\beta_1}$ determines $\beta_2$. Moreover $\beta_2(\sqrt{\beta_1}) = \beta_2(-\sqrt{\beta_1})$. The following table shows the

relationship between both coefficients:

| $\sqrt{\beta_1}$ | 0.0 | ±0.2 | ±0.5 | ±1.0 | ±1.5 | ±1.8 | ±2.0 |
|---|---|---|---|---|---|---|---|
| $\beta_2$ | 6.00 | 6.03 | 6.17 | 6.68 | 7.58 | 8.34 | 9.00 |

The values $\sqrt{\beta_1} = 0$ and $\beta_2 = 6$ correspond to $\theta = 1$, that is, the Laplace distribution. From an empirical point of view, if the sample skewness and kurtosis coefficients do not lie near the appointed values it would be a sign that the SKL distribution is inadequate for fitting our data.

Another measure of dispersion is $E|X - \mu|$, that is, the mean deviation with respect to the location parameter. It gives, $E|X - \mu| = (\alpha^2 + \beta^2)/(\alpha + \beta) = V(X)/(\alpha + \beta)$. Then the normalizing constant in (1) can be interpreted as the quotient $E|X - \mu|/V(X)$.

### 2.2 Generation of values

Given a random variable $X$, two-parameter exponentially distributed with starting point $x_0$ and expectation $E(X) = \tau + x_0$ then the moment generating function is $\Phi_X(t) = \exp(x_0 t)/(1 - \tau t)$. Hence, $\Phi_{-X}(t) = \exp(-x_0 t)/(1 + \tau t)$ and from (3) it can be readily deduced that the difference of two-parameters exponential independent random variables follows a SKL distribution. Now the parameters of the SKL have a new meaning, that is, $\alpha$ and $\beta$ are the means of each exponential after subtracting its starting points and $\mu$ measures the distance between these starting points.

This result leads to a first approach to simulate a SKL$(\mu, \alpha, \beta)$, by subtracting two independent exponentials starting at 0 with means $\alpha$ and $\beta$ respectively and adding the constant $\mu$. It can be summarized in the following formula:

$$X = \alpha \log(z_1) - \beta \log(z_2) + \mu = \log(z_1^{\alpha}/z_2^{\beta}) + \mu$$

where $z_1$ and $z_2$ are two independent uniform $(0, 1)$ variates.

A second approach comes from the mixture pattern model mentioned in Section 1. Consider that observations follow a normal distribution with mean $a + b\sigma^2$ and variance $\sigma^2$, where $\sigma^2$ is also a continuous random variable with density $g(x)$ over the positive reals. It can be easily shown that the moment generating function of the resulting distribution is $\Phi(t) = \int_0^{\infty} e^{(a+b\sigma^2)t+\sigma^2 t/2} g(\sigma^2) d\sigma^2$. If the mixing density is an exponential with mean $\tau$ then it gives the moment generating function in (3), and the relationship between the two parameterizations is $\mu = a$, $2\alpha\beta = \tau$ and $\beta - \alpha = b\tau$. This can be summarized in the following expression:

$$X = \mu + \frac{\beta - \alpha}{2\alpha\beta} x_1 + x_1 y_1$$

where $x_1$ is an exponential variate with mean $2\alpha\beta$ and $y_1$ is a standard normal, both being independent.

A third approach to simulate a SKL random variable is by using the classical inverse distribution method. It gives the following expression,

$$X = \begin{cases} \alpha \log(\frac{\alpha+\beta}{\alpha} z) + \mu & z \in (0, \frac{\alpha}{\alpha+\beta}) \\ \beta \log(\frac{\beta}{(1-z)(\alpha+\beta)}) + \mu & z \in (\frac{\alpha}{\alpha+\beta}, 1) \end{cases} \tag{5}$$

where $z$ is a uniform $(0, 1)$ variate. This method is better than the preceding ones because it only requires one uniform value for each SKL value.

## 3 Parameter estimation

Consider a sample $X = (x_1, \ldots, x_n)$, coming from a $SKL(\mu, \alpha, \beta)$. Our goal is to find the MLEs of the parameters. First suppose that $\mu$ is known and $x_{(r)} \leq \mu \leq x_{(r+1)}$, where $x_{(r)}$ indicates the r-th order statistic. The log-likelihood function can be written as

$$l(X; \mu, \alpha, \beta) = -\frac{1}{\alpha} L_r(\mu) - \frac{1}{\beta} U_r(\mu) - n \log(\alpha + \beta) \tag{6}$$

where $L_r(\mu) = \sum_{i=1}^{r} (\mu - x_{(i)})$ and $U_r(\mu) = \sum_{i=r+1}^{n} (x_{(i)} - \mu)$. The likelihood function is differentiable with respect to $\alpha$ and $\beta$ in the domain of the parameters. Then, solving the likelihood equations we obtain

$$\begin{aligned} \hat{\alpha_0}(\mu) &= \frac{\sqrt{L_r(\mu)}(\sqrt{L_r(\mu)} + \sqrt{U_r(\mu)})}{n} \\ \hat{\beta_0}(\mu) &= \frac{\sqrt{U_r(\mu)}(\sqrt{L_r(\mu)} + \sqrt{U_r(\mu)})}{n} \end{aligned} \tag{7}$$

Notice that if $\mu \leq x_{(1)}$ then $L_r(\mu) = 0$ and, similarly, if $\mu \geq x_{(n)}$ then $U_r(\mu) = 0$. Consequently it can be directly shown that (7) is also valid to describe the maximum likelihood estimators of $\alpha$ and $\beta$ in these situations.

Taking into account that $(L_r(\mu) + U_r(\mu))/n = \sum_{i=1}^{n} |x_i - \mu|/n = \Delta(\mu)$ and $(L_r(\mu) - U_r(\mu))/n = \mu - \bar{x}$, (7) can be written in a more suitable form, independent of where $\mu$ is located:

$$\begin{aligned} \hat{\alpha_0}(\mu) &= \tfrac{1}{2}(\Delta(\mu) - \bar{x} + \mu + \sqrt{\Delta^2(\mu) - (\bar{x} - \mu)^2}) \\ \hat{\beta_0}(\mu) &= \tfrac{1}{2}(\Delta(\mu) + \bar{x} - \mu + \sqrt{\Delta^2(\mu) - (\bar{x} - \mu)^2}) \end{aligned} \tag{8}$$

Notice that if $\mu \leq x_{(1)}$ then $\Delta(\mu) = \bar{x} - \mu$ and $\hat{\alpha_0}(\mu) = 0$. Similarly, if $\mu \geq x_{(n)}$ then

$\hat{\beta}_0(\mu) = 0$. For these situations the MLEs are degenerate in the sense that the estimations lie outside the domain of the parameters.

Now, by substituing (8) in (6), the maximum of the likelihood function is

$$l_M(X;\mu) = -n(\log(\Delta(\mu) + \sqrt{\Delta^2(\mu) - (\bar{x}-\mu)^2}) + 1) \tag{9}$$

Therefore, if all the parameters are unknown, the MLE of $\mu$ can be found by maximizing (9) or, equivalently, by minimizing $\psi(\mu) = \Delta(\mu) + \sqrt{\Delta^2(\mu) - (\bar{x}-\mu)^2}$. Observe that $\psi(\mu) = \bar{x} - \mu$ for $\mu < x_{(1)}$, and $\psi(\mu) = -\bar{x} + \mu$ for $\mu > x_{(n)}$. Then it is obvious that the minimum must be located in the interval $[x_{(1)}, x_{(n)}]$.

The function $\psi(\mu)$ is not differentiable at the points $\mu = x_i$, but the derivative can be computed at all other points. Then, for $x_{(r)} < \mu < x_{(r+1)}$, $r = 1, \ldots, n-1$, $\Delta(\mu) = \frac{2r-n}{n}\mu - 2\sum_{i=1}^{r} x_{(i)} + \bar{x}$ and

$$\psi'(\mu) = \frac{2r-n}{n} + \frac{\Delta(\mu)(2r-n)/n + \bar{x} - \mu}{\sqrt{\Delta^2(\mu) - (\bar{x}-\mu)^2}}$$

Straightforward calculations show that the unique solution of $\psi'(\mu) = 0$ is $\mu_0 = (r^2\bar{x} + (n-2r)\sum_{i=1}^{r} x_{(i)})/(r(n-r))$. If $\mu_0$ is not in $(x_{(r)}, x_{(r+1)})$, then $\psi(\mu)$ is monotone in this interval. Otherwise, further calculations show that

$$\psi''(\mu_0) = -\frac{n}{2r(\bar{x} - \sum_{i=1}^{r} x_{(i)}/r)}$$

Notice that $\bar{x} - \sum_{i=1}^{r} x_{(i)}/r \geq 0$, and equality occurs with probability 0. Consequently in $\mu_0$ we have a local maximum. Due to the continuity of $\psi(\mu)$, if the minimum is not attained inside the intervals $(x_{(r)}, x_{(r+1)})$, $r = 1, \ldots, n-1$, it must be attained at the borders, that is, in one (or several) of the sample values $x_{(i)}$, $i = 1, \ldots, n$. Now, we have proved the following theorem:

**Theorem 1 (Hinkley and Revankar, 1977)** *Let $x_1, \ldots, x_n$ be a sample coming from a SKL$(\mu, \alpha, \beta)$. The MLEs of the parameters are given by,*

$$\begin{aligned} \hat{\mu} &= x_j \\ \hat{\alpha} &= \tfrac{1}{2}(\Delta(\hat{\mu}) - \bar{x} + \hat{\mu} + \sqrt{\Delta^2(\hat{\mu}) - (\bar{x}-\hat{\mu})^2}) \\ \hat{\beta} &= \tfrac{1}{2}(\Delta(\hat{\mu}) + \bar{x} - \hat{\mu} + \sqrt{\Delta^2(\hat{\mu}) - (\bar{x}-\hat{\mu})^2}) \end{aligned}$$

*where $\Delta(\mu) = \sum_{i=1}^{n} |x_i - \mu|/n$ and $x_j$ is any sample value where the function $\psi(\mu) = \Delta(\mu) + \sqrt{\Delta^2(\mu) - (\bar{x}-\mu)^2}$ attains its unique minimum. Moreover the maximum of the log-likelihood function is*

$$l_M(X) = -n(\log(\psi(\hat{\mu})) + 1)$$

**Remark 1** Observe that the calculation of $\hat{\mu}$ is very simple because we only need to evaluate the function $\psi(\mu)$ at a finite number of points, that is, the sample values.

The MLEs are not necessarily unique but the function $\psi(\mu)$ has a unique absolute minimum. The points where $\psi(\mu)$ attains its minimum are not necessarily consecutive as happens with the Laplace distribution ($\alpha = \beta$). For example for the sample, $-1.085$, 0.043, 3.326, 3.954, 5.967, the maximum likelihood estimates of the location parameter are $\hat{\mu}_1 = -1.085$ and $\hat{\mu}_2 = 5.967$. We have observed this troublesome phenomenon only with small samples.

When $\hat{\mu} = x_{(1)}$ or $\hat{\mu} = x_{(n)}$ then $\hat{\alpha} = 0$ or $\hat{\beta} = 0$ and empirically, this means that data is fitted by the exponential or negative-exponential distribution. This situation can also be troublesome and unfortunately it can occur in moderately small samples with an appreciable probability. For instance, we have simulated 10000 samples of different sizes for a SKL(0,1,2). For $n = 5$ this anomaly has been observed in 96% of the samples. For $n = 10$ in 63% and for $n = 20$ in 22%. For $n = 50$ it only happened in 1% of the samples. Consequently, MLEs are not recommended for small samples.

The density of the SKL does not satisfy the standard conditions of regularity. However the consistency and the asymptotic efficiency of the MLE can be established using the very general conditions of Daniels (1961)(see also Hinkley and Revankar, 1977). The asymptotic variance $V$ can be calculated in a standard way by inverting the Fisher information matrix. It gives the following:

$$V = \begin{pmatrix} 2\alpha\beta & \alpha\beta & -\alpha\beta \\ \alpha\beta & \alpha(\alpha+\beta) & 0 \\ -\alpha\beta & 0 & \beta(\alpha+\beta) \end{pmatrix} \tag{10}$$

The asymptotic variance of some functions of the estimates of the parameters can be calculated from here. For instance, to test symmetry it is necessary to estimate $\theta = \beta/\alpha$. It can be shown that the asymptotic variance of $\hat{\theta} = \hat{\beta}/\hat{\alpha}$ is $V(\hat{\theta}) = \theta(1+\theta)^2$.

Notice that the MLE of the expectation is $\hat{E} = \hat{\mu} + \hat{\beta} - \hat{\alpha} = \overline{x}$, that is the sample mean. Then its variance is $V(\hat{E}) = (\alpha^2 + \beta^2)/n$ and approximate confidence intervals can be calculated easily.

In practice is important to decide if the skew-Laplace distribution is a good choice to fit a data set. In the next section some goodness of fit tests are presented.

## 4 Goodness of fit tests

Our goodness of fit tests will be based on statistics which compare the empirical distribution function (EDF) of the sample with the hypothesised distribution $F(x)$.

The EDF is defined by $F_n(x) = \frac{\#x_i \leq x}{n}$. The statistics considered are the Cramér-von Mises $W^2$ and the Anderson-Darling $A^2$:

$$W^2 = n \int_{-\infty}^{\infty} \{F_n(x) - F(x)\}^2 \, dF(x)$$

$$A^2 = n \int_{-\infty}^{\infty} \{F_n(x) - F(x)\}^2 \, \psi(x) dF(x)$$

where $\psi(x) = 1/[F(x)\{1 - F(x)\}]$.

Generally these tests are powerful. The percentage points for these and other EDF tests for a variety of distributions can be found in D'Agostino and Stephens (1986).

The tests procedure is as follows. Suppose the order statistics (ascending) of the sample are $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

a) Find the MLEs of the parameters $(\hat{\mu}, \hat{\alpha}, \hat{\beta})$ following the algorithm of Theorem 1.

b) Make the transformation $z_{(i)} = F(x_{(i)}; \hat{\mu}, \hat{\alpha}, \hat{\beta})$, for $i = 1, \ldots, n$. The $z_{(i)}$ will be in ascending order.

c) The Cramér-von Mises statistic is computed from

$$W^2 = \sum_{i=1}^{n} \{z_{(i)} - (2i - 1)/(2n)\}^2 + 1/(12n)$$

and the Anderson-Darling statistic from

$$A^2 = -n - (1/n) \sum_{i=1}^{n} (2i - 1) \left[\log(z_{(i)}) + \log(1 - z_{(n+1-i)})\right]$$

d) Estimate the coefficient of skewness by using the expression

$$\sqrt{\hat{\beta}_1} = \frac{2(\hat{\beta}^3 - \hat{\alpha}^3)}{(\hat{\alpha}^2 + \hat{\beta}^2)^{3/2}}$$

e) Look at the table 1 for the chosen statistic and interpolate to find the percentage point at a given significance level. If the value of the statistic is greater than this percentage point, then the null hypothesis is rejected at this level.

The distributions of $W^2$ and $A^2$ depend only on $|\beta/\alpha|$. Simulation studies show us that the tests performed by estimating the coefficient of skewness have a significance level slightly lower than expected.

The percentage points of the tables are those of the asymptotic distribution of $W^2$ and $A^2$ under the null hypothesis. They have been computed using the standard techniques

described in Stephens (1976) (see also Puig and Stephens, 2000). For finite samples the percentage points can be calculated by using Monte Carlo methods, but they are very close to the asymptotic for samples above $n = 20$.

***Table 1***: *Percentage points of the asymptotic distribution of $W^2$ and $A^2$ for different values of $\sqrt{\beta_1}$ and different significance levels (in boldface).*

| | $W^2$ | | | | $A^2$ | | | |
|---|---|---|---|---|---|---|---|---|
| $\sqrt{\beta_1}$ | **0.10** | **0.05** | **0.025** | **0.01** | **0.10** | **0.05** | **0.025** | **0.01** |
| 0.00 | .077 | .091 | .106 | .125 | .498 | .582 | .665 | .774 |
| ±0.20 | .077 | .092 | .107 | .126 | .498 | .583 | .666 | .776 |
| ±0.40 | .078 | .093 | .108 | .129 | .501 | .586 | .671 | .784 |
| ±0.60 | .079 | .095 | .111 | .133 | .504 | .592 | .680 | .796 |
| ±0.80 | .081 | .098 | .115 | .139 | .510 | .601 | .692 | .814 |
| ±1.00 | .084 | .102 | .121 | .148 | .519 | .614 | .710 | .840 |
| ±1.20 | .088 | .108 | .129 | .158 | .531 | .632 | .735 | .875 |
| ±1.40 | .093 | .116 | .140 | .172 | .550 | .659 | .771 | .924 |
| ±1.60 | .102 | .127 | .154 | .191 | .579 | .700 | .825 | .996 |
| ±1.80 | .116 | .146 | .177 | .220 | .634 | .775 | .922 | 1.121 |
| ±1.90 | .128 | .162 | .198 | .246 | .692 | .851 | 1.017 | 1.243 |
| ±1.95 | .139 | .176 | .215 | .268 | .748 | .925 | 1.108 | 1.359 |
| ±1.98 | .150 | .190 | .233 | .290 | .816 | 1.013 | 1.218 | 1.498 |
| ±1.99 | .156 | .199 | .243 | .303 | .861 | 1.072 | 1.290 | 1.588 |
| ±2.00 | .174 | .222 | .271 | .338 | 1.062 | 1.321 | 1.591 | 1.959 |

### *4.1 An example*

Bain and Engelhardt (1973) consider the following data set, consisting of 33 differences in flood levels between stations on a river:

1.96, 1.97, 3.60, 3.80, 4.79, 5.66, 5.76, 5.78, 6.27, 6.30, 6.76, 7.65, 7.84, 7.99, 8.51, 9.18, 10.13, 10.24, 10.25, 10.43, 11.45, 11.48, 11.75, 11.81, 12.34, 12.78, 13.06, 13.29, 13.98, 14.18, 14.40, 16.22, 17.06

They fit the data by using the Laplace distribution arguing that the observations could occur as the difference of two exponential distributions with the same mean. However, the fit does not work well as can be seen in Puig and Stephens (2000) who perform EDF tests for the Laplace distribution. Possibly the two exponentials do not have the same mean and consequently a reasonable alternative is the skew-Laplace distribution.

By using theorem 1, the MLEs are $\hat{\mu} = 11.75$, $\hat{\alpha} = 4.4654$ and $\hat{\beta} = 2.0691$. The estimated coefficient of skewness is $\sqrt{\hat{\beta}_1} = -1.345$ and the EDF test statistics are $W^2 = .097$ and $A^2 = .568$. From Table 1 the skew-Laplace assumption is not rejected for the Cramér-von Mises statistic even at a significance level of 0.10 and for the Anderson-Darling statistic at a level of 0.05.

Given the above, an approximate 95% confidence interval for the mean can be calculated from the expression

$$\bar{x} \pm 1.96 \sqrt{\frac{\hat{\alpha}^2 + \hat{\beta}^2}{n}} = 9.354 \pm 1.679$$

We then test the Laplace assumption against the skew-Laplace for this example. It is equivalent to consider $H_0 : \theta = 1$ against $H_1 : \theta \neq 1$. As has been pointed out in Section 3, an approximate 95% confidence interval for $\theta$ can be calculated from

$$\hat{\theta} \pm 1.96 \sqrt{\frac{\hat{\theta}}{n}(1 + \hat{\theta})} = 0.463 \pm 0.340$$

Consequently the Laplace assumption must be rejected in favour of the general skew-Laplace.

## Acknowledgements

This research was partially supported by grant MTM2006-01477 from the Ministry of Education of Spain.

## References

Bain, Lee J. and Englehardt, Max (1973). Interval estimation for the two-parameter double exponential distribution. *Technometrics*, 15, 875-887.

Barndorff-Nielsen, O. E. (1977). Exponentially decreasing distributions for the logarithm of a particle size. *Proceedings of the Royal Society (London), Ser. A*, 353, 401-419.

D'Agostino and Stephens, M. A. (1986). *Goodness of Fit Techniques*. Marcel Decker, New York.

Daniels, H. E. (1961). The asymptotic efficiency of a maximum likelihood estimator. *Proc. 4th Berkeley Sympos. Math. Statist. and Prob.*, Vol. I, 151-163 Univ. California Press, Berkeley, Calif.

Fieller, N. R. J., Flenley, E. C. and Olbricht, W. (1992). Statistics of particle size data. *Applied Statistics*, 41, 127-146.

Hinkley, D. V. and Revankar, N. S. (1977). On the estimation of the Pareto law from underreported data. A further analysis. *Journal of Econometrics*, 5, 1-11.

Julià, O. and Vives-Rego, J. (2005). Skew-Laplace distribution in Gram-negative bacterial axenic cultures: new insights into intrinsic cellular heterogeneity. *Microbiology*, 151, 749-755.

Kotz, S., Kozubowski, T. J. and Podgórski, K. (2001). *The Laplace Distribution and Generalizations*. Birkhäuser, Berlin.

Puig, P. and Stephens, M. A. (2000). Tests of fit for the Laplace distribution, with applications. *Technometrics* 42, 4, 417-424.

Stephens, M. A. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. *Annals of Statistics*, 4, 357-369.

SORT 31 (1) January-June 2007, 55-74

# Parameter estimation of S-distributions with alternating regression

I-Chun Chou[1], Harald Martens[2], Eberhard O. Voit[1,*]

[1] *Georgia Institute of Technology and Emory University*
[2] *CIGENE/IKBM, Norwegian U. of Life Sciences*

**Abstract**

We propose a novel *3-way alternating regression* (3-AR) method as an effective strategy for the estimation of parameter values in S-distributions from frequency data. The 3-AR algorithm is very fast and performs well for error-free distributions and artificial noisy data obtained as random samples generated from S-distributions, as well as for traditional statistical distributions and for actual observation data. In rare cases where the algorithm does not immediately converge, its enormous speed renders it feasible to select several initial guesses and search settings as an effective countermeasure.

*MSC:* 62G05, 62E17, 62J02, 62J05.

*Keywords:* Alternating regression, Parameter estimation, S-distribution, S-system.

## 1 Introduction

Motivated by a distribution family based on S-systems (Savageau, 1982), the S-distribution was introduced in the early 1990s as a convenient univariate, unimodal four-parameter distribution that is capable of modelling a wide range of shapes and skewness (Voit, 1992). Due to its rich shape flexibility and relatively simple mathematical

* *Address for correspondence:* I-Chun Chou, Eberhard O. Voit. The Wallace H. Coulter Department of Biomedical Engineering at Georgia Institute of Technology and Emory University, 313 Ferst Drive, Atlanta, GA, 30332, U.S.A. E-mail: gtg392p@mail.gatech.edu, Eberhard.Voit@bme.gatech.edu. Harald Martens. CIGENE/IKBM, Norwegian U. of Life Sciences, P.O. Box 5003, N-1432 Ås, Norway. Email: harald.martens@matforsk.no.
Received: January 2007
Accepted: April 2007

format, the S-distribution has been shown to constitute a good general-purpose default distribution, especially for data of unknown structure. The S-distribution may also be used in lieu of the traditional distributions, because it always has the same structure and, with an appropriate choice of parameter values, rather accurately approximates many continuous central and non-central distributions, as well as a wide variety of discrete distributions (Voit, 1992; Voit and Yu, 1994; Yu and Voit, 1996). In addition, the S-distribution allows for combinations of parameter values that do not correspond to traditional distributions and permits a spectrum of distributions with long or heavy tails and with skewness to the left or right. Thus, one might in many cases expect a better fit than is possible with traditional distributions. As a specific application of the combination of its flexibility and small number of parameters, the S-distribution is well suited for the non-trivial characterization of trends of distributions that change mean, variance, shape, and even skewness over time (Voit, 1996; Sorribas, March and Voit, 2000; Voit and Sorribas, 2000).

The S-distribution is formulated as a differential equation, which renders the estimation of parameter values from data a challenge. Several methods have been suggested for this task, including nonlinear regression (Voit, 1992; Sorribas, March and Voit, 2000), a graphical method (Voit, 1992), constrained maximum likelihood estimation (Voit, 2000), and techniques based on quantiles (Voit and Schwacke, 2000; Hernández-Bermejo and Sorribas, 2001). Here, we propose an entirely different method called *3-way alternating regression* (3-AR), which was motivated by a 2-way alternating regression method used for the estimation of parameters in multivariate S-systems (Chou, Martens and Voit, 2006). The main appeal of 3-AR is its enormous speed and robustness. In this article, we discuss the method and apply it to several artificial and actual examples.

## 2 S-distribution

The S-distribution is a four-variable distribution that emphasizes the cumulative density function (*cdf*) $F$, which is formulated as a differential equation with respect to random variable $X$ and reads

$$f = \frac{dF}{dX} = \alpha\left(F^g - F^h\right), \qquad F_0 = F(X_0) \in (0,1). \tag{1}$$

Because the probability density function (*pdf*) $f$ is the derivative of $F$, the S-distribution can be seen as an algebraic function $f(F)$. The first parameter of the distribution, $X_0$, characterizes the location of the distribution. The second parameter, $\alpha$, is a positive real number, which determines the scale. The remaining two parameters,

$g$ and $h$, may be any real numbers as long as $g < h$; they determine the shape of the distribution.[1]

## 3 Alternating Regression

Suppose the S-distribution is characterized through $N$ values of the random variable, $X_1$, $X_2$,..., $X_k$,..., $X_N$, and that $X_k$, $F(X_k)$ and $f(X_k)$ are observed or obtainable for each $k$ (see later sections for further discussion on the construction of *pdf*s and *cdf*s). For the purpose of parameter estimation, the original differential equation can then be analyzed in the form of $N$ uncoupled algebraic equations as

$$\begin{aligned} f(X_1) &\approx \alpha \left(F^g(X_1) - F^h(X_1)\right), \\ f(X_2) &\approx \alpha \left(F^g(X_2) - F^h(X_2)\right), \\ &\vdots \\ f(X_k) &\approx \alpha \left(F^g(X_k) - F^h(X_k)\right), \\ &\vdots \\ f(X_N) &\approx \alpha \left(F^g(X_N) - F^h(X_N)\right). \end{aligned} \tag{2}$$

The $\approx$ symbol is used because the data may only be representable in approximation by the S-distribution format. As a consequence of this decoupling step, substitution of the derivative of $F$ with $f$ allows us to estimate the S-distribution parameters $\alpha$, $g$, and $h$ in a purely algebraic system (*cf.* Voit and Almeida, 2000). We propose for this estimation purpose a new method called *3-way alternating regression* (3-AR).

In previous work, we have shown that alternating regression (AR), applied to S-system models of the form

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j^{h_{ij}}, \; i = 1, 2, ..., n, \tag{3}$$

and combined with methods for slope estimation and decoupling systems of differential equations, provides a fast tool for identifying parameter values from time series data (Chou, Martens and Voit, 2006). The key feature of AR is the reduction of the nonlinear inverse problem of parameter estimation into iterative steps of two phases of linear regression. In the first phase, the parameters of the $\beta$-term, $\beta_i$ and $h_{ij}$, are set to some reasonable values. Given measurements of all $X_i$ at $N$ time points and estimates $S_i(t_k)$ of

1. Throughout the paper, random variables and *cdf*s are represented as upper-case italics, while *pdf*s are given by the corresponding lower-case italic symbols ($X$, $F$, $f$). An upper-case boldface variable (**L**) represents a matrix of regressor columns and a lower-case boldface variable (**y**) represents a regressand column in a linear statistical regression model.

the slope of $X_i$ at these points, the $\beta$-term becomes a number at each time point, and this number is added to both sides of Equation (3). Taking the logarithm of the equation for each time point, one obtains a linear regression problem with the slope and the $\beta$-term as a real number on the left-hand side, and a linear expression on the right hand side:

$$\log\left(S_i(t_k) + \beta_i \prod_{j=1}^{n} X_j^{h_{ij}}(t_k)\right) \approx \log(\hat{\alpha}_i) + \sum_{j=1}^{n} \hat{g}_{ij} \log(X_j(t_k)) + \varepsilon_{i,k} \tag{4}$$

The regression with the $N$ equations of this type at time points $t_k$ now yields estimates for $\alpha_i$ and all $g_{ij}$. In the second step of AR, these estimates are used in an analogous fashion to compute $\beta_i$ and $h_{ij}$. The algorithm switches back and forth and usually converges fast (see Chou, Martens and Voit (2006) for details).

The S-distribution is obviously a special case of an S-system, with the notable feature that by definition $\alpha = \beta$. This feature is important for AR methods, because $\alpha$ and $\beta$ are no longer independent of each other, and it turns out to be inconvenient to constrain $\alpha$ to be the same in both phases of the regression. Therefore, we modify the 2-way AR approach here into a three-cycle 3-AR method specifically for S-distribution estimation.

Similar to the original AR, 3-AR works by iteratively cycling between phases of linear regression. The first phase begins with guesses of the values of $g$ and $h$ and uses these to solve for the value of parameter $\alpha$. Experience has shown that it is more expedient to start the algorithm with $g$ and $h$, rather than $g$ and $\alpha$ or $h$ and $\alpha$, presumably due to the fact that the typical ranges of $g$ and $h$ are much smaller than that of $\alpha$ and because $h$ is per definition constrained by $g$. The second phase takes estimates of $\alpha$ and $h$ to solve for $g$, while the third phase takes estimates of $\alpha$ and $g$ to solve for $h$ and thus improve the parameter guesses or estimates from the previous phases. The phases are iterated until a solution is found or AR terminates for other reasons. The overall flow of the method is shown in Figure 1, and specific steps of the 3-AR algorithm are detailed below.

### *Steps of the 3-AR Algorithm*

{1} Define $\mathbf{L_f}$ and $\mathbf{L_F}$ as $2 \times N$ matrices of logarithms of regressors $f$ and $F$, respectively:

$$\mathbf{L_f} = \begin{bmatrix} 1 & \log(f(X_1)) \\ 1 & \log(f(X_2)) \\ \vdots & \vdots \\ 1 & \log(f(X_k)) \\ \vdots & \vdots \\ 1 & \log(f(X_N)) \end{bmatrix} \tag{5}$$

$$\mathbf{L_F} = \begin{bmatrix} 1 & \log(F(X_1)) \\ 1 & \log(F(X_2)) \\ \vdots & \vdots \\ 1 & \log(F(X_k)) \\ \vdots & \vdots \\ 1 & \log(F(X_N)) \end{bmatrix} \tag{6}$$

$\mathbf{L_f}$ is used in the first phase of AR to determine $\alpha$, and $\mathbf{L_F}$ is used in the second and third phases of AR to determine $g$ and $h$.

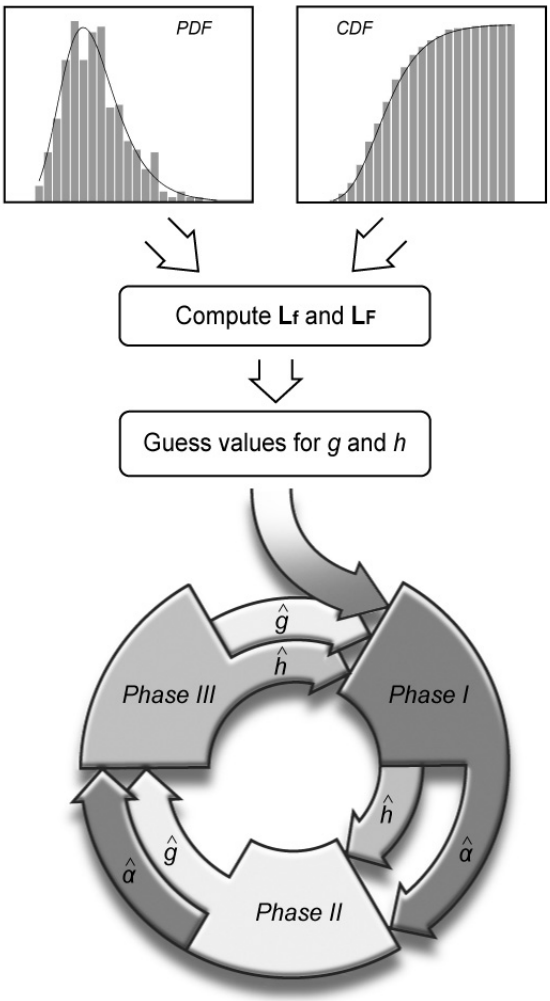


***Figure 1***: *Flow of parameter estimation by 3-way alternating regression.*

{2} Select values for $g$ and $h$ in accordance with experience about S-distribution parameters (see Voit (1992) for relationships between parameter values and distributional shape).

{3} For all $X_k$, $k = 1, 2, \ldots, N$, compute $F^{\hat{g}}(X_k) - F^{\hat{h}}(X_k)$, using values $F(X_k)$ from the data distribution. Here $\hat{g}$ and $\hat{h}$ denote the estimators of $g$ and $h$ after the $2^{nd}$ iteration, while during the $1^{st}$ iteration, $\hat{g}$ and $\hat{h}$ are the initial guesses for $g$ and $h$, respectively. Determine the index $I_\alpha$ of all positive quantities $F^{\hat{g}}(X_k) - F^{\hat{h}}(X_k)$. The number of *qualified points* then becomes $N_\alpha$, where $N_\alpha$ is the length of $I_\alpha$. Quantities restricted to $N_\alpha$ instead of all $N$ points are identified in the following with an additional subscript $\alpha$. Note: Theoretically $F^g(X_k)$ should always be greater than $F^h(X_k)$, because $g < h$, or at most equal, for $F = 0$ and $F = 1$. However, because of noise, this may not always be true, suggesting temporary exclusion of some data points.

{4} After logarithmic transformation and rearrangement, Equation (1) can be written as $\log\left(\frac{f}{\alpha}\right) = \log(F^g - F^h)$. Therefore, compute the $N_\alpha$-dimensional vector $\mathbf{y_\alpha} = \log\left(F_\alpha^{\hat{g}} - F_\alpha^{\hat{h}}\right)$ for $N_\alpha$ points, as well as $\mathbf{L_{f_\alpha}}$, where the subscript $\alpha$ limits the computation to qualified points.

{5} Based on the linear regression model

$$\mathbf{y_\alpha} = \mathbf{L_{f_\alpha}}\hat{\mathbf{b}}_\alpha + \boldsymbol{\varepsilon}_\alpha, \tag{7}$$

estimate the regression coefficient vector $\hat{\mathbf{b}}_\alpha = [\hat{b}_{\alpha_1}, \hat{b}_{\alpha_2}]^{\mathrm{T}}$ over the $N_\alpha$ qualified points, to obtain an estimate of $\alpha$. In other words, this equation may be written as $\mathbf{y_\alpha} \approx \log\left(\frac{1}{\hat{\alpha}}\right) + \log(f_\alpha) + \varepsilon_\alpha$ so that $\hat{b}_{\alpha_1}$ is equivalent to $\log\left(\frac{1}{\hat{\alpha}}\right)$ and $\hat{b}_{\alpha_2}$ is the coefficient of $\log(f_\alpha)$, which is expected to converge to 1. Thus, $\hat{\mathbf{b}}_\alpha$ is estimated with any of the methods of linear regression, *e.g.*, by ordinary least squares regression (OLSR) as

$$\hat{\mathbf{b}}_\alpha = \left(\mathbf{L}_{\mathbf{f}_\alpha}^{\mathrm{T}}\mathbf{L}_{\mathbf{f}_\alpha}\right)^{-1}\mathbf{L}_{\mathbf{f}_\alpha}^{\mathrm{T}}\mathbf{y_\alpha}. \tag{8}$$

As an alternative to OLSR, weighted or robust estimators could be used. If $\mathbf{L_{f_\alpha}}$ does not have full column rank, *i.e.*, if $\mathbf{L}_{\mathbf{f}_\alpha}^{\mathrm{T}}\mathbf{L}_{\mathbf{f}_\alpha}$ has a small eigenvalue, one could also use a small ridge regression constant $\kappa$ for stabilization and compute $\hat{\mathbf{b}}_\alpha$ as

$$\hat{\mathbf{b}}_\alpha = \left(\mathbf{L}_{\mathbf{f}_\alpha}^{\mathrm{T}}\mathbf{L}_{\mathbf{f}_\alpha} + \kappa\mathbf{I}\right)^{-1}\mathbf{L}_{\mathbf{f}_\alpha}^{\mathrm{T}}\mathbf{y_\alpha}. \tag{9}$$

{6} For the estimation of $g$, reformulate Equation (1) as $\frac{f}{\alpha} + F^h = F^g$. Thus, using values of $f(X_k)$ and $F(X_k)$ that are directly obtained from the data (see later sections), compute $\frac{f(X_k)}{\hat{\alpha}} + F^{\hat{h}}(X_k)$ for all $X_k$, $k = 1, 2, \ldots, N$. Here $\hat{h}$ denotes the estimator of $h$ after the $2^{nd}$ iteration, while during the $1^{st}$ iteration, $\hat{h}$ is the initial guess for $h$. Find the index $I_g$ of

positive quantities $\frac{f(X_k)}{\hat{\alpha}} + F^{\hat{h}}(X_k)$. The number of qualified points for this step becomes $N_g$, where $N_g$ is the length of $I_g$.

{7} Compute the $N_g$-dimensional vector $\mathbf{y_g} = \log\left(\frac{f_g}{\hat{\alpha}} + F_g^{\hat{h}}\right)$ for $N_g$ points, as well as $\mathbf{L_{F_g}}$.

{8} Based on the linear regression model

$$\mathbf{y_g} = \mathbf{L_{F_g}}\hat{\mathbf{b}}_\mathbf{g} + \boldsymbol{\varepsilon}_\mathbf{g}, \tag{10}$$

and in analogy to step {5}, estimate the regression coefficient vector $\hat{\mathbf{b}}_\mathbf{g} = [\hat{b}_{g_1}, \hat{b}_{g_2}]^{\mathrm{T}}$ by regression over the $N_g$ time points as

$$\hat{\mathbf{b}}_\mathbf{g} = \left(\mathbf{L}_{\mathbf{F_g}}^{\mathrm{T}}\mathbf{L}_{\mathbf{F_g}}\right)^{-1}\mathbf{L}_{\mathbf{F_g}}^{\mathrm{T}}\mathbf{y_g}, \tag{11}$$

or with an alternative regression method. The estimator $\hat{b}_{g_2}$ is the parameter of interest, $\hat{g}$; estimator $\hat{b}_{g_1}$ is expected to be zero in the model.

{9} For the estimation of $h$, reformulate Equation (1) as $F^g - \frac{f}{\alpha} = F^h$ and compute $F^{\hat{g}}(X_k) - \frac{f(X_k)}{\hat{\alpha}}$ for all $X_k$, $k = 1, 2, \ldots, N$, again using the values of $f(X_k)$ and $F(X_k)$. Determine the index $I_h$ of positive quantities $F^{\hat{g}}(X_k) - \frac{f(X_k)}{\hat{\alpha}}$. The number of qualified points for this step becomes $N_h$, where $N_h$ is the length of $I_h$.

{10} Compute the $N$-dimensional vector $\mathbf{y_h} = \log\left(F_h^{\hat{g}} - \frac{f_h}{\hat{\alpha}}\right)$ for $N_h$ points, as well as $\mathbf{L_{F_h}}$.

{11} Based on the linear regression model

$$\mathbf{y_h} = \mathbf{L_{F_h}}\hat{\mathbf{b}}_\mathbf{h} + \boldsymbol{\varepsilon}_\mathbf{h}, \tag{12}$$

and in analogy to steps {5} and {8}, estimate the regression coefficient vector $\hat{\mathbf{b}}_\mathbf{h} = [\hat{b}_{h_1}, \hat{b}_{h_2}]^{\mathrm{T}}$ by regression over the $N_h$ time points as

$$\hat{\mathbf{b}}_\mathbf{h} = \left(\mathbf{L}_{\mathbf{F_h}}^{\mathrm{T}}\mathbf{L}_{\mathbf{F_h}}\right)^{-1}\mathbf{L}_{\mathbf{F_h}}^{\mathrm{T}}\mathbf{y_h}, \tag{13}$$

or with an alternative regression method. The estimator $\hat{b}_{h_2}$ is the parameter of interest, $\hat{h}$; estimator $\hat{b}_{h_1}$ is expected to be zero in the model.

{12} Iterate steps {3} – {11} until a solution is found or some termination criterion is satisfied.

At each phase of 3-AR, lack-of-fit criteria are estimated and used for monitoring the iterative process and to define termination conditions. We use here specifically the logarithm of the sums of squared $y$-errors ($SSE_\alpha$, $SSE_g$, and $SSE_h$) as optimization criteria

for the three regression phases. Upon convergence, we also compute the residual error *SSE* of the fit and the standard deviation $S.D. = \sqrt{SSE/(N-p)}$ of the *pdf*, as well as the *cdf* and $f$-$F$ plots, where $p$ is the number of estimated parameters, which in all cases here is 3.

The location parameter $X_0$ is not explicit in the method, because it does not appear in the algebraic formulation of the *pdf* as a function of the *cdf*. However, it is easily estimated directly as the observed or estimated median or by optimizing the horizontal position of the distribution with parameters $\hat{\alpha}$, $\hat{g}$, and $\hat{h}$ (Voit, 2000).

## 4 Results

We tested the 3-AR method with a large number of representative cases, including estimations based on "data" from error-free distributions, artificial noisy data obtained as random samples generated from S-distributions with known parameters, traditional statistical distributions (using Matlab®), and from actual observation data. Representative details of each case are discussed in this section.

### *4.1 Fitting distributions without noise*

In order not to confuse the features of 3-AR with possible effects of noise in the data, we begin the exploration of convergence properties by using true S-distribution *cdf*s and *pdf*s, which are evaluated directly from Equation (1) at a number of values for the random variable. Specifically, we choose 50 equally spaced instances of the random variable and compute the corresponding $f$ and $F$ values from Equation (1) to obtain the "true" *pdf* and *cdf*. Figure 2 shows an example of a typical convergence pattern. Starting from the (essentially arbitrary) initial guesses $g = 3$ and $h = 6$, it takes the 3-AR algorithm just 51 iterations to converge to the true solution, requiring 0.0742 seconds on a Pentium® D (~3.4GHz) machine. Since we use noise-free data, the residual error should approach 0, which corresponds to $-\infty$ in logarithmic coordinates. We use $-9$ instead as one of the termination criteria, which corresponds to a result very close to the true value, but allows for issues of machine precision and numerical inaccuracies. The low error tolerance causes the algorithm to need 51 iterations. However, as Figure 2 indicates, the estimates are already very close to the true optimum after just a few initial iterations. Big jumps in the beginning do not negatively affect convergence time. For instance, using the same error tolerance and initial guesses $g = 10$, $h = 10.5$ or $g = 100$ and $h = 120$, respectively, the algorithm needs 57 iterations (0.0535 second) or 63 iterations (0.0567 second) to converge to the true parameter values. Thus, somewhat different from results for general S-systems (Chou, Martens and Voit, 2006),

the speed of convergence here does not depend much on initial guesses. Also in contrast to observations with S-systems, the convergence patterns for $\alpha$, $g$, and $h$ are often not monotonic, and each parameter may temporarily increase or decrease during the initial iterations.

While convergence is almost always extremely fast, as in the example described above, some initial values cause 3-AR not to converge at all. In such rare cases, the value of $\alpha$ typically increases without bound, while $g$ and $h$ converge toward each other and ultimately become the same. This case corresponds to the trivial solution $\frac{f}{\alpha} \rightarrow 0 \leftarrow F^g - F^h$ in Equation (1) and is easy to detect and discard.

Figure 3 combines results for several noise-free S-distributions and essentially exhaustive sets of initial guesses for $g$ and $h$ satisfying $g < h$, as required. The selected distributions are representative for different shapes and skewness, which are reflected in different categories of parameter values (*cf.* Voit, 2000):

1. $g > 0$ and $h > 0$: as exemplified in Figure 3A and 3B;
2. $g < 0$ and $h > 0$: as exemplified in Figure 3C;
3. $g < 0$ and $h < 0$.

In addition, samples from all categories must by definition satisfy the condition $g < h$.

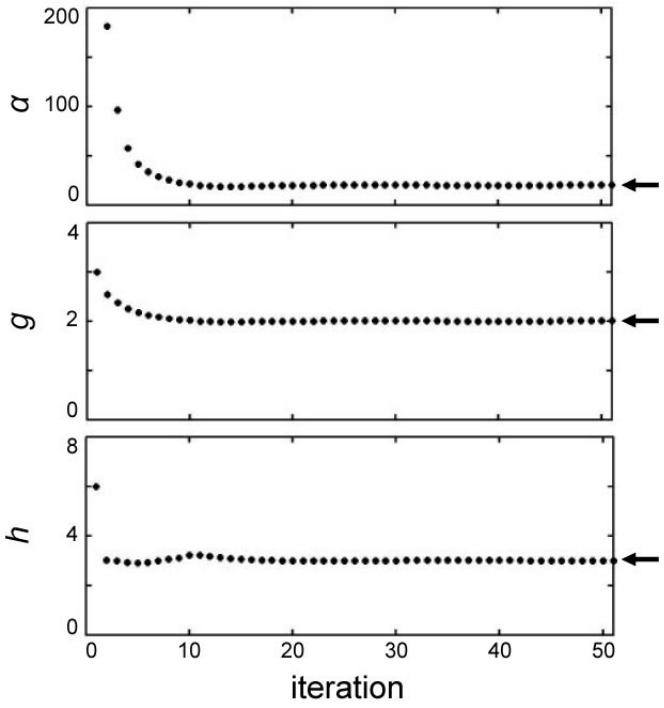


***Figure 2***: *Convergence pattern of 3-AR. For this example, 50 instances of the random variable were chosen from a parent distribution with parameters* $\alpha = 20$, $g = 2$, $h = 3$, *and* $F_0 = 0.01$. *Initial guesses were chosen as* $g = 3$ *and* $h = 6$, *but do not affect convergence much. No initial guess for* $\alpha$ *is needed in 3-AR.*

The left panels in Figure 3 exhibit the *cdf* and *pdf* of each distribution. Inserts show the so-called $f$-$F$ plots, where the *pdf* is plotted against the corresponding *cdf*. These plots are important because they are the basis for 3-AR and many other estimation methods for S-distributions. The right-hand panels present "heat maps" of convergence: the *x*- and *y*-axes represent the initial guesses of *h* and *g*, respectively, and the gray bar represents the logarithm (base 10) of the number of iterations needed for convergence. Once the predetermined error level is reached, 3-AR stops and the number of iterations is recorded as a measure for the speed of convergence. In each case shown here, 25 instances of the random variable were chosen and the corresponding noise-free $f$ and $F$ values were obtained according to the selected random variables. Black areas represent divergence to the trivial solution $\alpha \approx \infty$, $g \approx h$.

As discussed above, the convergence time for a given distribution does not vary much with different initial guesses, and the basin of convergence within each heat map is therefore almost monochrome. However, the heat maps of different distributions are quiet different. For instance, the times needed to generate the heat maps in Figures 3A, 3B, and 3C for a total of 57,600 initial values shown are 14,957, 1,197, and 1,094 seconds on a single PC, respectively, thus yielding average convergence times of 0.26, 0.021, and 0.019 seconds per case. While reasons for the wide variations in convergence times among distributions are unclear, the convergence *patterns* are similar in all cases: 3-AR takes big steps during the first few iterations, already coming very close to the true solution, and then spends many iterations on fine-tuning. The convergence area in each case is relatively large, and it seems to be a good general strategy to choose rather large, similar initial values for *g* and *h*, such as 10 and 10.5, to avoid divergence. Of importance is that each iteration consists essentially of three linear regressions, which are very fast. Thus, even if one encounters a rare case of divergence, the choice of alternative initial settings is computationally cheap and provides for effective estimation results.

Examples with $g < 0$ and $h < 0$ or with different $\alpha$ values are not shown in Figure 3, but 3-AR performed in a similar fashion for all cases tested. Most of the estimation tasks were solved very effectively, except for cases where the difference between *g* and *h* is large, for instance, $g = 0.1$ and $h = 6$. In such cases, the algorithm sometimes converges to sets of values between the true *g* and *h* and oscillates between them. A possible reason for this behaviour may be that in the $3^{rd}$ phase of regression (estimation of *h*), the slope of the regression line in the $\mathbf{y_h}$-$\mathbf{L_{F_h}}$ plot (which is reflected in the high value of *h*) is large and greatly affected by small errors, especially when $f$ and $F$ values are small so that their logarithms dominate the regression. In this case, the algorithm may not converge to exactly the right solution, but the oscillation happens within a reasonable range of parameter values. If it is desirable to obtain only one *g* and *h*, instead of ranges of oscillation that bound these values, a possible solution is to exclude some of the small $F$ values. In the cases we tested, this omission heuristically resulted in the algorithm converging to the true optimum.

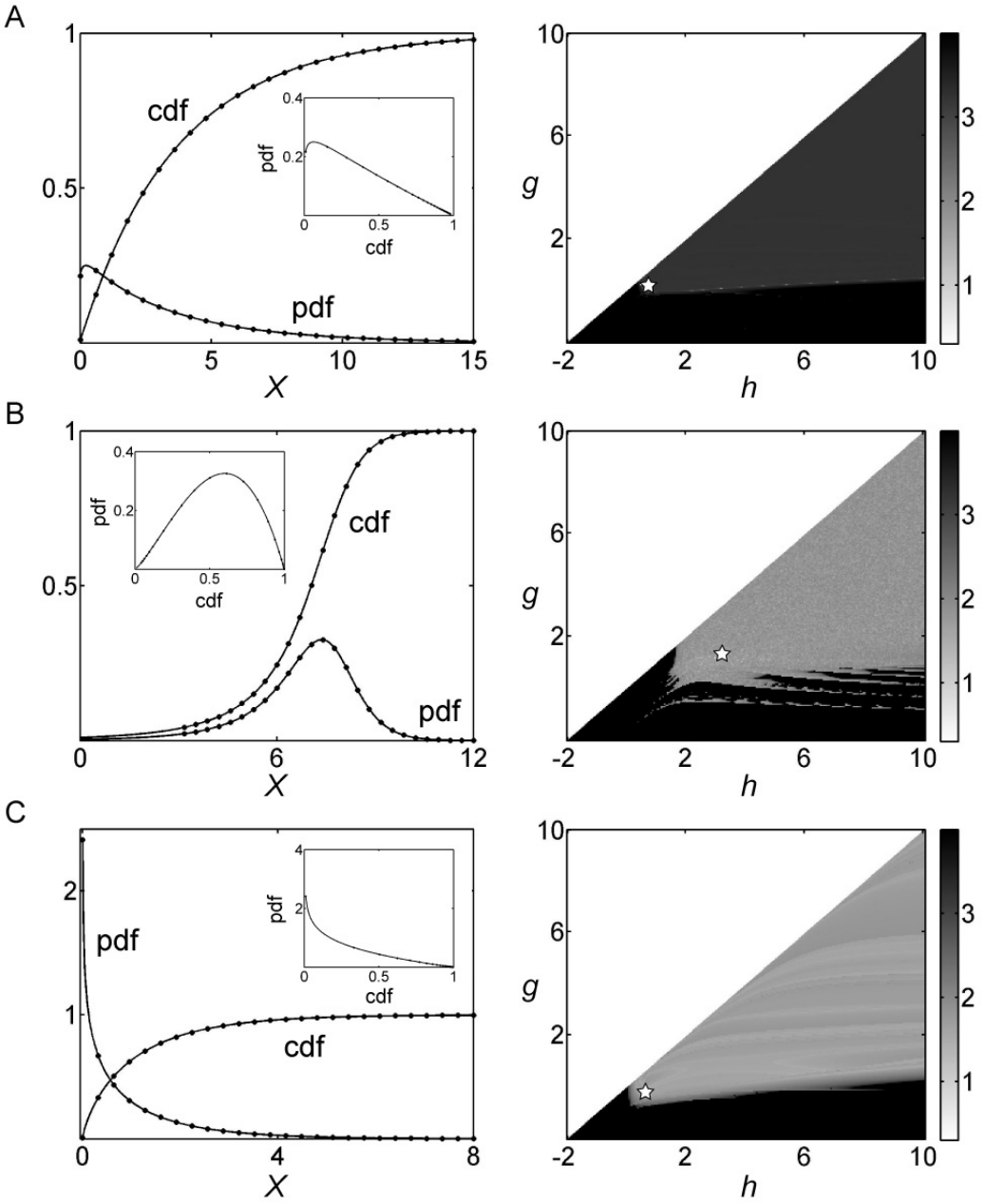


***Figure 3***: *Summary of convergence patterns of 3-AR. Panels on the left show the pdf, cdf, and $f$-$F$ plot (insert) of each distribution. Panels on the right present heat maps of convergence as functions of starting values of g and h, with gray bar indicating the logarithm (base 10) of the number of iterations needed for convergence. Each asterisk represents the true value of g or h. Case A:* $\alpha = 1$, $g = 0.25$, $h = 0.5$, $F_0 = 0.01$. *Case B:* $\alpha = 1$, $g = 1.2$, $h = 3$, $F_0 = 0.01$. *Case C:* $\alpha = 1$, $g = -0.2$, $h = 0.5$, $F_0 = 0.01$. *Twenty-five instances of the random variable were chosen in each case.*

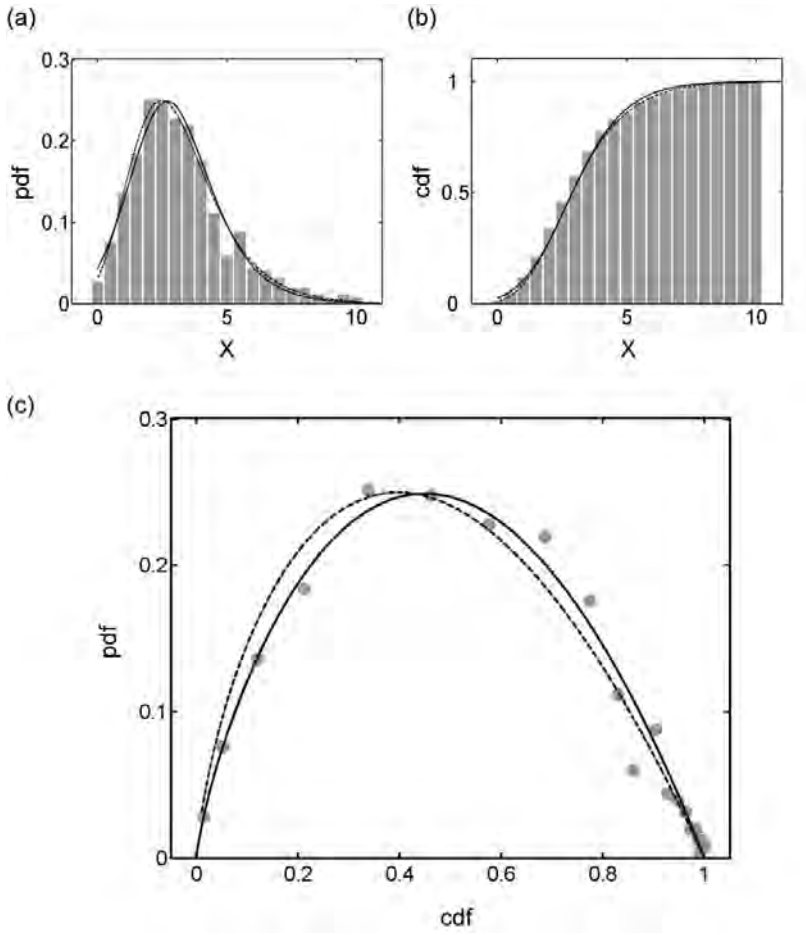


***Figure 4***: *Data sampled from an S-distribution with parameter values* $\alpha = 1$, $g = 0.75$, $h = 1.5$ *and fits with the parent S-distribution (dashed lines) and with an S-distribution obtained with 3-AR and initial guesses* $g = 10$ *and* $h = 10.5$ *(solid lines). Optimal parameter estimates are obtained as* $\alpha = 0.80$, $g = 0.78$, $h = 1.87$. *(a) pdfs; (b) cdfs; (c)* $f$*-*$F$ *plot showing the pdf as algebraic function of the cdf. SSE of the 3-AR optimized distribution is 0.0041 (*$S.D. = 0.0151$*), while SSE for the parent S-distribution is 0.0064 (*$S.D. = 0.0189$*).*

### *4.2 Fitting distributions with noise*

The preceding section discussed 3-AR for error-free samples from S-distributions. In this section we analyze finite random samples from S-distributions, which result in artificial datasets that appear noisy. To create these data, we use the quantile method, as discussed in Voit (2000). Specifically, we consider the inverted *cdf* equation

$$\frac{dX}{dF} = \frac{1}{\alpha\,(F^g - F^h)}, \qquad F(0.5) = median \tag{14}$$

and draw random numbers $R_i$ from the uniform distribution over (0,1), which are used as quantiles. Solving Equation (14) numerically upwards or downwards from the median to $F = R_i$ yields in $X_i$ the desired S-distributed random number. The S-distributed random numbers are collected and form the equivalent of an observed data sample, whose "noise" depends on the sample size.

The performance of 3-AR in fitting these artificial data is shown in Figure 4 with an example, where five hundred random numbers were generated from an S-distribution and categorized into 21 bins of a relative frequency histogram (Figure 4a). The *pdf* was constructed from the resulting histogram without smoothing and easily yielded the *cdf* (Figure 4b). The 3-AR algorithm converged within 47 iterations from the initial guesses $g = 10$ and $h = 10.5$ to the estimated solution. Interestingly, the fit with this solution is associated with a lower *SSE* than a fit with the parent S-distribution, from which the "data" were sampled, which confirms similar earlier observations (*e.g.*, Sorribas, March and Voit, 2000). To assess dependence on sample size, we also tested the algorithm with smaller sample sizes, *e.g.*, $n = 100$, and 3-AR performed similarly well.

To explore the flexibility of the S-distribution, we repeated the example shown in Figure 4 several times with 500 points each. The results (Figure 5) show slightly different fits with *SSE*s around 0.0045-0.0047 (Figure 5A), 0.0054-0.0057 (Figure 5B), and 0.0096 (Figure 5C), which are driven by the degree with which each random sample truly represents the underlying distribution. Within each class, the relationships between the estimates $\alpha$, $g$, and $h$ are similar, again confirming earlier results (Sorribas, March and Voit, 2000), where classes of quasi-equivalent S-distributions with quite similar *SSE*s were produced by fixing the value of $\alpha$ and fitting $g$ and $h$. In each class, $g$ and $h$ exhibit an almost linear relationship between each other and with $\log(\alpha)$ and converge to each other when $\alpha$ becomes larger. Even though the parameter sets within each class are clearly different, the resulting distributions are essentially indistinguishable.

In some cases, the 3-AR algorithm does not converge to a single value. Instead, it oscillates between reasonable candidate solutions. This is probably due to noise in the data, causing 3-AR to find the best "local" fit for each phase, which however is not the best fit for other phases. This behaviour is commonly seen in nonlinear algorithms. It is easy to find a suitable solution by choosing from among the candidate solutions, based on their *SSE*s.

### 4.3 Fitting traditional statistical distributions

The selection of a traditional distribution for fitting data is often difficult because the "true" parent distribution is typically not known. Testing candidate distributions one by one is cumbersome, and all-encompassing distribution families (*e.g.*, Savageau, 1982) often contain so many parameters that over-fitting and redundancy become complicating issues. Instead, the S-distribution may be used as an inclusive model that is capable of

A

| α | g | h | SSE |
|---|---|---|---|
| 1.134 | 0.884 | 1.748 | 0.0045 |
| 2.219 | 1.028 | 1.375 | 0.0046 |
| 1.005 | 0.848 | 1.805 | 0.0046 |
| 0.743 | 0.767 | 1.982 | 0.0047 |

B

| α | g | h | SSE |
|---|---|---|---|
| 0.611 | 0.704 | 2.098 | 0.0054 |
| 1.318 | 0.898 | 1.555 | 0.0055 |
| 1.830 | 0.959 | 1.429 | 0.0056 |
| 4.151 | 1.055 | 1.254 | 0.0057 |

C

| α | g | h | SSE |
|---|---|---|---|
| 1.221 | 0.899 | 1.785 | 0.0096 |
| 1.537 | 0.946 | 1.472 | 0.0096 |
| 2.403 | 1.004 | 1.360 | 0.0096 |

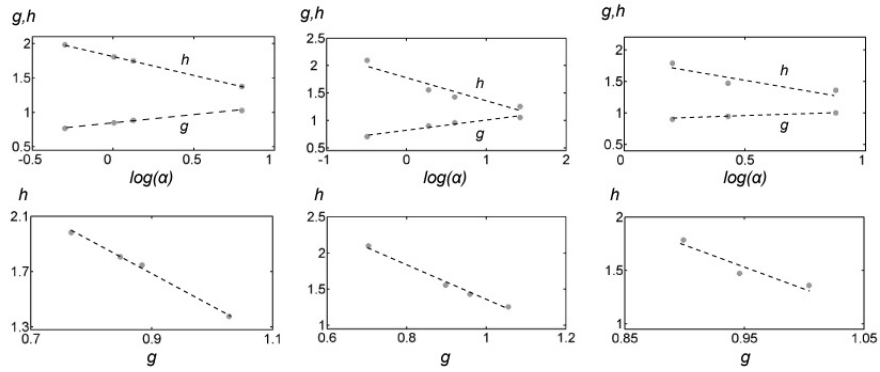


***Figure 5***: *Quasi-equivalent S-distributions. Parameters are estimated for different samples randomly generated from a given distribution ($\alpha = 1$, $g = 0.75$, $h = 1.5$). The residual errors SSEs are recorded and classified into three classes based on the value of SSE. The plots of g or h versus* $\log(\alpha)$ *and of g versus h are generated in each class. A: SSE between 0.0045 and 0.0047; B: SSE between 0.0054 and 0.0057; C: SSE equal to 0.0096.*

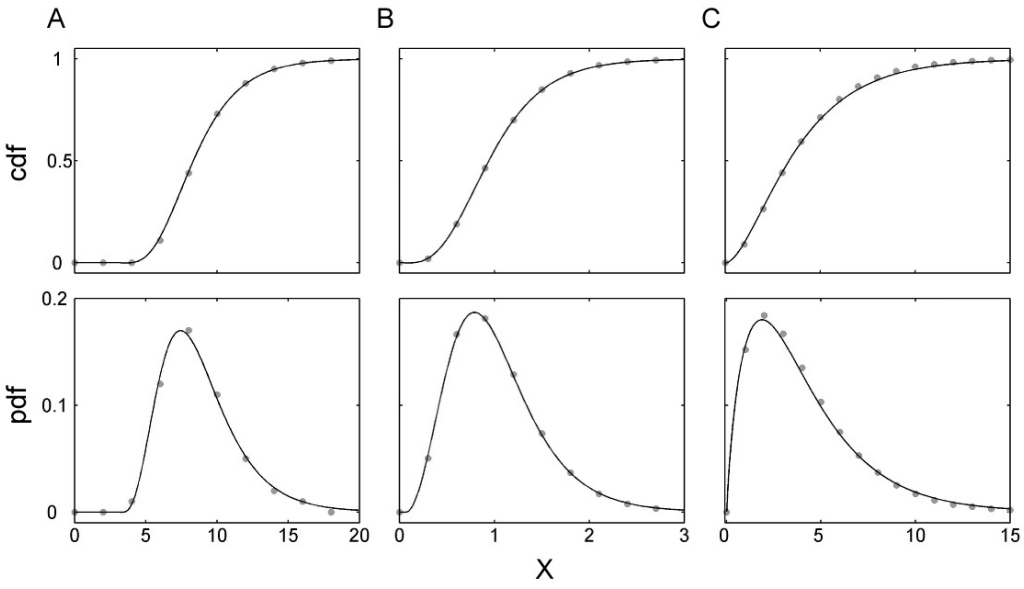


***Figure 6***: *Fitting traditional distributions. The gray dots represent data used in the regressions, while the solid curves represent the estimated S-distributions. The SSEs are calculated for the $f$-$F$ plot.A: noncentral $t_{8,8}$-distribution, SSE = 0.00007, S.D. = 0.0032; B: $F_{10,100}$-distribution, SSE = 0.00066, S.D. = 0.0097; C: $\chi^2_4$-distribution, SSE = 0.00026, S.D. = 0.0045.*

representing many traditional statistical distributions in sufficiently close approximation. The strategy thus becomes to fit data of unknown structure with an S-distribution and to identify which traditional distributions have similar shapes (Voit, 1992; Voit and Yu, 1994; Yu and Voit, 1996). This section explores how well 3-AR identifies S-distributions for random samples from traditional distributions.

The S-distribution contains only two classical distributions as special cases: the exponential distribution for $g = 0$ and $h = 1$ and the logistic distribution for $g = 1$ and $h = 2$. Fitting these two distributions yield *SSE*s equal to 0 (results not shown). All other classical distributions incur some unavoidable approximation error when modelled as S-distributions. Figure 6 shows the results of 3-AR fitting of three examples that are not special cases, namely a noncentral *t*-distribution, an *F*-distribution, and a $\chi^2$-distribution; the initial guesses were again chosen as $g = 10$ and $h = 10.5$. As before, 3-AR converges to a solution within a few iterations for these and many other examples. The only convergence problems occurred when fitting traditional distributions requiring $g \approx h$ (see Voit (1992) for these uncommon cases). A possible reason is presumably that the S-distribution is not a very good model for such distributions.

### 4.4 Fitting observed data

The ultimate measure of success of any fitting algorithm is the modelling of actual data. Figure 7 shows the performance of 3-AR in fitting an S-distribution to weight data of males ages 20 to 29 (data from *NHANES III* (National Center for Health Statistics, 1996)). The observed distribution contains 574 males, classified into bins of 3 kg. The *pdf* and *cdf* histograms were constructed in the same fashion as in Section 4.2. The *SSE* of the fit is similar to the result of using a constrained maximum likelihood estimator (Voit, 2000), although the parameter values are somewhat different, exhibiting again the flexibility and quasi-redundancy inherent in S-distributions. Visually, and judged by the *SSE*, the fit obtained here is satisfactory and obtained in less than a second.

## 5 Discussion

The S-distribution is a four-variable distribution that combines mathematical simplicity with superior flexibility in modelling data. A crucial prerequisite for using the distribution in practical applications is the availability of effective methods for estimating optimal parameter values from observed frequency data. Addressing this issue, we introduced here a method called *3-way alternating regression* (3-AR) that is extremely fast and robust. The 3-AR method constitutes a modification of a 2-way alternating regression method that was recently proposed for parameter estimation in S-systems (Chou, Martens and Voit, 2006), of which S-distributions are special cases.

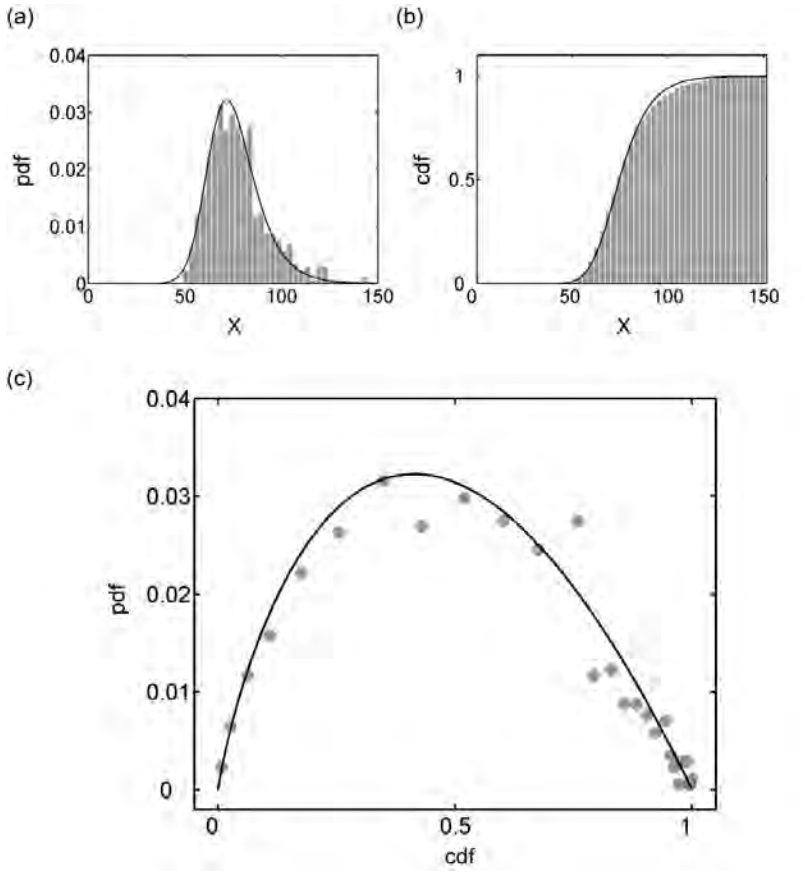


***Figure 7***: *Fitting observed data. Observed distribution (bars and dots) of weights of 574 males, ages 20-29 (National Center for Health Statistics, 1996) and S-distribution fit (lines) obtained with 3-AR and initial guesses* $g = 10$, $h = 10.5$. *Estimated parameter values:* $\alpha = 0.270$, $g = 0.958$, $h = 1.328$, $X_{0.5} = 74.37$. *(a) pdf (SSE = 0.000143,* $S.D. = 0.0023$*); (b) cdf (SSE =0.009629,* $S.D. = 0.0189$*); (c)* $f$*-*$F$ *plot (SSE = 0.000187,* $S.D. = 0.0026$*).*

The 3-AR method performs well in all typical scenarios, namely for estimating parameters from error-free distributions, from random samples generated from S-distributions, from traditional statistical distributions, and from actual data. The basin of convergence is rather large, and convergence speed is essentially independent of initial guesses that are selected to start the 3-AR algorithm. Therefore, even if one selects initial guesses quite far away from the true optimum, the algorithm only takes a few iterations to converge to points very close to the true solution and refines this solution with a relatively small number of further cycles. An exception is the situation where 3-AR converges to the trivial solution where $\alpha$ increases without bound and $g$ approaches $h$. This scenario is easy to spot and the choice of another initial guess typically remedies the situation. A second exception to rapid convergence may occur if the true $g$ and $h$ are very different. In this rather unusual case, the algorithm sometimes converges to values

between the true $g$ and $h$ and oscillates between them. In this case, one may select values from within the oscillation range or redo the estimation by omitting some of the very small values of the *pdf* and *cdf*.

The 3-AR fitting of data from traditional distributions works well in most cases, except for distributions that are not well approximated by S-distributions and where the relatively best fit requires $g \approx h$, as described in Section 4.3.

For finite random samples, the estimated solution is also obtained very quickly, but its parameters depend on the particular sample. As a consequence, the computed estimates may be rather different, even though the *SSE*s are very similar and the shapes of the resulting distributions are essentially indistinguishable. This finding is a manifestation of the shape flexibility and quasi-redundancy of S-distributions and confirms similar observations in the literature (*e.g.*, Sorribas, March and Voit, 2000).

The 3-AR algorithm provides a strategy for parameter estimation with S-distributions that is genuinely different from all other published methods. While some issues associated with the basin of convergence should be investigated further, our results shown here provide strong indication that this algorithm is much faster than the currently available alternatives.

An issue that seems generic to S-distributions and has been observed in other contexts is the covariance among the parameters $\alpha$, $g$, and $h$ (*e.g.*, Sorribas, March and Voit, 2000). While each set of these parameters determines a unique distribution, the covariance permits distinct sets leading to solutions that are so similar that their differences are often smaller than the noise in the data. This quasi-equivalence will require future work. For instance, it might be possible to specify the theoretical uncertainty variances of the estimated parameters or analytically study the uncertainty variance by principal component analysis or linear series expansion of the model around the convergence point ($\alpha$, $g$ and $h$).

Quasi-equivalence also poses problems when it is necessary to determine the uncertainty in the estimated parameters, for instance in the context of significance testing. The quasi-equivalent different parameter sets, which yield essentially indistinguishable distributions, are not arbitrary, but form slightly curved, essentially one-dimensional manifolds in the parameter space, as we and others have discussed in the literature several times. These manifolds may be similar to quasi-solution sets recently derived from Newton flow methods (see Dedieu and Shub, 2005). Whatever the structure of the quasi-solution sets may be, it is quite evident that equivalence tests focusing on one parameter at a time will not be useful. Instead, one will have to compare solutions globally, for instance based on Hellinger or Kullbach-Leibler distances (see Balthis, 1998) or on some measure of maximal distance, such as $Q_2 = sup_X|F_1(X) - F_2(X)|$. To calculate a confidence interval for these distances, one would probably use the bootstrap. One could similarly use bootstrap methods to calculate $p$-values for the null hypothesis that two S-distributions are the same, although the bootstrap sampling for hypothesis testing would be slightly different than that used

for confidence intervals. Furthermore, one could use Monte Carlo simulation methods to construct power curves for the alternative significance tests, under different true scenarios.

A related issue needing future attention will be the characterization of the intrinsic features of the 3-AR estimator, including its biasedness, consistency, and efficiency. These characterizations appear to be complex and may have to be postponed until the convergence behaviour of 3-AR is more fully understood.

Finally, a future extension of 3-AR might be its generalization to the more comprehensive GS-distribution (Muiño, Voit and Sorribas, 2006), which is characterized by increased flexibility in shape, in particular, for symmetric distributions, at the cost of one additional parameter. The inclusion of this additional parameter will require modifications to the 3-AR algorithm that need to be investigated in detail.

## Acknowledgments

This work was supported in part by a National Heart, Lung and Blood Institute Proteomics Initiative (Contract N01-HV-28181; D. Knapp, PI), a grant from the National Science Foundation (MCB 0517135; E. O. Voit, PI), and an endowment from the Georgia Research Alliance. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsoring institutions.

## References

Balthis, W. L. (1998). *Application of Hierarchical Monte Carlo Simulation to the Estimation of Human Exposure to Mercury via Consumption of King Mackerel (Scomberomorus cavalla)*, Ph.D. Dissertation, Medical University of South Carolina, Charleston, SC.

Chou, I. C., Martens, H. and Voit, E. O. (2006). Parameter estimation in biochemical systems models with alternating regression. *Theoretical Biology and Medical Modelling*, 25.

Dedieu, J.-B. and Shub, M. (2005). Newton flow and interior point methods in linear programming. *International Journal of Bifurcation and Chaos*, 15, 827-840.

National Center for Health Statistics. (1996). Analytic and Reporting Guidelines: The Third National Health and Nutrition Examination Survey, NHANES III (1988-1994). U.S. Department of Health and Human Services, Public Health Service, Center for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD.

Hernández-Bermejo, B. and Sorribas, A. (2001). Analytical quantile solution for the S-distribution, random number generation and statistical data modelling. *Biometrical Journal*, 43, 1007-1025.

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

Savageau, M. A. (1982). A suprasystem of probability distributions. *Biometrical Journal*, 24, 323-330.

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

Voit, E. O. (1992). The S-distribution. A tool for approximation and classification of univariate, unimodal probability distributions. *Biometrical Journal*, 34, 855-878.

Voit, E. O. (1996). Dynamic trends in distributions. *Biometrical Journal*, 38, 587-603.

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

Voit, E. O. and Almeida, J. (2000). Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics*, 20, 1670-81.

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

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

Voit, E. O. and Yu, S. (1994). The S-distribution. Approximation of discrete distributions. *Biometrical Journal*, 36, 205-219.

Yu, S. S. and Voit, E. O. (1996). A graphical classification of survival distributions. In: *Lifetime Data: Models in Reliability and Survival Analysis*, Jewell, N. P., Kimber, A. C., Lee, M-L. T. and Whitmore, G. A., Kluwer Academic Publishers, Dordrecht, 385-392.

SORT 31 (1) January-June 2007, 75-96

# Nonparametric bivariate estimation for successive survival times

Carles Serrat* and Guadalupe Gómez**

*Universitat Politècnica de Catalunya*

**Abstract**

Several aspects of the analysis of two successive survival times are considered. All the analyses take into account the dependent censoring on the second time induced by the first. Three nonparametric methods are described, implemented and applied to the data coming from a multicentre clinical trial for HIV-infected patients. Visser's and Wang and Wells methods propose an estimator for the bivariate survival function while Gómez and Serrat's method presents a conditional approach for the second time given the first. The three approaches are compared and discussed at the end of the paper.

*MSC:* 62N02, 62G05

*Keywords:* AIDS; Conditional survival; Dependent censoring; Inverse probability of censoring weighted estimators; Successive survival times.

## 1 Introduction

The survival experience of a population often involves two times of interest. The estimation of their joint survivor function is of intrinsic interest since it is useful in predicting the joint survival experience, in estimating the degree of dependence, in model building and testing and in strengthening marginal analysis.

* *Address for correspondence:* Departament de Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Avda. Dr. Marañón, 44–50, 08028 Barcelona, Spain. Tel: +34 93 405 44 86. Fax +34 93 401 63 06. E-mail: carles.serrat@upc.edu

** Guadalupe Gómez. Departament d'Estadística i I.O., Universitat Politècnica de Catalunya, Edifici C5, Despatx 210, Jordi Girona 1–3, 08034 Barcelona, Spain. E-mail: lupe.gomez@upc.edu

Received: November 2006
Accepted: February 2007

These two survival times of interest could be naturally paired, for instance, in twin studies when analyzing time to death of each sibling, in oncology studies when the interest is the time to cancer detection in the left and right breast, in ophthalmology when recording the time to severe visual loss on the left and right eyes. In these cases, several possibly dependent failure processes act concurrently, and henceforth, they ought to be modelled jointly.

In many other situations there is a natural ordering of the times of occurrence of events. For instance, in any clinical study, time to diagnosis precedes time to start treatment which in turns precedes times to cure. In AIDS studies, time to HIV infection precedes the time to AIDS diagnosis, which in turn precedes the time to death due to AIDS. In food science studies, when referring to climacteric fruits, the time to maturation precedes the time to senescentia.

In univariate survival studies, right censoring usually precludes the complete observation of the time to event variable. When we have two survival times of interest the censoring mechanism could either be the same for both variables or act differently on each one. For instance, when analyzing the joint behaviour of the ages of cancer diagnosis in each breast the censoring – due to loss of follow-up or end of study – acts simultaneously on each breast. However, when studying the population of twins who have suffered a heart attack, the follow-up time, and hence the censoring, could be independent for each twin. When one event precedes the second, the censoring mechanism acting on the second and subsequent times will depend not only on the total time of follow-up but also in the value of the first and preceding times. When this situation arises the methods to estimate the joint survival, or functions of the joint, have to handle the special case of dependent and informative censoring induced by the previous failure times.

The motivation of the paper comes from the Tibet (Guided-**T**reatment **I**nterruption **Be**nefi**t**) study. Tibet is a multicentre, open label clinical trial with blinded and centralized randomization conducted to investigate the safety and clinical benefits of an intermittent antiretroviral therapy guided by CD4+ T-cell counts and plasma HIV-1 RNA in patients with chronic HIV-1 infection with more than 500 CD4+ counts/mm$^3$ and undetectable HIV-1 RNA. Patients were randomized to follow either the intermittent guided therapy which is described below or to continue with their prior HAART (Highly Active Antiretroviral Therapy). Details of the study are described in Ruiz *et al* (2007).

In this work we restrict our attention to the interruption group in which the patients interrupt therapy until CD4+ counts reach values equal or inferior to 350 cells/mm$^3$, plasma viral load increases to 100000 copies/ml or a severe acute retroviral syndrome takes place or an AIDS-defining illness occurs. If any of these events occurs, the prior HAART is reinitiated and maintained until CD4+ counts increases to 500 cells/mm$^3$ or more and viral load reaches 80 copies/ml, at which time HAART is again discontinued as previously described.

This intermittent therapeutic strategy process defines, for each patient, a sequence of alternative stages without HAART (OFF) and with HAART (ON). Various lifetime variables can be defined within this process, for instance, $T_1$ is the first time OFF, that is, the time (in weeks) from randomization (and therefore interruption of HAART therapy) to first reinitiation of treatment, $T_2$ is the first time ON, that is the time from the first treatment reinitiation until the next interruption, $T_3$ is the second time OFF, that is, the number of weeks from the second HAART interruption to the second reinitiation of treatment, and so on.

Apart from the number of scientific and clinical questions that such an study poses, there are also a number of relevant statistical issues which arise due to this particular data set. In particular, the dependent censoring mechanism that affects $T_2$, and subsequent times, as a consequence of having an administrative censoring time $C$, invalidates the standard methods of survival analysis for right-censored data and requires alternative approaches.

We now review the most relevant papers concerning the estimation of the bivariate survival function as well as the estimation of the conditional survival function from pairs of random variables which might be right-censored. Campbell (1981) and Campbell and Földes (1982) propose several nonparametric estimators for the bivariate survival function in the presence of independent pairs of censoring variables that are independent of the failure times. Their main idea is based on the factorization of the bivariate distribution function as a product of the distribution function for the bivariate vector of interest and the distribution function for the censoring variables. These estimators are shown to be strongly uniform consistent at a rate of convergence equal to that of the empirical distribution function. All the estimators they propose, however, are not legitimate survival functions since they are not necessarily monotone increasing in both coordinates. Tsai, Leurgans and Crowley (1986) propose a family of closed form estimators that are always survival functions based on a decomposition of the bivariate survival in terms of identifiable survival and subsurvival functions. Their estimators are fairly complicated and have a rate of convergence slower than Campbell and Földes. Burke (1988) proposes an estimator based on the representation of the bivariate distribution function as the convolution of the subdistribution function, which can be naturally estimated by the observed data, and the inverse of the bivariate distribution function for the censoring times. Burke's approach only uses the information provided by the uncensored observations, throwing away the relevant information of censored data points. Dabrowska (1988) proposes an estimator for the bivariate survival function based on an empirical estimator for the bivariate cumulative hazard. This estimator is almost surely consistent and weakly convergent. Unfortunately, its computation is quite complicated and the covariance function of these estimators cannot be estimated analytically. More details can be found in Gómez *et al.* (2004).

Visser (1996), Wang and Wells (1998) and Gómez, Serrat and Ruiz (2007) approach different aspects of the nonparametric bivariate survival estimation problem. Visser

derives the nonparametric maximum likelihood estimate for the conditional hazard of $T_2$ given a fixed value of $T_1$ under the assumption that the two durations are discrete. Wang and Wells present an estimator for the cumulative conditional hazard of $T_2$ given $T_1 > t_1$ following Nelson-Aalen's construction of the cumulative hazard estimator but where each observation has been weighted using the information on the first duration to unbias the effect of dependent censoring. Due to the limited applicability of Visser's estimator, since lifetime data are genuinely continuous, and the lack of interpretability of Wang and Wells's parameter of interest in the case of two ordered times $T_1$ and $T_2$ where the observation of the second time, $T_2$, is conditioned on the observation of the first time, Gómez, Serrat and Ruiz (2007) propose a weighted conditional estimation for the survival of $T_2$ on a given category of $T_1$.

We introduce the notation and assumptions for the rest of the paper in Section 2 and develop the methods of Visser, Gómez and Serrat and Wang and Wells in Sections 3, 4 and 5 respectively. Each of these three sections starts with a description of the method, continues with some software considerations and ends with a specific analysis of the Tibet clinical trial. The three approaches are compared and discussed at the end of the paper.

## 2 Notation

Assume that $T_1$ and $T_2$ represent two consecutive duration variables corresponding to two different events at times $T_1$ and $T_1 + T_2$, respectively, which are measured from the start of the follow-up. The follow-up time is subject to independent right censoring by $C$. Note that $T_1$, $T_2$ and $T_1 + T_2$ are independent of $C$. However $T_2$, which is subject to right censoring by $C - T_1$, is not independent of $C - T_1$ unless $T_1$ is independent of $T_2$. In this situation, we cannot use conventional survival methods for independent or noninformative censorship models. Whenever the censoring random variable for a given time depends on other random times we say that we are in the framework of a dependent censoring mechanism.

Define the marginal and bivariate survival functions for $(T_1, T_2)$ as $S_1(t_1) = \Pr\{T_1 > t_1\}$ and $S_{12}(t_1, t_2) = \Pr\{T_1 > t_1, T_2 > t_2\}$. Denote by $G(t) = \Pr\{C > t\}$ the survival function corresponding to the total time of follow-up $C$. If $\tau_C = \sup\{t : G(t) > 0\}$ is the maximum follow-up time, the bivariate survival function, $S_{12}(t_1, t_2)$, is only estimable for $t_1 + t_2 \leq \tau_C$. This restriction is analogous to the non-estimability of the Kaplan-Meier estimator beyond those values larger than the total follow-up time. It follows as well that the marginal distributions for $T_2$ cannot be estimated by the Kaplan-Meier method. Note that if $T_1$ and $T_2$ are positively correlated, even under independent censoring, persons with long $T_1$'s are more likely to have long $T_2$'s and hence more likely to be censored.

For a given individual we observe a vector $(Y_1, Y_2, D_1, D_2)$ where for every $j = 1, 2$, $Y_j = \min\{T_j, C_j\}$, $D_j = 1\{T_j \leq C_j\}$, $C_1 = C$, $C_2 = (C - T_1)1\{T_1 \leq C\}$.

Note that when

i. $D_1 = 0 = D_2$: the two durations are right-censored and thus $Y_1 = C$, $Y_2 = 0$ and no information about $T_2$ is available

ii. $D_1 = 1, D_2 = 0$, $T_1$ is observed while $T_2$ is right-censored by $C - T_1$, which implies that $T_2$ is right-censored by a dependent variable if $T_1$ and $T_2$ are correlated.

iii. $D_1 = 1, D_2 = 1$, $T_1$ and $T_2$ are observed.

Our estimation problem is to be based on a random sample $\{(T_{1i}, T_{2i}, C_i), i = 1, \ldots, n\}$ of $(T_1, T_2, C)$ from which the observed sample is $\mathcal{S} = \{(Y_{1i}, Y_{2i}, D_{1i}, D_{2i}),\ i = 1, \ldots, n\}$. We also consider $\mathcal{S}^*$, a subset of $\mathcal{S}$, consisting of those observations for which $T_1$ is observed, that is, $\mathcal{S}^* = \{(Y_{1i}, Y_{2i}, D_{1i}, D_{2i}) \in \mathcal{S} | D_{1i} = 1, \quad i = 1, \ldots, n\} \subset \mathcal{S}$.

Note that when $D_{1i} = 0$, no crude information about $T_{2i}$ is available. However, these subjects provide information about $T_1$, which is supposed to be dependent on $T_2$. Thus, these missing data ($\{i : D_{1i} = 0\}$), are not at random because the probability of being observed for $T_2$ depends on $T_1$. As a consequence, inferences for $T_2$ cannot be only based on the subset $\mathcal{S}^*$, and we will have to use these partially observed individuals to infer about the law of $T_2$.

## 3 Visser's method. A discrete approach

### 3.1 Introduction to the methodology

Visser (1996) proposes a nonparametric estimator for the bivariate survival function when the two duration variables are always observed in a particular order and the censoring mechanism acts on their sum.

Visser starts assuming that $T_1, T_2$ and $C$ are discrete random variables taking values in $\{0, 1, 2, \ldots, K\}$, and therefore $Y_1, Y_2$, defined in Section 2 as $Y_j = \min\{T_j, C_j\}$, $j = 1, 2$, are discrete as well. Due to the fact that the random variables $T_1, T_2$ and $C$ are supposed to be discrete and take a finite number of values, Visser defines the corresponding survival distributions at each time $t$ as the probability of being greater or equal than $t$ as follows

$$\begin{aligned} S_{T_1,T_2}(k, l) &= \Pr\{T_1 \geq k, T_2 \geq l\} \\ S_{T_1}(k) &= \Pr\{T_1 \geq k\} \end{aligned}$$

$$\begin{aligned} \lambda_{T_1}(k) &= \Pr\{T_1 = k | T_1 \geq k\} \\ G(k) &= \Pr\{C \geq k\}. \end{aligned}$$

Visser factorizes $S_{T_1,T_2}(k,l)$ as the product of the conditional and the marginal as follows,

$$S_{T_1,T_2}(k,l) = S_{T_1}(k) S_{T_2|T_1}(l|k) \tag{1}$$

On the other hand, the product expression of the survival functions in terms of the hazard functions allows to write for $k, l = 1, 2, \ldots, K$:

$$S_{T_1}(k) = (1 - \lambda_{T_1}(0)) \ldots (1 - \lambda_{T_1}(k-1)) \tag{2}$$

$$S_{T_2|T_1=k}(l) = \Pr\{T_2 \geq l | T_1 = k\} = (1 - \lambda_{T_2|T_1=k}(0)) \ldots (1 - \lambda_{T_2|T_1=k}(l-1)) \tag{3}$$

where $\lambda_{T_2|T_1=k}(l) = \Pr\{T_2 = l | T_1 = k, T_2 \geq l\}$.

Remark as well that $S_{T_2|T_1}(l|k)$ can be written as follows:

$$\begin{aligned} S_{T_2|T_1}(l|k) &= \Pr\{T_2 \geq l | T_1 \geq k\} = \frac{\Pr\{T_2 \geq l, T_1 \geq k\}}{\Pr\{T_1 \geq k\}} \\ &= (S_{T_1}(k))^{-1} \sum_{j=k}^{K} S_{T_2|T_1=j}(l)(S_{T_1}(j) - S_{T_1}(j+1)) \end{aligned} \tag{4}$$

Equalities (1) and (4) imply that in order to estimate $S_{T_1,T_2}(k,l)$ we only need to estimate $S_{T_1}(k)$ and $S_{T_2|T_1=j}(l)$. The estimation of $S_{T_1}(k)$ is straightforward through the Kaplan-Meier estimator.

Denote by $n_{1k}, n_{2kl}, n_{3kl}$ the following counting processes: $n_{1k} = \sum_{i=1}^{n} 1\{ Y_{1i} = k, \delta_i = 1\}$, $n_{2kl} = \sum_{i=1}^{n} 1\{Y_{1i} = k, Y_{2i} = l, \delta_i = 2\}$ and $n_{3kl} = \sum_{i=1}^{n} 1\{Y_{1i} = k, Y_{2i} = l, \delta_i = 3\}$. That is, $n_{1k}$ counts the number of censored individuals at $k$ months (for these individuals $T_1 > k$ and $T_2$ is not defined), $n_{2kl}$ counts the number of individuals whose first duration is equal to $k$ months and who are censored after $k + l$ months (for these individuals $T_1 = k$ and $T_2 > l$) and $n_{3kl}$ counts the number of subjects with a first duration equal to $k$ months and a second duration equal to $l$ months (for these individuals $T_1 = k$ and $T_2 = l$). Denote as well $n_{k\cdot} = \sum_{l=1}^{K}(n_{2kl} + n_{3kl})$ which counts the total number of individuals whose $T_1 = k$ irrespective of their status on $T_2$.

Visser proves (see Appendix for more details) that the nonparametric MLE for $\lambda_{T_1}(k)$ is given by

$$\hat{\lambda}_{T_1}(k) = \frac{\sum_{i=1}^{n} 1\{Y_{1i} = k, \delta_i \geq 2\}}{\sum_{i=1}^{n} 1\{Y_{1i} \geq k\}} = \frac{\sum_{l=1}^{K}(n_{2kl} + n_{3kl})}{n_{1k} + \sum_{l=1}^{K}(n_{2kl} + n_{3kl})} \tag{5}$$

which yields the discrete time Kaplan-Meier estimator for $S_{T_1}(k)$ after replacing it in (2). On the other hand, the nonparametric MLE for $\lambda_{T_2|T_1=k}(l)$ is given by

$$\hat{\lambda}_{T_2|T_1=k}(l) = \sum_{i=1}^{n} \frac{1\{Y_{1i} = k, Y_{2i} = l, \delta_i = 3\}}{\sum_{i=1}^{n} 1\{Y_{1i} = k, Y_{2i} \geq l\}}. \tag{6}$$

Replacing $\hat{\lambda}_{T_2|T_1=k}(l)$ in (3) provides the MLE for $S_{T_2|T_1=k}(l)$, which in turn can be replaced in (4) to obtain an estimator for $S_{T_2|T_1}(l|k)$. Finally, everything could be replaced in (1) to get the bivariate nonparametric estimator for $S_{T_1,T_2}(k, l)$.

Visser proves that both estimators, $\hat{\lambda}_{T_1}(k)$ and $\hat{\lambda}_{T_2|T_1=k}(l)$, are consistent asymptotically normal after normalizing by $\sqrt{n}$, and asymptotically independent. These facts, together with the $\delta$ method, imply that $\sqrt{n}(\hat{S}_{T_1,T_2}(k, l) - S_{T_1,T_2}(k, l))$ is asymptotically normal, mean zero and with an asymptotic variance that can be estimated replacing the unknown functions by their estimators.

The survival function $G$ of the censoring variable appears in the expression for the variances. It may be estimated by the product-limit method.

### 3.2 Implementation

We have implemented in S-PLUS a function `bwv21` that computes the conditional survival for $T_2$, given a value $T_1 = t_1$, according to expression (3). The function uses as parameters the observed values of $T_1$ and $T_2$, as well as, the corresponding censoring indicators ($D_1$ and $D_2$).

After estimating the conditional survival we can compute the joint survival $S_{T_1,T_2}(k, l)$ in (1), by using the function `c2jv` that implements the expression given in (4).

All the S-PLUS functions that we have implemented are available at the web page of the GRASS group at `http://www-eio06.upc.es/grass`.

### 3.3 TIBET project: A discrete time analysis

In the Tibet clinical trial, one hundred HIV-patients were recruited between May 2001 and January 2002 and randomly assigned to interrupt HAART. The interim closing date for the study was July 15, 2004.

Figure 1 shows the empirical survival estimator corresponding to the follow-up time of each patient. Based on this estimation, the probability of being followed 96 weeks or more is 93%, the median follow-up time is 130 weeks, the third quartile is 146 weeks and the maximum follow-up time being 188 weeks. Furthermore, the effective minimum follow-up has been 96 weeks.

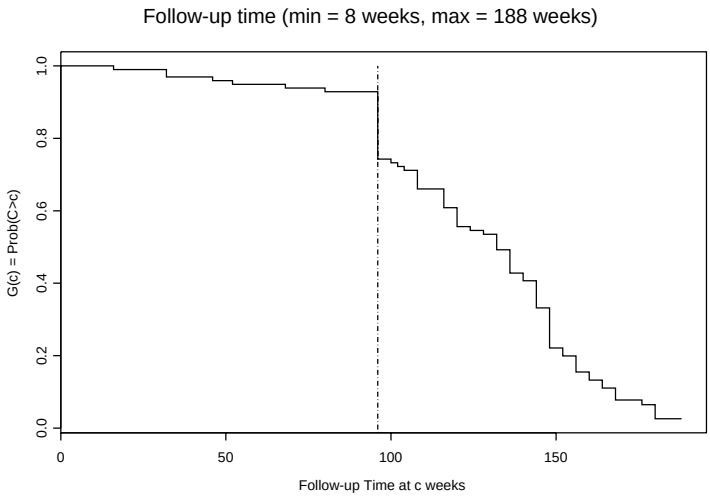


***Figure 1***: *Survival estimator for the time to follow-up.*

### 3.3.1 Conditional estimation of the first time ON given the first time OFF

Using the function `bwv21` we have obtained the results in Table 1. This table illustrates the finite grid of 30 observed times for $T_1$ by 23 observed times for $T_2$, and the corresponding estimation of $S_{T_2|T_1=t_1}(t_2)$ for each pair $(t_1, t_2)$ of the grid. For instance, the median time of being with treatment among those who have been 8 weeks interrupted is approximately 23 weeks. Note, however, that although the median and even the third quartile are estimable for $t_1 = 8$, for longer interrupted times the estimation is either not possible or quite rough.

As a matter of fact, Visser's discrete method does not provide efficient estimates of the conditional survival function due to the drastic reduction of the sample size. This drawback is due to the fact that for a fixed time $T_1 = t_1$ the sample size is not large enough to make inferences on $T_2|T_1 = t_1$. In particular, the sample size is dramatically small for $t_1 > 40$. Furthermore, the small number of events for $T_2$ makes the estimation of $S_{T_2|T_1=t_1}(.)$ quite hopeless. This fact is still more problematic for high values of $t_1$, where the proportion of censoring for $T_2$ is higher (in some cases 100%).

### 3.3.2 Joint distribution estimation

Based on the results in Table 1 we have computed the joint survival $S_{T_1,T_2}(k, l)$ in (1). Table 2 shows the results of this estimation for a selection of times in $T_1$ and $T_2$.

It is important to remark that expression in (4) can not be directly computed from the data because, as we have seen in Table 2, $S_{T_2|T_1=j}(l)$ is not estimable for some pairs

***Table 1***: *Estimates for the conditional survival of $T_2$ given values in $T_1$, $\hat{S}_{T_2|T_1=t_1}(t_2)$.*

| | $t_1$=6 | 8 | 9 | 10 | 12 | 13 | 16 | 20 | 24 | 26 | 28 | 32 | 36 | 40 | 44 | 48 | 56 | 60 | 64 | 68 | 72 | 76 | 88 | 92 | 96 | 100 | 108 | 116 | 136 | 144 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $t_2$=8 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | – | 1 | – | 1 | 1 | 1 |
| 10 | 0.83 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | – | 1 | – | 1 | 1 | 1 |
| 14 | 0.83 | 1 | 1 | 0.50 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | – | 1 | – | 1 | 1 | 1 |
| 15 | 0.83 | 1 | 0.50 | 0.50 | 1 | 1 | 1 | 1 | 1 | – | 1 | 1 | 1 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | – | 1 | – | 1 | 1 | 1 |
| 16 | 0.83 | 0.94 | 0.50 | 0.50 | 1 | 1 | 0.67 | 1 | 1 | – | 1 | 1 | 1 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | – | 1 | – | 1 | 1 | 1 |
| 19 | 0.83 | 0.88 | 0.50 | 0.50 | 1 | 1 | 0.67 | 1 | 1 | – | 1 | 1 | 1 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | – | 1 | – | 1 | 1 | 1 |
| 20 | 0.83 | 0.76 | 0.50 | 0.50 | 1 | 1 | 0.33 | 1 | 1 | – | 1 | 1 | 1 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.50 | 1 | – | 1 | – | 1 | 1 | 1 |
| 22 | 0.50 | 0.76 | 0.50 | 0.50 | 1 | 1 | 0.33 | 1 | 1 | – | 1 | 1 | 1 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | – | 0.50 | 1 | – | – | – | 1 | 1 | – |
| 24 | 0.50 | 0.41 | 0.50 | 0.50 | 1 | 1 | 0.33 | 1 | 1 | – | 0.75 | 1 | 0.50 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | – | 0.50 | 1 | – | – | – | 1 | 1 | – |
| 25 | 0.50 | 0.41 | 0.0 | 0.50 | 1 | 1 | 0.33 | 1 | 1 | – | 0.75 | 1 | 0.50 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | – | 0.50 | 1 | – | – | – | 1 | – | – |
| 26 | 0.17 | 0.41 | – | 0.50 | 1 | 1 | 0.33 | 1 | 1 | – | 0.75 | 1 | 0.50 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | – | 0.50 | 1 | – | – | – | 1 | – | – |
| 28 | 0.17 | 0.29 | – | 0.50 | 1 | 1 | 0.33 | 0 | 1 | – | 0.50 | 1 | 0.50 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | – | 0.50 | 1 | – | – | – | 1 | – | – |
| 32 | 0.17 | 0.18 | – | 0.50 | 1 | 1 | 0.33 | – | 0.67 | – | 0.25 | 0 | 0.50 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | – | 0.50 | 1 | – | – | – | 1 | – | – |
| 34 | 0 | 0.18 | – | 0.50 | 1 | 1 | 0.33 | – | 0.67 | – | 0.25 | – | 0.50 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | – | 0.50 | 1 | – | – | – | 1 | – | – |
| 36 | – | 0.12 | – | 0.50 | 1 | 1 | 0.33 | – | 0.67 | – | 0.00 | – | 0.50 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | – | 0 | 0 | – | – | – | 1 | – | – |
| 39 | – | 0.12 | – | 0.50 | 1 | 0 | 0.33 | – | 0.67 | – | – | – | 0.50 | 0.67 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | – | – | – | – | – | – | 1 | – | – |
| 40 | – | 0.12 | – | 0.50 | 1 | – | 0.33 | – | 0.33 | – | – | – | 0.50 | 0.67 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | – | – | – | – | – | – | 1 | – | – |
| 44 | – | 0.12 | – | 0.50 | 1 | – | 0.33 | – | 0.33 | – | – | – | 0.50 | 0.67 | – | 1 | 1 | 0 | 1 | 1 | 1 | – | – | – | – | – | – | 1 | – | – |
| 46 | – | 0.12 | – | 0 | 1 | – | 0.33 | – | 0 | – | – | – | 0.50 | 0.67 | – | 1 | 1 | – | 1 | 1 | 1 | – | – | – | – | – | – | – | – | – |
| 48 | – | 0.12 | – | – | 1 | – | 0.33 | – | – | – | – | – | 0.50 | 0.67 | – | 1 | 1 | – | 0 | 1 | 1 | – | – | – | – | – | – | – | – | – |
| 52 | – | 0.12 | – | – | 1 | – | 0.33 | – | – | – | – | – | 0.50 | 0.67 | – | 1 | 0.50 | – | – | 1 | 1 | – | – | – | – | – | – | – | – | – |
| 60 | – | 0.06 | – | – | 1 | – | 0.33 | – | – | – | – | – | 0.50 | 0.67 | – | 1 | 0.50 | – | – | 1 | – | – | – | – | – | – | – | – | – | – |
| 100 | – | 0.06 | – | – | – | – | – | – | – | – | – | – | 0.50 | – | – | – | – | – | – | 0 | – | – | – | – | – | – | – | – | – | – |

***Table 2***: *Estimates for some selected times of the joint survival of* $(T_1, T_2)$ *using Visser's approach.*

| | $t_1 = 6$ | 12 | 24 | 48 | 72 | 96 | 120 |
|---|---|---|---|---|---|---|---|
| $t_2 = 12$ | 0.878 | 0.678 | 0.598 | 0.460 | 0.377 | 0.325 | 0.312 |
| 24 | 0.428 | 0.315 | 0.241 | 0.164 | 0.080 | 0.054 | 0.054 |
| 36 | 0.206 | 0.198 | 0.155 | 0.116 | 0.033 | 0.022 | 0.022 |
| 48 | 0.100 | 0.098 | 0.084 | 0.062 | 0.010 | – | – |
| 60 | 0.078 | 0.077 | 0.064 | 0.042 | – | – | – |
| 100 | 0.016 | 0.015 | 0.015 | – | – | – | – |

$(j, l)$. As a first approximation we have omitted the contribution of these terms in those pairs in which the estimation of the conditional survival has not been possible. This fact produces two important drawbacks: on one hand, the method is not efficient and, on the other hand, in general there is an underestimation of the corresponding bivariate survival distribution.

## 4 Gómez and Serrat's method. A stratified approach

Driven by the Tibet clinical trial where it is of special interest to characterize the number of weeks on treatment that a patient needs in order to recover their virological and immunological levels given that he/she has spent a certain number of weeks without treatment, we propose an estimator for the survival of $T_2$ on a given category of $T_1$.

Let $0 < \tau_1 < \tau_2 < \ldots < \tau_M$ be the $M$ times of interest for a particular study. For convenience define $\tau_0 = 0$ and consider $\tau_{M+1}$ as an arbitrary value larger than $\tau_M$.
Let $T_1^*$ be a discrete version of $T_1$ defined as follows:

$$T_1^* = \begin{cases} \tau_k & \text{if} \quad \tau_{k-1} < T_1 \leq \tau_k \quad k = 1, \cdots, M \\ \tau_{M+1} & \text{if} \quad T_1 > \tau_M. \end{cases}$$

Note that the election of the representative of each class is not relevant for the results. Denote the conditional cumulative hazard function for $T_2$ given $T_1^* = \tau_k$ by $\Lambda_{T_2|T_1^*=\tau_k}(db)$ and the conditional survival function for $T_2$ given $T_1^* = \tau_k$ by

$$S_{T_2|T_1^*=\tau_k}(v) = \Pr(T_2 > v|T_1^* = \tau_k) = \Pr(T_2 > v|\tau_{k-1} < T_1 \leq \tau_k)$$

for $k = 1, \ldots, M$ and $v > 0$. The factorization of the survival function in terms of the conditional cumulative hazard is straightforward:

$$S_{T_2|T_1^*=\tau_k}(v) \quad = \quad \prod_{b \leq v} \{1 - \Lambda_{T_2|T_1^*=\tau_k}(db)\}. \tag{7}$$

When estimating the survival of $T_2$ given a certain category of $T_1$ we could legitimately apply the conditional Kaplan-Meier if censoring for $T_2$ would be non informative, that is, if individuals with different $T_1$ values, within a given category, had the same chances of being at risk for different values of $T_2$. However, this might not be the case if the categories are quite wide and, in this case, we will have to take into account the effect that the dependent censoring caused by $T_1$, within each strata, is producing on $T_2$.

In the next subsection we show how would affect the dependent censoring on the estimations and we propose how to adjust the Kaplan-Meier survival estimates for $T_2$ on each strata to unbias the effect produced by $T_1$ and we propose a weighted conditional estimator for $S_{T_2|T_1^*=\tau_k}(t_2) = \Pr(T_2 > t_2|\tau_{k-1} < T_1 \leq \tau_k)$, adjusted by the dependent censoring.

### 4.1 Weighted conditional methodology

Denote by $R_{T_2}(b|\tau_k)$ the risk set of $T_2$ at time $b$ given $T_1^* = \tau_k$. Under the dependent censoring structure the risk set $R_{T_2}(b|\tau_k)$ for estimating $\Lambda_{T_2|T_1^*=\tau_k}(db)$ may not be homogeneous, as is shown in Theorem 1.

**Theorem 1** *The probability of being at risk at time b for the second duration $T_2$ for an individual with first duration equal to $T_1 = t_{1i}$ depends on $G(t_{1i} + b)$.*

*Proof:* An observation $i$ with the first duration $T_1 = t_{1i}$ affects the probability of the corresponding $T_{2i}$ being included in $R_{T_2}(b|\tau_k)$, as we see in the following expression:

$$\begin{aligned}\Pr\{i \in R_{T_2}(b|\tau_k)\} &= \Pr\{Y_{1i} \in t_{1i}, \tau_{k-1} < t_{1i} \leq \tau_k, D_{1i} = 1, Y_{2i} \geq b\} \\ &= \Pr\{T_1 \in t_{1i}, \tau_{k-1} < t_{1i} \leq \tau_k, T_2 \geq b\} G(t_{1i} + b).\end{aligned}$$

where $T_1 \in t_{1i}$ is the abbreviation of $T_1 \in (t_{1i}, t_{1i} + \Delta)$ as $\Delta \to 0$. □

Therefore, the conditional Kaplan-Meier produces biased results because the value of $t_{1i}$ affects the probability of the corresponding $T_{2i}$ being included in $R_{T_2}(b|\tau_k)$. To adjust this heterogeneity, one can weight each observation in $R_{T_2}(b|\tau_k)$ by an estimate of the reciprocal of $G(t_{1i} + b)$.

We define the conditional cumulative hazard estimator as follows:

$$\begin{aligned}\widehat{\Lambda}_{T_2|T_1^*=\tau_k}(\Delta b) &= \frac{\sum_{i \in R_{T_2}(b|\tau_k)} 1\{Y_{2i} = b, D_{2i} = 1\}/\hat{G}(t_{1i} + b)}{\sum_{i \in R_{T_2}(b|\tau_k)} 1\{Y_{2i} \geq b\}/\hat{G}(t_{1i} + b)} \\ &= \frac{\sum_{i=1}^{n} 1\{\tau_{k-1} < Y_{1i} \leq \tau_k, D_{1i} = 1, Y_{2i} = b, D_{2i} = 1\}/\hat{G}(Y_{1i} + b)}{\sum_{i=1}^{n} 1\{\tau_{k-1} < Y_{1i} \leq \tau_k, D_{1i} = 1, Y_{2i} \geq b\}/\hat{G}(Y_{1i} + b)} \qquad (8)\end{aligned}$$

for every $b$ such that $\max_{1\leq i\leq n} Y_{1i} + b < \hat{\tau}_C$ where $\hat{\tau}_C = sup\{t : \hat{G}(t) > 0\}$ is the observed maximum follow-up time and where $\hat{G}(\cdot)$ is the empirical survival computed from the follow-up times.

This estimator has a potential problem when $\hat{G}(\cdot) = 0$. The convention $0/0 = 0$ is used to avoid the misdefinition. However, in many clinical trials, and in particular in the one that motivated our work, the follow-up time $C$ is a continuous variable which is observed for all the individuals and hence $\hat{G}(\cdot) \neq 0$, except for the largest follow-up time.

A nonparametric estimator, $\widehat{S}_{T_2|T_1^*=\tau_k}(v)$, for the conditional survival function is obtained by plugging (8) into (7) as follows:

$$\widehat{S}_{T_2|T_1^*=\tau_k}(v) \quad = \quad \prod_{b\leq v}\{1 - \widehat{\Lambda}_{T_2|T_1^*=\tau_k}(db)\}. \tag{9}$$

Asymptotic properties of $\widehat{\Lambda}_{T_2|T_1^*=\tau_k}(\Delta b)$ and $\widehat{S}_{T_2|T_1^*=\tau_k}(v)$, as well as related issues to the estimation of the variance of $\widehat{S}_{T_2|T_1^*=\tau_k}(v)$ via a bootstrapping methodology can be found in Gómez *et al.* (2004 and 2007). A simulation study illustrating its good behaviour when the sample size is moderate is included in Gómez *et al.* (2007).

### 4.2 Implementation

We have implemented in S-PLUS the inverse probability of censoring weighted (IPCW) conditional methodology introduced in the previous section. The main function in the library is called `bwwce21` and its syntax is the following:

```
bwwce21(vartimes1,varcens1,vartimes2,varcens2,breaks,wmet,vtfw,vcfw)
```

where

```
vartimes1 = first time variable (T1 by default),
varcens1 = censoring indicator for the first time (D1 by default),
vartimes2 = second time variable (T2 by default),
varcens2 = censoring indicator for the second time (D2 by default),
breaks = partition values ({12, 24, 48, 96} by default),
wmet = weighting method for the dependent censoring
   (0=no weights, 1=follow-up -default-, 2=T1+T2, 3=T1+T2+T3),
vtfw = follow-up time variable (TFW by default),
vcfw = censoring indicator for the follow-up time (DFW by default).
```

Function `bwwce21` allows to reproduce the conditional Kaplan-Meier estimator attending the categories in the variable $T_1$, by setting no weights (`wmet`=0) in the call.

### 4.3 TIBET project: Conditional estimation of the first time ON given the first time OFF

We illustrate the conditional estimator, given in Subsection 4.1, for the estimation of the survival of the first time with treatment conditioned to the first time without treatment. Clinicians were very interested in the survival pattern of the first time with treatment, $T_2$, for patients who had been short, medium and long times without treatment, $T_1$. Based on these considerations they fixed the times of interest for the conditional analysis in $\tau_1 = 12$ (one trimester), $\tau_2 = 48$ (one year) and $\tau_3 = 96$ (two years), and according to this partition we have the following three categories in $T_1$: $T_1 \leq 12$, $12 < T_1 \leq 48$ and $48 < T_1 \leq 96$. Among the 100 patients, there are 31 with $T_1$ right-censored ($D_1 = 0$). Among the 69 patients with $T_1$ observed ($D_1 = 1$), there are 15 patients with $T_2$ right–censored ($D_2 = 0$) and 51 patients with $T_1$ and $T_2$ observed ($D_1 = D_2 = 1$).

In Figure 1 we observed that, except for a few number of subjects, there is a common minimum follow-up of 96 weeks. As a consequence, if we are estimating the survival at time $T_2 = v$ in the category $T_1^*$, the standard Kaplan-Meier estimator would be enough if $v + T_1^* \leq 96$ (for example, for $v \leq 84$ weeks when $T_1^* = 12$ or for $v \leq 48$ weeks when $T_1^* = 48$) and, on the other hand, we will appreciate the correction of the bias due to the dependent censoring when $v + T_1^* \geq 96$ by using the proposed weighted methodology.

***Table 3***: *Estimates and standard errors (computed using bootstrap) for some selected times of the conditional survival of $T_2$ given the following three categories: $T_1$: $T_1 \leq 12$, $12 < T_1 \leq 48$ and $48 < T_1 \leq 96$*

| $t_2$ | (0,12] | (12,48] | (48,96] |
|---|---|---|---|
| 12 | 0.964 (0.0328) | 0.954 (0.0463) | 1 (0) |
| 24 | 0.464 (0.1273) | 0.727 (0.1048) | 0.802 (0.1265) |
| 36 | 0.143 (0.0695) | 0.410 (0.1028) | 0.571 (0.1938) |
| 48 | 0.107 (0.0588) | 0.228 (0.0889) | 0.386 (0.1674) |
| 60 | 0.071 (0.0539) | 0.228 (0.0885) | 0.298 (0.1566) |
| 100 | 0.071 (0.0511) | 0.228 (0.0881) | 0 (0.1889) |

Table 3 provides the estimates and standard errors for some selected times of the conditional survival of $T_2$ given categories in $T_1$. The standard errors have been computed using bootstrap. Based on these results (see Table 4) we estimate that while a patient who needs treatment quite fast ($0 < T_1 \leq 12$) will be as well fast in recovering his/her CD4 and viral load levels (median equals to 23), those patients which are able to stay a bit longer without treatment ($12 < T_1 \leq 48$) take longer time to recover levels (median equals to 30) and those patients which are able to stay much longer without treatment ($48 < T_1 \leq 96$) take much longer time to recover levels (median equals to 42). This behaviour can be explained by introducing as a covariate the cause of treatment reinitiation (plasma viral load > 100000 copies/ml and/or CD4+ counts ≤ 350 cells/mm$^3$). Those patients that reinitiate treatment because viral load has become

***Table 4***: *Description of $T_2$ given categories of $T_1$*

| | $T_1 \leq 12$ | $12 < T_1 \leq 48$ | $48 < T_1 \leq 96$ |
|---|---|---|---|
| Size | 29 | 22 | 13 |
| Events | 26 | 17 | 8 |
| 1st Q | 20 | 23 | 35 |
| Median | 23 | 30 | 42 |
| 3rd Q | 30 | 45 | 80 |

higher than 100000 copies/ml, do so quite fast and their immunological system has not had time to be deteriorated. Since treatment is design to control viral replications these patients need shorter times to reach an undetectable viral load and still CD4 > 500. On the other hand, although some patients are able to stay without treatment long enough because they can keep viral load below 100000, the immunological system is slowly, but constantly, deteriorating. As a consequence, once they start treatment they need a longer period to recover the immunological level.

We present in Figure 2 the standard conditional Kaplan-Meier estimator together with the proposed weighted estimator of the conditional survival function $\widehat{S}_{T_2|T_1^*=\tau_k}(v)$ given in (9), for each of these categories. Note that in the first two categories both estimators coincide due to the long common minimum follow-up as we have previously noticed. Figure 2 also illustrates that the time that a patient needs to recover their immunological and virological levels depends on the time that he/she has been without treatment, as we have already observed in the previous paragraph. We can also clearly see a different behaviour between the survival of the times on treatment for patients that

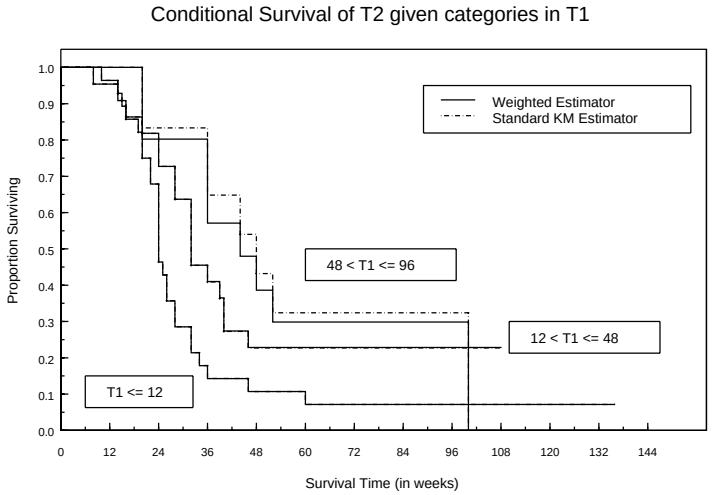


***Figure 2***: *Plots of the conditional survival function of $T_2$ given $T_1$ in the following three categories: $T_1 \leq 12$, $12 < T_1 \leq 48$ and $48 < T_1 \leq 96$ for both the standard conditional Kaplan-Meier and the proposed weighted estimator given in (9).*

stay without treatment less than 12 weeks as opposed to those that stay without treatment more than 48 weeks.

## 5 Wang and Wells' method. A continuous approach

### 5.1 Wang and Wells' estimator

Wang and Wells (1998) propose a path-dependent (nonparametric) estimate for the joint survival function of two duration variables.

According to the notation introduced in Section 2, let the observed sample be $\mathcal{S} = \{(Y_{1i}, Y_{2i}, D_{1i}, D_{2i}), i = 1, \ldots, n\}$, and $\mathcal{S}^* = \{(Y_{1i}, Y_{2i}, D_{1i} = 1, D_{2i}), i = 1, \ldots, n\}$ the subset of $\mathcal{S}$ consisting of those observations for which $T_1$ is observed. Wang and Wells consider the following path-dependent decomposition

$$\begin{aligned} S_{12}(t_1, t_2) &= \Pr\{T_2 > t_2 | T_1 > t_1\} \Pr\{T_1 > t_1\} \\ &= \prod_{v \le t_2} \{1 - \Lambda_{T_2|T_1>t_1}(dv)\} S_1(t_1) \end{aligned} \tag{10}$$

where $\Lambda_{T_2|T_1>t_1}(dv)$ is the cumulative conditional hazard of $T_2$ given $T_1 > t_1$. Wang and Wells propose to estimate $S_{12}(t_1, t_2)$ via estimable components for $\Lambda_{T_2|T_1>t_1}(dv)$ and for $S_1(t_1)$ and to plug them into (10). The estimation of the marginal $S_1(t_1)$ is accomplished using the Kaplan-Meier estimator based on the observables $(Y_{1i}, D_{1i})$ $(i = 1, \ldots, n)$.

The estimator for $\Lambda_{T_2|T_1>t_1}(dv)$ extends Campbell and Földes estimator so that dependent censoring is taking into account. First note that if we let $R_{T_2}(v|t_1)$ be the risk set of $T_2$ at time $v$ given $T_1 > t_1$, if $v > 0$ then $R_{T_2}(v|t_1) \subset \mathcal{S}^*$. An observation $i$ with the first duration $T_1 = t_{1i}$ affects the probability of the corresponding $T_{2i}$ being included in $R_{T_2}(v|t_1)$ as we see in the following expression

$$\begin{aligned} \Pr\{i \in R_{T_2}(v|t_1)\} &= \Pr\{Y_{1i} \in t_{1i}, t_{1i} > t_1, D_{1i} = 1, Y_{2i} \ge v\} \\ &= \Pr\{T_1 \in t_{1i}, t_{1i} > t_1, T_2 \ge v\} G(t_{1i} + v). \end{aligned}$$

Hence they adjust this heterogeneity by weighting each observation in $R_{T_2}(v|t_1)$ by an estimate of $1/G(t_{1i} + v)$.

Wang and Wells' estimator for $\Lambda_{T_2|T_1>t_1}(dv)$ can be expressed as follows:

$$\begin{aligned} \hat{\Lambda}^{WW}_{T_2|T_1>t_1}(\Delta v) &= \frac{\sum_{i \in R_{T_2}(v|t_1)} 1\{Y_{2i} = v, D_{2i} = 1\} / \hat{G}(t_{1i} + v)}{\sum_{i \in R_{T_2}(v|t_1)} 1\{Y_{2i} \ge v\} / \hat{G}(t_{1i} + v)} \\ &= \frac{\sum_{i=1}^{n} 1\{Y_{1i} > t_1, D_{1i} = 1, Y_{2i} = v, D_{2i} = 1\} / \hat{G}(Y_{1i} + v)}{\sum_{i=1}^{n} 1\{Y_{1i} > t_1, D_{1i} = 1, Y_{2i} \ge v\} / \hat{G}(Y_{1i} + v)} \end{aligned} \tag{11}$$

where $\hat{G}(\cdot)$ is an appropriate estimator of $G(\cdot)$ computed from the follow-up data. For example, $\hat{G}(\cdot)$ can be the Kaplan-Meier estimator of $G(\cdot)$ computed from the data $(Y_{1i} + Y_{2i}, 1 - D_{1i}D_{2i})$ $(i = 1, \ldots, n)$.

Wang and Wells' estimator for $S_{T_2|T_1>t_1}(t_2) = \Pr\{T_2 > t_2 | T_1 > t_1\}$ is given by plugging $\hat{\Lambda}^{WW}_{T_2|T_1>t_1}(\Delta v)$ into (10):

$$\widehat{S}^{WW}_{T_2|T_1>t_1}(t_2) \quad = \quad \prod_{v \le t_2} \{1 - \hat{\Lambda}^{WW}_{T_2|T_1>t_1}(dv)\}. \tag{12}$$

and the corresponding estimator for $S_{12}(t_1, t_2)$ is given by

$$\hat{S}_{12}(t_1, t_2) \quad = \quad \widehat{S}^{WW}_{T_2|T_1>t_1}(t_2)\hat{S}_1(t_1) \tag{13}$$

Their estimator uses the information on the first duration to weight each observation to unbias the effect of dependent censoring. This estimator has a potential problem with the existence of $\hat{S}_{12}(t_1, t_2)$ when $\hat{G}(\cdot) = 0$. If the largest value of $Y_{1i} + Y_{2i}$, say $c_{(n)}$, is censored ($D_{1i}D_{2i} = 0$), then the largest observation of the censoring variables is observed ($1 - D_{1i}D_{2i} = 1$) and hence $\hat{G}(c_{(n)}) = 0$. However, in this case the numerator in (11) is also 0 and the convention $0/0 = 0$ can be used. Note that the marginal survivor function can be estimated by $\hat{S}_2(t_2) = \hat{S}_{12}(0, t_2)$.

Wang and Wells show that $\hat{S}_{12}(t_1, t_2)$ converges in probability to $S_{12}(t_1, t_2)$ and claim that the limit distribution of $\sqrt{n}(\hat{S}_{12}(t_1, t_2) - S_{12}(t_1, t_2))$ converges weakly to a zero-mean Gaussian process, but the variance of the limiting process is quite complex and is not given.

### 5.2 Joint survival considerations for ($T_1$, $T_2$)

The bivariate estimator is useful in predicting the joint survival experience, in estimating the degree of dependence, in model building and testing and in strengthening marginal analysis. Furthermore, it is a necessary step if we want to compare $\widehat{S}_{T_2|T_1^*=\tau_k}(v)$, given in (9), to the estimator of $S_{T_2|T_1^*=\tau_k}(v)$ obtained from Wang and Wells' approach.

An estimator, $\hat{S}^{WW}_{12}(t_1, t_2)$, for the bivariate survival function of $(T_1, T_2)$ is obtained plugging $\widehat{S}_1(t)$ and $\widehat{S}^{WW}_{T_2|T_1>t_1}(t_2)$, given in (12), into (13). This estimator suffers from two drawbacks: it is not a legitimate survival function and is dependent on the selected path and ordering of the components. We propose to isotonize $\hat{S}^{WW}_{12}$ so that the survival function is monotone in both components. Denote by $\hat{S}^{isot}_{12}$ the isotonic version of $\hat{S}^{WW}_{12}$.

### 5.3 Related issues

On one hand, by using the joint survival $\hat{S}^{isot}_{12}$ introduced in the previous subsection, we can also derive an estimator for the survival of $T_2$ conditioned on the categories in $T_1$

just defining $\hat{S}^{isot}_{T_2|T_1^*=\tau_k}(v)$ as follows:

$$\hat{S}^{isot}_{T_2|T_1^*=\tau_k}(v) = \frac{\hat{S}^{isot}_{12}(\tau_{k-1}, v) - \hat{S}^{isot}_{12}(\tau_k, v)}{\hat{S}_1(\tau_{k-1}) - \hat{S}_1(\tau_k)}, \tag{14}$$

This is an alternative estimator for $S_{T_2|T_1^*=\tau_k}(v)$.

On the other hand, when the investigator is interested in the bivariate survival distribution in some prefixed intervals of time (for instance, we might be interested in the survival behaviour for every year) the estimation of the following joint function $f_{T_1^*,T_2}(\tau_k, v) = \Pr(T_1^* = \tau_k, T_2 > v)$ is particularly appealing. $f_{T_1^*,T_2}(\tau_k, v)$ can be factorized following the path-dependent decomposition: $f_{T_1^*,T_2}(\tau_k, v) = \Pr\{T_2 > v | T_1^* = \tau_k\} \cdot \Pr\{T_1^* = \tau_k\} = S_{T_2|T_1^*=\tau_k}(v) \cdot \Pr\{T_1^* = \tau_k\}$.

As a consequence, $f_{T_1^*,T_2}(\tau_k, v)$ is estimated straightforwardly using the nonparametric estimation of $S_{T_2|T_1^*=\tau_k}(v)$ provided in (9) and from which we know asymptotic properties, and an estimator, $\widehat{\Pr}\{T_1^* = \tau_k\}$, for $\Pr\{T_1^* = \tau_k\}$. To estimate $\Pr\{T_1^* = \tau_k\}$ we simply estimate the marginal survival function of $T_1$, $S_1(\tau_k)$, using the Kaplan-Meier estimator, $\widehat{S}_1(\tau_k)$, and replace accordingly, that is, $\widehat{\Pr}\{T_1^* = \tau_k\} = \hat{S}_1(\tau_{k-1}) - \hat{S}_1(\tau_k)$.

### 5.4 Implementation

In a similar way that for the weighted conditional estimator in Section 4 we implemented in S-PLUS the function `bwww21` to estimate the joint survival distribution of $(T_1, T_2)$ according to the Wang and Wells estimator in (13). The basic syntax of `bwww21` is: `bwww21(vartimes1, varcens1, vartimes2, varcens2, wmet, vtfw, vcfw)` where the parameters for the function are the same as the ones described for the `bwwce21` function in Subsection 4.2.

Specific computations for $\hat{\Lambda}^{WW}_{T_2|T_1>t_1}(dv)$ in (12) have been implemented in the function `lww21`. The function `bwww21` also uses the function `isoton` that performs the isotonization of a matrix so that the corresponding survival function is monotone in both components. It is important to note that, in order to avoid successive steps isotonizing by rows and columns alternatively, with non-unique results, our algorithm applies in one single-step an upper-left triangular minimization (see the corresponding code below). Related functions are `is.isoton` and `isotonv` that have been implemented to check if a matrix is isotonic and to isotonize a vector, respectively.

```
isoton <- function(mat) {
 n.r <- dim(mat)[1]
 n.c <- dim(mat)[2]
 mati <- mat
 for(j in 2:n.c) mati[1,j] <- min(mati[1, c(j-1,j)])
 for(i in 2:n.r) mati[i,1] <- min(mati[c(i-1, i),1])
```

```
 for(j in 2:n.c) for(i in 2:n.r)
   mati[i,j] <- min(mati[i-1,j], mati[i,j], mati[i,j-1])
 mati
}
```

Finally, the conditional survival $\hat{S}^{isot}_{T_2|T_1^*}$ in (14) has been implemented in the function `j2c`.

### 5.5 TIBET project: Joint survival estimation

In the same way that in Section 4 the follow-up time variable, $TFW$, is the information on the censoring that we have for each patient.

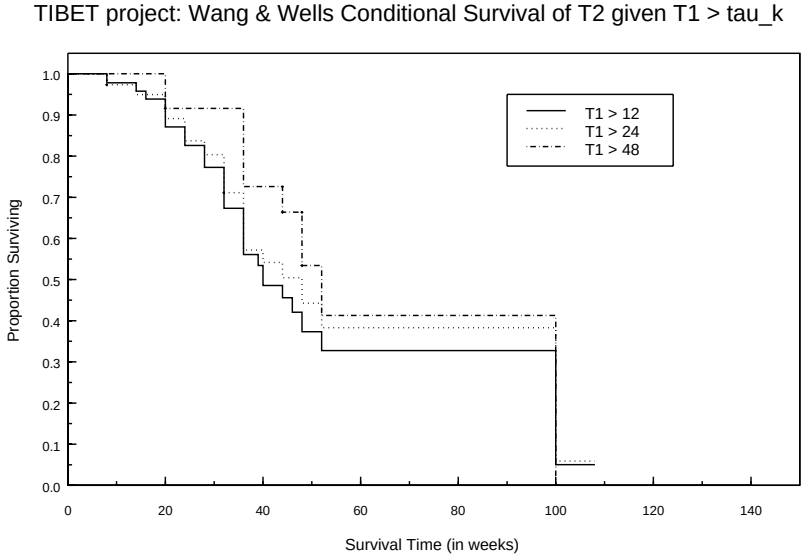


***Figure 3***: *Plots of the conditional survival function of $T_2$ given categories in $T_1$. $T_1$ represents the number of weeks without treatment and is splitted into three categories: $T_1 > 12$, $T_1 > 24$ and $T_1 > 48$. Each curve represents the survival function of the first time with treatment on each of the categories.*

Figure 3 illustrates Wang and Wells estimator $\widehat{S}^{WW}_{T_2|T_1>t_1}(t_2)$ given in (12) for the following three categories defined by $T_1$: $T_1 > 12$, $T_1 > 24$ and $T_1 > 48$. We see from these curves that patients who stay without treatment more than 48 weeks, will stay with treatment longer times than those patients who stayed OFF more than 24 weeks.

In Table 5 we illustrate the isotonic joint survival estimator for $\hat{S}_{12}(t_1, t_2)$, $\hat{S}^{isot}_{12}$, proposed in Subsection 5.2. We can see that the joint estimation it is not feasible for those pairs $(t_1, t_2)$ with no events for $T_2$, with $T_2 > t_2$, in the category $T_1 > t_1$.

***Table 5***: *Estimates for some selected times of the joint survival of* $(T_1, T_2)$ *using Wang and Wells method.*

| | $t_1 = 0$ | 6 | 12 | 24 | 48 | 72 | 96 | 120 |
|---|---|---|---|---|---|---|---|---|
| $t_2 = 12$ | 0.972 | 0.926 | 0.694 | 0.613 | 0.489 | 0.405 | 0.353 | 0.312 |
| 24 | 0.691 | 0.666 | 0.586 | 0.527 | 0.448 | 0.360 | 0.353 | 0.312 |
| 36 | 0.394 | 0.394 | 0.394 | 0.360 | 0.355 | 0.207 | 0.207 | – |
| 48 | 0.266 | 0.266 | 0.265 | 0.256 | 0.256 | – | – | – |
| 60 | 0.219 | 0.219 | 0.219 | 0.219 | 0.202 | – | – | – |
| 100 | 0.050 | 0.050 | 0.036 | 0.033 | 0.000 | – | – | – |

With respect to isotonizing the resulting joint survival in (13), note that we did not need to isotonize in more than 60% of the points. On the other hand, the resulting differences in the rest of the points have $-0.157$, $-0.025$ and $-0.010$ as quartiles.

Table 6 gives the estimates for the conditional survival of $T_2$ given categories in $T_1$, $\hat{S}^{isot}_{T_2|T_1^*=\tau_k}(v)$, derived from Wang and Wells method. In more than 70% of the points isotonization has not been necessary.

***Table 6***: *Estimates for some selected times of the conditional survival of* $T_2$ *given categories in* $T_1$ *after estimating the joint survival distribution of* $(T_1, T_2)$ *using Wang and Wells method.*

| $t_2$ | (0,12] | (12,48] | (48,96] |
|---|---|---|---|
| 12 | 0.960 | 0.930 | 1 |
| 24 | 0.360 | 0.627 | 0.697 |
| 36 | 0.000 | 0.130 | 0.697 |
| 48 | 0.000 | 0.005 | – |
| 60 | 0.000 | 0.005 | – |
| 100 | 0.000 | 0.005 | – |

After comparing with the weighted conditional estimator that we have proposed in the Section 4 (see Table 3), we can see that the conditional estimator derived from Wang and Wells approach underestimates, in general, the corresponding survival.

## 6 Discussion

In this paper we have illustrated three different approaches to analyze two successive survival times. The main difficulty in this type of study is the presence of the dependent censoring induced by the potential relationship between both times of interest. All the approaches consider the estimation either of the joint distribution of $(T_1, T_2)$ or the conditional distribution of $T_2$ given $T_1$. The main difference between the proposed methods is on the conditioning strategy and the way of considering the correction of the bias due to the dependent censoring.

Visser's method is based on the direct estimation of the conditional survival given an specific value for $T_1$, and it does not correct for the effect of the dependent censoring. As we mentioned in Section 3 the main restriction of this methodology is that it needs an important initial sample in order to obtain, after conditioning, a sample size that is large enough to estimate efficiently the conditional survival.

On the other hand, the weighted conditional estimation proposed by Gómez and Serrat provides an unbiased estimator for the conditional survival function for the second survival time given the categories in the first survival time. The main interest of this approach is that it takes into account the heterogeneity due to the dependent censoring by using all the information provided by $T_1$ to weight the observed data. In this sense, Gómez and Serrat's estimator is a good alternative to Visser's method because it does not need a discretization of the time variables and it allows to perform the estimation when the sample size is not very large. We remark here that although our parameter of interest is based on the categories of a first survival time, we use the continuous survival times, $t_{1i}$, without discretizing them, to contribute to the inverse weight, $G(t_{1i}+b)^{-1}$, and furthermore we do not need to discretize the second survival time $T_2$.

Concerning Wang and Wells' estimator, it is important to remark that the conditioning part is based on the subsets $T_1 > t_1$ and that the methodology corrects for the dependent censoring. However, the resulting estimator for the joint survival is not isotonic and, as a consequence, it is not a proper distribution. Hence the derivation of other functions of the bivariate survival distribution, for instance the conditional survival in the Tibet clinical trial, is questionable. As we noticed in Subsection 5.2 an alternative could be to isotonize the resulting estimates, however, as we can see after comparing Tables 3 and 6, this strategy provides a quick-to-zero survival distribution that underestimates the parameter of interest. In this sense, the weighted conditional estimation is also an interesting alternative to Wang and Wells' estimator because it avoids the non-desirable effects of the isotonization.

It is important to note that the proposed Gómez and Serrat's estimator can depend on the partition and the resulting estimates can be sensitive to the sample size in each category as well as to the number of different observed times $T_2$ in the category. In practice and for the Tibet clinical trial study, we have also analyzed the dataset using other partitions, for instance splitting $T_1$ into the following four categories: $T_1 < 12$, $12 < T_1 \leq 24$, $24 < T_1 \leq 48$ and $48 < T_1 \leq 96$, and similar results are obtained. In fact, in order to choose a partition for the analysis, it is necessary to take into account not only the resulting sample size in each category but also the number of different observed times for $T_2$ in the category.

Extensions of Gómez and Serrat's approach to the estimation of the survival function for other successive duration times given the information on the first are under consideration for the authors, by studying the effect of the intermediate events in the estimation of the appropriated weights for each subject. In the Tibet study it could be of interest, for instance, to estimate the duration of the second period OFF, $T_3$, given the category of the duration of the first period OFF, $T_1$.

All the approaches in this paper have been implemented in S-PLUS and they are easily exportable to other available software or platforms. The respective functions are available at the web page of the GRASS group at `http://www-eio06.upc.es/grass`.

## Acknowledgements

This work has been partially supported by grant MTM2005–08886 from the Ministerio de Ciencia y Tecnología. Authors are grateful to the GRASS group for fruitful discussions and interesting suggestions, and to the Fundació Lluita contra la SIDA for providing the data set. We thank the reviewers and the editor for constructive suggestions in the previous version of this manuscript.

## Appendix: Visser's likelihood

To simplify the expressions we introduce the following complementary notation. For a given individual a different way of representing the observables is using $(Y_1, Y_2, \delta)$, where

$$\delta = \begin{cases} 1 & \text{if} \quad T_1 > C \qquad (i.e. \quad D_1 = 0 = D_2) \\ 2 & \text{if} \quad T_1 \leq C < T_1 + T_2 \quad (i.e. \quad D_1 = 1, D_2 = 0) \\ 3 & \text{if} \quad T_1 + T_2 \leq C \qquad (i.e. \quad D_1 = 1, D_2 = 1). \end{cases}$$

The likelihood for the $n$ observations is as follows:

$$\begin{aligned} L \quad = \quad & \prod_{i=1}^{n} \Big\{ \Pr\{T_1 > y_{1i}, C = y_{1i}\}^{1\{\delta_i = 1\}} \\ & \Pr\{T_1 = y_{1i}, T_2 > y_{2i}, C - y_{1i} = y_{2i}\}^{1\{\delta_i = 2\}} \\ & \Pr\{T_1 = y_{1i}, T_2 = y_{2i}, C > y_{1i} + y_{2i}\}^{1\{\delta_i = 3\}} \Big\} \end{aligned}$$

and the corresponding log-likelihood, $\mathcal{L} = \log L$, since variables are discrete, the possible values for $y_{1i}, y_{2i}$ and $C$ are only $\{0, 1, 2, \ldots, K\}$ and $C$ is independent of $(T_1, T_2)$, looks like as

$$\begin{aligned} \mathcal{L} \quad = \quad & \sum_{k=1}^{K} \{n_{k\cdot} \log \Pr\{T_1 = k\} + n_{1k} \log \Pr\{T_1 > k\}\} + \\ & \sum_{k=1}^{K} \sum_{l=1}^{K} \{n_{3kl} \log \Pr\{T_2 = l | T_1 = k\} + n_{2kl} \log \Pr\{T_2 > l | T_1 = k\}\} + \end{aligned}$$

$$\sum_{k=1}^{K}\{n_{1k}\log\Pr\{C=k\}+\sum_{l=1}^{K}\{n_{2kl}\log\Pr\{C=k+l\}+n_{3kl}\log\Pr\{C>k+l\}\}\}$$
$$= \mathcal{L}_{T_1}+\mathcal{L}_{T_2|T_1}+\mathcal{L}_C.$$

All the expressions for the probabilities can be replaced by functions containing uniquely $\lambda_{T_1}(k)$ and to $\lambda_{T_2|T_1=k}(l)$. For instance,

$$\begin{aligned}\mathcal{L}_{T_1} &= \sum_{k=1}^{K}\{n_{k\cdot}(\log\lambda_{T_1}(k)+\log\prod_{j=0}^{k-1}(1-\lambda_{T_1}(j)))+n_{1k}\log\prod_{j=0}^{k}(1-\lambda_{T_1}(j))\}\\ &= \sum_{k=1}^{K}n_{k\cdot}\log\lambda_{T_1}(k)+\sum_{k=1}^{K}(n_{1k}+n_{k\cdot})\sum_{j=0}^{K-1}\log(1-\lambda_{T_1}(j))+\sum_{k=1}^{K}n_{1k}\log(1-\lambda_{T_1}(k)).\end{aligned}$$

The nonparametric estimators for the hazard functions are obtained after maximizing the log-likelihood $\mathcal{L}=\mathcal{L}_{T_1}+\mathcal{L}_{T_2|T_1}+\mathcal{L}_C$. Note that we are in fact maximizing $\log L$ with respect to $\lambda_{T_1}(k)$ and to $\lambda_{T_2|T_1=k}(l)$, and because the terms act additively we can maximize first with respect to $\lambda_{T_1}(k)$ and then with respect to $\lambda_{T_2|T_1=k}(l)$.

## References


Burke, M. D. (1988). Estimation of a bivariate distribution function under random censorship. *Biometrika*, 75, 379-382.

Campbell, G. (1981). Nonparametric bivariate estimation with randomly censored data. *Biometrika*, 68, 417-422.

Campbell, G. and Földes, A. (1982). Large-sample properties of nonparametric bivariate estimators with censored data. *Colloquia Mathematica Societatis Janos Bolyai*, 32, 103-121.

Dabrowska, D. M. (1988). Kaplan-Meier estimate on the plane. *The Annals of Statistics*, 16, 1475-1489.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Boca Raton: Chapman & Hall/CRC.

Gómez, G., Calle, M. L., Serrat, C. and Espinal, A. (2004). Review of Multivariate Survival Data. Technical Report DR2004/15. Departament d'Estadística i Investigació Operativa. Universitat Politècnica de Catalunya.

Gómez, G., Serrat, C. and Ruiz, L. (2007). Weighted conditional survival estimator for ordered failure times subject to a common censoring process. *Submitted.*

Ruiz, L., Paredes, R., Gómez, G. *et al*. (2007). Antiretroviral Therapy Interruption Guided by CD4 T cell Counts and Plasma HIV-1 RNA Levels in chronically HIV-1 Infected Patients. *AIDS*, 21, 169-178.

Tsai,W-Y., Leurgans, S. and Crowley, J. (1986). Nonparametric estimation of a bivariate survival function in the presence of censoring. *The Annals of Statistics*, 14, 1351-1365.

Visser, M. (1996). Nonparametric estimation of the bivariate survival function with an application to vertically transmitted AIDS. *Biometrika*, 83, 507-518.

Wang, M.-C. and Wells, M. T. (1998). Nonparametric Estimation of successive duration times under dependent censoring. *Biometrika*, 85, 561-572.

# Corrections and errata

SORT 31 (1) January-June 2007, 99-100

# Correction on "The importance of being the upper bound in the bivariate family"

Carles M. Cuadras[1] and Ali Dolati[2]

[1] *University of Barcelona and* [2] *Shimz University*

## 1 Introduction

Let us perform two corrections on Cuadras (2006) and add a short note. First, the cumulative distribution function of the Pareto distribution should be $F(x) = 1 - (x/a)^{-c}$ if $x > a$, instead of $1 - (x/a)^{c}$. This misprint does not alter the formulas for the Lorenz curve, Gini coefficient and the role of the upper bound for evaluating the social inequality.

Second, a general expression for a measure of stochastic dependence between two random variables with cdf $H$ was proposed in the same paper. For uniform marginals this measure reduces to

$$A(U, V) = c \int_{\mathbf{I}^2} (C_H(u, v) - uv)\, d\mu,$$

where $c$ is a normalizing constant, $\mu$ is a suitable measure and $C_H$ is the copula related to $H$. Then it is said that Kendall's $\tau$ and Spearman's $\rho_s$ are coefficients of dependence computed from the copula $C_H$ by using $d\mu = dC_H$ and $d\mu = dudv$, respectively. However, the first statement is not true. Thus, while the second equation in

$$\begin{aligned} \tau &= 4 \int_{\mathbf{I}^2} (C_H(u, v) - uv) dC_H(u, v) \\ &= 4 \int_{\mathbf{I}^2} C_H(u, v) dC_H(u, v) - 1, \end{aligned}$$

is correct, the first one is incorrect.

As it was proved by Behboodian *et al.* (2005), if we impose the condition $A(U,V) = 1$ for the upper bound $C^+ = \min\{u,v\}$, then $c = 6$ and $A(U,V)$ is maximum. Therefore from

$$\begin{aligned} \rho_s &= 12 \int_{\mathbf{I}^2} C_H(u,v)dudv - 3, \\ 3\tau &= 12 \int_{\mathbf{I}^2} C_H(u,v)dC_H(u,v) - 3, \end{aligned}$$

and

$$\int_{\mathbf{I}^2} C_H(u,v)dudv = \int_{\mathbf{I}^2} uvdC_H(u,v),$$

where $C_H$ is any copula, this measure can be expressed as

$$\begin{aligned} A(U,V) &= 6 \int_{\mathbf{I}^2} (C_H(u,v) - uv)dC_H(u,v) \\ &= \frac{3\tau - \rho_s}{2}. \end{aligned}$$

Thus the above proposed measure can not give $\tau$. Actually $A(U,V)$ combines both coefficients Kendall's $\tau$ and Spearman's $\rho_s$ and is an example of average quadrant dependence measure.

Finally note that $A(U,V)$ is not a measure of concordance in the sense of Scarsini (1984). Specifically, it does not preserve the concordance ordering property, that is, if $C_1, C_2$ are two copulas such that $C_1 \leq C_2$ then not necessarily $A_1 \leq A_2$. See Section 5 and Example 5.1 in Behboodian *et al.* (2005) for further details. In fact, $A(U,V)$ is just a measure of association.

## References

Behboodian, J., Dolati, A. and Úbeda-Flores, M. (2005). Measures of association based on average quadrant dependence. *Journal of Probability and Statistical Science*, 3, 161-173.

Cuadras, C. M. (2006). The importance of being the upper bound in the bivariate family. *SORT*, 30, 55-84.

Scarsini, M. (1984). On measures of concordance. *Stochastica*, 8, 201-218.

## ERRATA

- In volume 30, number 2, page 173 the first formula should not appear in the text. The correct paragraph would be:

The exponentiated-Weibull distribution considered in Mudhokar et al. (1995) with parameters $\alpha$, $\theta$ and $\sigma$ considers that life time $T$ has a density function given by

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

where $\alpha > 0$, $\theta > 0$ are shape parameters and $\sigma > 0$ is a scale parameter.

- In volume 30, number 2, page 205, another author should be included: Joan Fiblà, Universitat de Lleida

We apologise for these errors

# Book reviews

***ANALYSIS OF INTEGRATED AND COINTEGRATED TIME SERIES WITH R***

Bernhard Pfaff

Springer Science+Business Media, Inc. New York, 2006, 139 pages

Integration and cointegration of time series have become an essential part of the modern Econometry. This book presents a good overview on those topics. As the author points out: "although the book's content is not a pure theoretical exposition of integration analysis, it is particularly suited as an accompanying text in applied computer laboratory classes". It combines theoretical concepts with applied examples coded in R, allowing the reader not only to grasp the concepts but also to apply them to real cases. At the end of each chapter there is a Summary, emphasizing the most important ideas exposed in the chapter. Additionally, a set of exercises is proposed with the idea that the readers apply, by means of R contributed packages, the notions covered in the chapter. All tests and models used in this book are included in the "*urca*" package. Bernhad Pfaff is the author and maintainer of the CRAN contributed packages "*urca*" and "*vars*".

This book is appropriate for graduate students and professionals in applied econometrics and also is suitable as a support for the computer sessions lab.

The book is divided into 9 chapters and organized into three parts:

- Theoretical Concepts
- Unit Root Tests
- Cointegration

Chapters 1-3 comprise the first part. This first part is a review of the basic ideas of time series models, unit roots and cointegration. Chapter 1 gives the background to ARMA models, introducing the notation that will be used around the book. Chapter 2 reviews the key concepts of nonstationary time series models such that integrated, seasonally integrated and fractionally integrated univariate time series models. Finally, Chapter 3 is focused on multivariate relationships between time series. Spurious regression is presented before discussing the concept of Cointegration and Error-Correction models, finalizing this Chapter with vector correction models (VECM). At this point the author discusses the decomposition of a time series models into deterministic and stochastic component and the extension of this decomposition to the multivariate case, before

presenting VECM models. The concepts presented in Chapters 2 and 3 will be discussed in more depth in parts 2 and 3 of the book.

Part 2 is dedicated to Unit Root Test and includes Chapter 4 and 5. Chapter 4 introduces the most used test in econometrics, the Dickey-Fuller test and its extensions (ADF test). Next, the author proposes a sequential test strategy with the aim of deciding whether to fit a trend to the data or to difference the process in order to work with an integrated process. Next, the following unit root tests used in the time series area, are introduced: Philips-Perron Test (P-P Test), ERS Test, Schmidt-Phillips (SP)-Test and KPSS-Test, emphasizing the form of the null hypothesis for each test. With the purpose of showing the pros and cons of those tests, the author applies those tests to two time series: Consumption in United Kingdom and the nominal GNP in the United States. This practical way of introducing those tests makes clear the advantages and disadvantages of them. Structural breaks in time series and the Zivot-Andrews test to detect them are considered in Chapter 5, showing how a structural break can modify the order of integration of a process. In this chapter seasonal unit roots are also introduced.

Finally, part 3 concentrates on the cointegration methodology, starting with the case of single equation models in Chapter 6, and ending up with more sophisticated vector error correction model (VECM) in Chapter 7. This part will be particularly helpful to beginners to better understand the difficult concept of cointegration. Chapter 6 introduces the reader to this topic with the intuitive Engel-Granger Two Step Procedure and Chapter 7 deals with the VECM model, particularly useful for a better comprehension of the nature of nonstationarity among several time series. The detection of cointegration rank is tested following the Johansen and Juselius approach. The data used in this chapter is the same data used in Johansen and Juselius paper (Oxford Bulletin of Economics and Statistics, 52, 2, 1990), making it easy to follow the results published in the mentioned paper. Finally, this Chapter presents the detection of structural breaks in a VAR model and how they affect VECM models.

The book concludes with a very useful appendix about time series data, tools and contributed packages stored in CRAN.

Hence, I recommend this book as a companion text in lab sessions to better understand the concepts of Integrated and Cointegrated Time Series.

M. Pilar Muñoz
Department of Statistics and Operation Research
Universitat Politècnica de Catalunya

# 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 $2_\varepsilon$.

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

### Publishing rights and authors' opinions



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 2007 (volume 31)

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

**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 2005 (volume 29)

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

**Bank copy**

Authorisation for automatic bank withdrawal in payment for

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

The undersigned _____

authorises Bank/Financial institution _____

located at (Street Address) _____

Zip/postal code _____ City _____

Country _____

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

number ☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

Date _____

Signature

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

# 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