Bayesian Statistics and its Applications
(S. K. Upadhyay, U. Singh and D. K. Dey, (eds.)
New Delhi: Anamaya Pub., 2006, 110-121.

Intrinsic Point Estimation of the Normal Variance

JOSE M. BERNARDO
Universitat de Valencia, Spain
<jose.m.bernardo@uv.es>

SUMMARY

Point estimation of the normal variance is surely one of the oldest non-trivial problems in mathematical statistics and yet, there is certainly no consensus about its more appropriate solution. Formally, point estimation may be seen as a decision problem where the action space is the set of possible values of the quantity on interest; foundations then dictate that the solution must depend on both the utility function and the prior distribution. An estimator 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 function, the *reference prior*, produces a general Bayesian objective solution to the problem of point estimation. In this paper, point estimation of the normal variance is considered in detail, and the behaviour of the solution found is compared with the behaviour of alternative conventional solutions from both a Bayesian and a frequentist perspective.

Keywords: DECISION THEORY; INTRINSIC DISCREPANCY; INTRINSIC LOSS; NONINFORMATIVE PRIOR; REFERENCE ANALYSIS; POINT ESTIMATION.

1. INTRODUCTION

Point estimation of the normal variance has a long, fascinating history which is far from settled. As mentioned by Maata and Casella (1990) in their lucid discussion of the frequentist decision-theoretic approach to this problem, the list of contributors to the twin problems of point estimation of the normal mean and point estimation of the normal variance reads like a *Who's Who* in modern 20th century statistics.

In this paper, an objective Bayesian decision-theoretic solution to point estimation of the normal variance when the mean is unknown is presented. In marked contrast with most approaches, this solution is invariant under one-to-one parametrization. The behaviour of this new solution is compared to the behaviour of known alternatives from both a Bayesian and a frequentist viewpoint.

José M. Bernardo is Professor of Statistics at the University of Valencia. Research partially funded with grant MTM2004-05956 of the MEyC, Madrid, Spain.

1.1. *Notation*

A brief review of notation is needed to proceed. 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 $f_{\boldsymbol{x}}(\cdot)$ and $g_{\boldsymbol{x}}(\cdot)$ of $\boldsymbol{x} \in \mathcal{X}$, respectively by $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ will often be used. Thus, the conditional probability density of observable data $\boldsymbol{x} \in \mathcal{X}$ given $\boldsymbol{\theta}$ will be represented by by either $p_{\boldsymbol{x}}(\cdot \,|\, \boldsymbol{\theta})$ or $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, with $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \geq 0$, $\boldsymbol{x} \in \mathcal{X}$, and $\int_{\mathcal{X}} p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x} = 1$, and the posterior density of a non-observable parameter vector $\boldsymbol{\theta} \in \Theta$ given $\boldsymbol{x}$ will be represented by either $\pi_{\boldsymbol{\theta}}(\cdot \,|\, \boldsymbol{x})$ or $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$, with $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \geq 0$, $\boldsymbol{\theta} \in \Theta$, and $\int_{\Theta} \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta} = 1$. Density functions of specific distributions are denoted by appropriate names. In particular, if x has a normal distribution with mean μ and variance σ^2, its probability density function will be denoted $\mathrm{N}(x \,|\, \mu,\, \sigma)$. The maximum likelihood estimators of μ and σ^2 given a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from $\mathrm{N}(x \,|\, \mu,\, \sigma)$ will respectively be denoted by

$$\hat{\mu} = \overline{x} = \frac{1}{n} \sum\nolimits_{j=1}^{n} x_j \;, \qquad \hat{\sigma}^2 = s^2 = \frac{1}{n} \sum\nolimits_{j=1}^{n} (x_j - \overline{x})^2.$$

The more common point estimators of the normal variance are members of the family of *affine invariant* estimators

$$\tilde{\sigma}^2_{\nu} = \frac{n s^2}{\nu} = \frac{1}{\nu} \sum\nolimits_{j=1}^{n} (x_j - \overline{x})^2, \quad \nu > 0; \tag{1}$$

In particular, the maximum likelihood estimator (MLE) is $s^2 = \tilde{\sigma}^2_n$, and the unique unbiased estimator is $\tilde{\sigma}^2_{n-1}$.

1.2. *Point Estimation*

The basic facts on frequentist point estimation of the normal variance are well known. The MLE of σ^2 is s^2 and, since maximum likelihood estimation is invariant under one-to-one transformations, the MLE of σ is s. The uniformly minimum variance unbiased estimator (UMVUE) of σ^2 is $\tilde{\sigma}^2_{n-1} = n s^2/(n-1)$, but the UMVUE of σ is not its squared root, but $s\, \sqrt{(n/2)}\, \Gamma((n-1)/2)/\Gamma(n/2)$ (see *e.g.*, Lehmann and Casella, 1998, p. 91); for $n = 2$ these respectively yield $\sqrt{2}\, s$ and $\sqrt{\pi}\, s$, with $s = |x_1 - x_2|/2$, a 25% difference using precisely the same procedure; this a good example of methodological inconsistency.

Despite many warnings on its inappropriate behaviour ("I find it hard to take the problem of estimating σ^2 with quadratic loss very seriously" Stein, 1964; see also Brown, 1968, 1990), decision theoretical approaches to the normal variance estimation are typically based on the standardized quadratic loss function

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

where overestimation of σ^2 is much more severely penalized than underestimation, thus leading to presumably too small estimates. Indeed, the best invariant estimator (minimum risk equivariant estimator, MRE) of σ^2 under this loss, which is also minimax, is $\tilde{\sigma}^2_{n+1}$ (see *e.g.*, , Lehmann and Casella, 1998, p. 172), smaller than both the MLE and the unbiased estimator, and it is often considered the "straw man" to beat in this problem (George, 1990). By considering a

larger class of estimators than (1), namely those of the form $\phi_n(z)\, ns^2$ where ϕ_n is a real valued function and $z = \overline{x}/s$, which are also are scale invariant, Stein (1964) found that

$$\tilde{\sigma}^2_{stein} = \min\left\{\tilde{\sigma}^2_{n+1},\, \tilde{\sigma}^2_{(n+2)/(1+z^2)}\right\}, \tag{3}$$

dominates $\tilde{\sigma}^2_{n+1}$ under the standardized quadratic loss (2) and, thus, $\tilde{\sigma}^2_{n+1}$ is inadmissible under that loss. The intuition behind this is that small z values indicate that μ may be close to 0, and then $\tilde{\sigma}^2_{n+2}$, which would the best affine invariant estimator under the quadratic loss (2) if μ were known to be zero, might be better than $\tilde{\sigma}^2_{n+1}$. This prompted a whole class of so-called *preliminary test* estimators where the estimator takes one of typically two different forms, depending of the value of z (Brown, 1968; Brewster and Zidek, 1974). For a review of their performance, see Csörgö and Faraway (1996). Notice however that $\tilde{\sigma}^2_{stein}$ must also be inadmissible, for admissible estimators are limits of Bayes estimators, and so must be analytic.

The results mentioned above are all obtained under the mathematically convenient—but otherwise rather unsatisfactory—standardized quadratic loss. James and Stein (1961) suggested the use of the far more appropriate *entropy loss*

$$\ell_{entropy}\{\tilde{\sigma}^2, \sigma^2\} = \int_{\Re} \mathrm{N}(x \,|\, \mu, \sigma) \log \frac{\mathrm{N}(x \,|\, \mu, \sigma)}{\mathrm{N}(x \,|\, \mu, \tilde{\sigma})}\, \mathrm{d}x = \frac{1}{2}\left[\frac{\tilde{\sigma}^2}{\sigma^2} - 1 - \log \frac{\tilde{\sigma}^2}{\sigma^2}\right] \tag{4}$$

and showed that the best invariant estimator for this loss is $\tilde{\sigma}^2_{n-1}$ (the unbiased estimator). Brown (1968) proved that, from a frequentist decision theoretic viewpoint, this may be also improved by appropriate preliminary test estimators. In particular,

$$\tilde{\sigma}^2_{brown} = \min\left\{\tilde{\sigma}^2_{n-1},\, \tilde{\sigma}^2_{n/(1+z^2)}\right\}, \tag{5}$$

dominates $\tilde{\sigma}^2_{n-1}$ under Stein loss (4), and thus $\tilde{\sigma}^2_{n-1}$ is inadmissible under that loss. On the other hand, Lin and Pal (2005) recently found that, from a frequentist viewpoint, $\tilde{\sigma}^2_{n-1}$ may be argued to be a good compromise estimator, for it performs moderately well under several alternative criteria (risk, Pitman nearness and stochastic domination), while the performance of the other estimators they consider dramatically depends of the criterion used.

It may certainly be argued that the frequentist emphasis of the concept described by the emotionally charged word "inadmissible" may well be inappropriate. Indeed, for some assumed loss function, a particular estimator is deemed as *inadmissible* (do not dare to use it!) only because its *average* loss under repeated sampling is larger than that of another estimator; whether or not the "inadmissible" estimator is actually better for most regions of the sampling space (which are often identifiable) is simply ignored. Yet, one would certainly expect that decent people would prefer a country where there is no poverty to one with a larger *average* income induced a small group of very rich people which over compensates the small income of the poor people, thus producing a larger average. In more technical terms, one should certainly analyse the sampling properties of any statistical procedure, and avoid procedures which for some possible parameter values would give misleading conclusions most of the time (the weak repeated sampling principle, Cox and Hinkley, 1974, p. 45); however, the crucial properties of a statistical procedure are those *conditional* on the observed data, not *average* over them: a sensible procedure should produce appropriate answers whatever the data obtained, with special attention to the parameter values supported by the observed data. But, of course, this type of analysis requires a Bayesian approach.

Conventional Bayesian point estimation typically consists of some location measure of the marginal posterior distribution of the quantity of interest. The solution naturally depends on

the prior used. Objective Bayesian estimators, which do not involve any information about the parameters beyond that contained in the assumed model—and may therefore be meaningfully compared with their frequentist counterparts—require an objective prior, that is a positive prior function to be formally used in Bayes theorem, which only depends on the assumed model and on the quantity of interest. In the particular problem where data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ consist of a random sample from a normal $\mathrm{N}(x \,|\, \mu, \sigma)$ distribution and the quantity of interest is either σ, or any one-to-one function of σ (say the variance, σ^2, the precision $1/\sigma^2$, or the approximate location parameter $\log \sigma$), there is a clear consensus on the objective prior function to use, namely

$$\pi(\mu, \sigma) = \sigma^{-1} \tag{6}$$

i.e., a uniform prior on both μ and $\log \sigma$. Indeed, this was already suggested by Barnard (1952) on invariance arguments, and recommended by Jeffreys (1939, p. 138), Lindley (1965, p. 37) and Box and Tiao (1973, p. 49) in their pioneering books. As one would expect, this is also the relevant reference prior (Bernardo, 1979).

The corresponding marginal posterior of the standard deviation, the reference posterior distribution of σ is

$$\pi(\sigma \,|\, \boldsymbol{x}) = \propto \int_{-\infty}^{-\infty} \prod_{j=1}^{n} \mathrm{N}(x_j \,|\, \mu, \sigma)\, \pi(\mu, \sigma)\, \mathrm{d}\mu \propto \sigma^{-n} e^{-\frac{n}{2}\frac{s^2}{\sigma^2}} \tag{7}$$

and, hence, the reference posterior of the variance σ^2 is

$$\pi(\sigma^2 \,|\, \boldsymbol{x}) \propto (\sigma^2)^{-(n+1)/2} e^{-\frac{n}{2}\frac{s^2}{\sigma^2}}, \tag{8}$$

an inverted gama $\mathrm{Ig}(\sigma^2 \,|\, (n-1)/2, ns^2/2)$, and the reference posterior of $\tau = ns^2/\sigma^2$ is $\pi(\tau \,|\, n) = \chi^2_{n-1}$, a central chi-square with $n-1$ degrees of freedom. Naïve Bayesian estimators of σ and σ^2 are given by the corresponding posterior means

$$\mathrm{E}[\sigma \,|\, \boldsymbol{x}] = \sqrt{\frac{n}{2}} \frac{\Gamma[(n-2)/2]}{\Gamma[(n-1)/2]}\, s\,, \qquad \mathrm{E}[\sigma^2 \,|\, \boldsymbol{x}] = \frac{n\, s^2}{n-3} = \tilde{\sigma}^2_{n-3}, \tag{9}$$

and posterior modes,

$$\mathrm{Mo}[\sigma \,|\, \boldsymbol{x}] = s\,, \qquad \mathrm{Mo}[\sigma^2 \,|\, \boldsymbol{x}] = \frac{n\, s^2}{n+1} = \tilde{\sigma}^2_{n+1}. \tag{10}$$

Note that these estimation procedures are *not* consistent under reparametrization. It may be appreciated that there are many direct relations between frequentist and naïve objective Bayesian estimators. For instance, the mode of the reference posterior of σ is s, its MLE, and the mode of the reference posterior of σ^2 is $\tilde{\sigma}^2_{n+1}$, its MRE.

A more formal Bayesian approach to point estimation is to consider this problem as a decision problem where the action space is the set of possible values of the quantity of interest. This requires to specify a loss function. Naïve loss functions reproduce naïve estimators. Thus, the optimal Bayes estimator under quadratic loss is the posterior expectation, leading to the results in (9), and the optimal Bayes estimator under a zero-one loss is the posterior mode, leading to the results in (10). Bayesian decision-theoretic point estimation is considered in detail in Section 2.1.

In this paper, a particular objective Bayesian solution to point estimation of any one-to-one function of the normal variance is presented. Section 2 is a review of the methodology used,

intrinsic estimation (Bernardo, 1999; Bernardo and Rueda, 2002; Bernardo and Juárez, 2003; Bernardo, 2005), an objective Bayesian decision-theoretic approach which uses an information-theory based loss function and a reference prior. Section 3 contains the derivation on the intrinsic point estimator of the normal variance with unknown mean and, by invariance of the argument used, that of any one-to-one transformation of the variance. In Section 4, the results obtained are compared with other solutions in the literature, from both a Bayesian and a frequentist viewpoint.

2. INTRINSIC ESTIMATION

2.1. *Point Estimation as a Decision Problem*

Let $\boldsymbol{x}$ be the available data, which are assumed to consist of one observation from model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{x} \in \mathcal{X}, \boldsymbol{\omega} \in \Omega\}$, and let $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \Theta$ be the vector of interest. Often, but not necessarily, the data $\boldsymbol{x}$ consist of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of some simpler model $\{q(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{x} \in \mathcal{X}, \boldsymbol{\omega} \in \Omega\}$, in which case $p(\boldsymbol{x} \,|\, \boldsymbol{\omega}) = \prod_{j=1}^{n} q(x_j \,|\, \boldsymbol{\omega})$. Without loss of generality, the original model $\mathcal{M}$ may be written as $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta, \boldsymbol{\lambda} \in \Lambda\}$, 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 $\tilde{\boldsymbol{\theta}}(\boldsymbol{x}) \in \Theta$ 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 $\boldsymbol{\theta}$.

For each given data set $\boldsymbol{x}$, to choose a point estimate $\tilde{\boldsymbol{\theta}}$ is a *decision problem*, where the action space is the class Θ of possible $\boldsymbol{\theta}$ values. 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* $\ell\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ measuring the consequences of acting *as if* the true value of the quantity of interest were $\tilde{\boldsymbol{\theta}}$, when the actual parameter values are $(\boldsymbol{\theta}, \boldsymbol{\lambda})$. Given data $\boldsymbol{x}$, the loss to be expected if $\tilde{\boldsymbol{\theta}}$ were used as the true value of the quantity of interest is

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

where $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})\, \pi(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is the joint posterior density of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$, and $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is the joint prior of the unknown parameters. Given data $\boldsymbol{x}$, the *Bayes estimate* is that $\tilde{\boldsymbol{\theta}}$ value which minimizes in Θ the (posterior) expected loos $l\{\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}\}$. The *Bayes estimator* is the function

$$\boldsymbol{\theta}^*(\boldsymbol{x}) = \arg \min_{\tilde{\boldsymbol{\theta}} \in \Theta} l\{\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}\}. \tag{11}$$

For any given model, the Bayes estimator depends on both the loss function $\ell\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ and the prior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda})$. In the case of the normal variance considered in this paper, data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ are assumed to be a random sample from $\mathrm{N}(x \,|\, \mu, \sigma)$, the parameter of interest is either σ or some one-to-one function of σ, and μ is a nuisance parameter. It has already been mentioned that the undisputed objective prior for this problem is $\pi(\mu, \sigma) = \sigma^{-1}$, which leads to the reference posterior distributions (7) and (8).

2.2. *The Loss Function*

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. The more common of these conventional loss functions (which often ignore the possible 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 is then the (marginal) *posterior mean* $\boldsymbol{\theta}^* = E[\boldsymbol{\theta} \,|\, \boldsymbol{x}]$, assuming that the mean exists (see, *e.g.*, Bernardo and Smith, 1994, p. 257).

In the case of the normal variance, the conventional quadratic loss $\ell(\tilde{\sigma}^2, \sigma^2) = c\,(\tilde{\sigma}^2 - \sigma^2)^2$ leads to the reference posterior expectation $\mathrm{E}[\sigma^2 \,|\, \boldsymbol{x}] = \tilde{\sigma}^2_{n-3}$ quoted in (9). If the slightly more sophisticated standardized quadratic loss (2) is used, the Bayes estimator is

$$\arg\min_{\tilde{\sigma}^2>0} \int_0^\infty [(\tilde{\sigma}^2/\sigma^2) - 1]^2 \, \pi(\sigma^2 \,|\, \boldsymbol{x}) \, \mathrm{d}\sigma^2 = \frac{n\,s^2}{n+1} = \tilde{\sigma}^2_{n+1}, \tag{12}$$

which is also the MRE of σ^2 under this loss. Using the entropy loss (4) yields

$$\arg\min_{\tilde{\sigma}^2>0} \int_0^\infty [(\tilde{\sigma}^2/\sigma^2) - 1 - \log(\tilde{\sigma}^2/\sigma^2)] \, \pi(\sigma^2 \,|\, \boldsymbol{x}) \, \mathrm{d}\sigma^2 = \frac{n\,s^2}{n-1} = \tilde{\sigma}^2_{n-1}, \tag{13}$$

which is also both the MRE of σ^2 under this loss, and the unbiased estimator.

Conventional loss functions are typically *not* invariant under reparametrization. As a consequence, the Bayes estimator $\boldsymbol{\psi}^*$ of a one-to-one transformation $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta})$ of the original parameter $\boldsymbol{\theta}$ is not necessarily $\boldsymbol{\psi}(\boldsymbol{\theta}^*)$. Yet, scientific applications require this type of invariance; indeed, it would be hard to argue that the best estimate of, say a galaxy speed, is θ^* but that the best estimate of the logarithm of that speed 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 $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \ell\{p_{\boldsymbol{x}}(. \,|\, \tilde{\boldsymbol{\theta}}),\, p_{\boldsymbol{x}}(. \,|\, \boldsymbol{\theta})\}$ 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* (Bernardo and Rueda, 2002) is an intrinsic loss with very attractive properties. The intrinsic discrepancy between two models $p_{\boldsymbol{x}}(\cdot \,|\, \boldsymbol{\theta}_1)$ and $p_{\boldsymbol{x}}(\cdot \,|\, \boldsymbol{\theta}_2)$ for data $\boldsymbol{x} \in \mathcal{X}$ is

$$\begin{aligned} \delta(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \delta\{p_{\boldsymbol{x}}(. \,|\, \boldsymbol{\theta}_1),\, p_{\boldsymbol{x}}(. \,|\, \boldsymbol{\theta}_2)\} = \min\{\kappa(\boldsymbol{\theta}_1 \,|\, \boldsymbol{\theta}_2),\, \kappa(\boldsymbol{\theta}_2 \,|\, \boldsymbol{\theta}_1)\}, \\ \kappa(\boldsymbol{\theta}_i \,|\, \boldsymbol{\theta}_j) &= \int_{\mathcal{X}} p_{\boldsymbol{x}}(\boldsymbol{x} \,|\, \boldsymbol{\theta}_j) \log \frac{p_{\boldsymbol{x}}(\boldsymbol{x} \,|\, \boldsymbol{\theta}_j)}{p_{\boldsymbol{x}}(\boldsymbol{x} \,|\, \boldsymbol{\theta}_i)} \, \mathrm{d}\boldsymbol{x}, \end{aligned} \tag{14}$$

that is, the minimum Kullback-Leibler logarithmic divergence between them. This is a proper discrepancy measure; indeed, (i) it is symmetric, (ii) it is non-negative and (iii) it is zero if, and only if, $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1) = p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_2)$ almost everywhere. The intrinsic discrepancy, is invariant under one-to-one transformations of either the parameter vector $\boldsymbol{\theta}$ of the data set $\boldsymbol{x}$. Moreover, the intrinsic discrepancy is well defined in irregular models, where the sample space may depend on the parameter value and, thus, the support of, say, $p_{\boldsymbol{x}}(. \,|\, \boldsymbol{\theta}_1)$ may be strictly smaller that that of $p_{\boldsymbol{x}}(. \,|\, \boldsymbol{\theta}_2)$.

In the context of point estimation, this leads to the (invariant) intrinsic loss function

$$\delta\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \min_{\tilde{\boldsymbol{\lambda}} \in \Lambda} \delta\{(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}), (\boldsymbol{\theta}, \boldsymbol{\lambda})\} \tag{15}$$

which measures the discrepancy between the *model* $p_{\boldsymbol{x}}(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ and its closest approximation within the family $\{p_{\boldsymbol{x}}(\cdot \,|\, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}),\, \tilde{\boldsymbol{\lambda}} \in \Lambda\}$ of all models with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$. Notice that the value of $\delta\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ does *not* depend on the particular parametrization chosen to describe the problem.

Given data $\boldsymbol{x}$ generated by $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})$ and no subjective prior information, the (reference posterior) expected intrinsic loss from using $\tilde{\boldsymbol{\theta}}$ as a proxy for $\boldsymbol{\theta}$ is the *intrinsic statistic function*

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

where $\delta\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\omega})\}$ is the intrinsic discrepancy of the true model from the family of models with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$, and $\pi(\boldsymbol{\theta}, \boldsymbol{\omega} \,|\, \boldsymbol{x})$ is the joint posterior distribution which results form formal use of Bayes theorem with the reference prior $\pi(\boldsymbol{\theta})\, \pi(\boldsymbol{\omega} \,|\, \boldsymbol{\theta})$ associated to model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})$ when $\boldsymbol{\theta}$ is the quantity of interest.

It immediately follows from (14), (15) and (16) that $d\{\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}\}$ is simply the posterior expectation of the minimum expected likelihood ratio between the true model and a model with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$. This is an explicit measure of the (expected posterior) loss associated to any particular estimate $\tilde{\boldsymbol{\theta}}$.

The *intrinsic estimate* is the value $\boldsymbol{\theta}^*$ which minimizes $d\{\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}\}$, that is, the Bayes estimate with respect to the intrinsic discrepancy loss and the reference posterior. Formally, the intrinsic estimator is then

$$\boldsymbol{\theta}^*(\boldsymbol{x}) = \arg\min_{\tilde{\boldsymbol{\theta}} \in \Theta} d\{\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}\}, \tag{17}$$

where $d\{\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}\}$ is given by (16). Since both the intrinsic loss function and the reference prior are invariant under one-to-one reparametrization, the intrinsic estimator $\boldsymbol{\phi}^*(\boldsymbol{x})$ of any one-to-one function $\boldsymbol{\phi}\{\boldsymbol{\theta}\}$ of $\boldsymbol{\theta}$ will simply be $\boldsymbol{\phi}^*(\boldsymbol{x}) = \boldsymbol{\phi}\{\boldsymbol{\theta}^*(\boldsymbol{x})\}$.

3. INTRINSIC POINT ESTIMATION OF NORMAL VARIANCE

In this section, the intrinsic point estimator of a normal variance (and, by the invariance of the argument used, the point estimators of any one-to-one function of the variance) are derived.

The intrinsic discrepancy $\delta\{p_1, p_2\}$ between two normal densities $p_1(x)$ and $p_2(x)$, with $p_i(x) = \mathrm{N}(x \,|\, \mu_i, \sigma_i)$, is

$$\begin{aligned} \delta\{p_1, p_2\} &= \min\{\kappa\{p_1 \,|\, p_2\},\ \kappa\{p_2 \,|\, p_1\}\}, \\ \kappa\{p_i \,|\, p_j\} &= \int_{-\infty}^{\infty} p_j(x) \log \frac{p_j(x)}{p_i(x)}\, \mathrm{d}x = \frac{1}{2}\Big\{ \log \frac{\sigma_i^2}{\sigma_j^2} + \frac{\sigma_j^2}{\sigma_i^2} - 1 + \frac{(\mu_i - \mu_j)^2}{\sigma_i^2} \Big\}. \end{aligned} \tag{18}$$

The behaviour of the intrinsic discrepancy is *very* different from the conventional quadratic distance, $\ell_{quad}\{p_1,\, p_2\} = c\{(\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2\}$. In Figure 1, the discrepancy between a normal distribution $\mathrm{N}(x \,|\, \mu, \sigma)$ and a standard normal $\mathrm{N}(x \,|\, 0, 1)$ is represented for both the intrinsic discrepancy and the quadratic distance as a function of μ and $\log \sigma$; a useful range of parameter values, $\mu \in [-3, 3]$ and $\sigma \in [e^{-3}, e^3] \approx [0.05, 20.1]$ has been used and, to facilitate comparison, the constant c in the quadratic distance has been chosen such that both surfaces have the same value at the extreme point $(3, 3)$. It is apparent from Figure 1 that, as one would expect from its analytical form, the quadratic distance essentially ignores discrepancies due to small σ values: the quadratic distance is simply not appropriate to describe the divergence between two normal distributions.

It may be verified that the minimum logarithmic discrepancy of $\{\mathrm{N}(x \,|\, \mu_i, \sigma_i), \mu_i \in \Re\}$ from $\mathrm{N}(x \,|\, \mu_j, \sigma_j)$ is achieved when $\mu_i = \mu_j$, so that

$$\min_{\mu_i \in \Re} \kappa\{p_i \,|\, p_j\} = \kappa\{p_i \,|\, p_j\}\Big|_{\mu_i = \mu_j} = \frac{1}{2}\Big\{ \log \frac{\sigma_i^2}{\sigma_j^2} + \frac{\sigma_j^2}{\sigma_i^2} - 1 \Big\}. \tag{19}$$

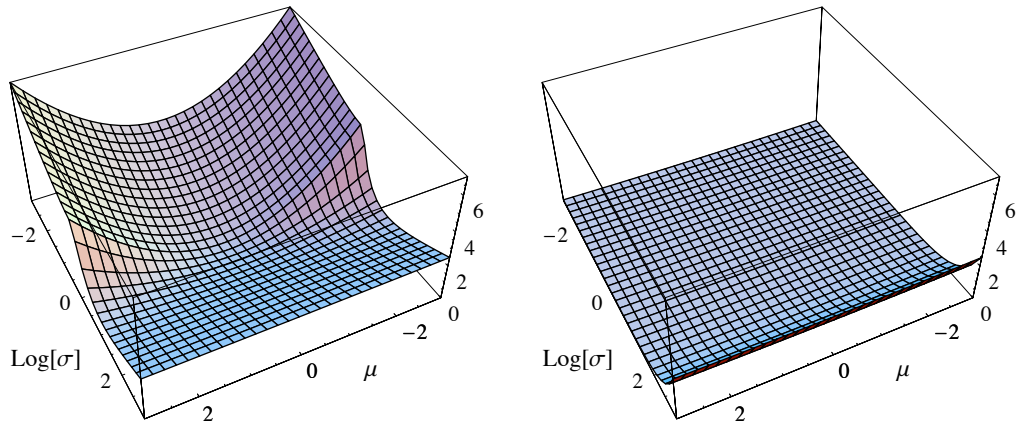

Figure 1. *Intrinsic discrepancy (left panel) and quadratic distance (right panel) between* $\mathrm{N}(x \,|\, \mu, \sigma)$ *and* $\mathrm{N}(x \,|\, 0, 1)$ *as a function of* μ *and* $\log \sigma$.

Thus, the intrinsic discrepancy between the normal $\mathrm{N}(x \,|\, \mu, \sigma)$ and the set of normals with standard deviation $\tilde{\sigma}$, $\mathcal{M}_{\tilde{\sigma}} \equiv \{\mathrm{N}(x \,|\, \tilde{\mu}, \tilde{\sigma}),\ \tilde{\mu} \in \Re)\}$ is achieved when $\tilde{\mu} = \mu$, and, using (18) and (19)

$$\delta\{\mathcal{M}_{\tilde{\sigma}},\ \mathrm{N}(x \,|\, \mu, \sigma)\} = \delta(\theta) = \begin{cases} \frac{1}{2}\left[\log \theta^{-1} + \theta - 1\right], & \text{if } \theta < 1; \\ \frac{1}{2}\left[\log \theta + \theta^{-1} - 1\right], & \text{if } \theta \geq 1. \end{cases} \tag{20}$$

which only depends on $\theta = \tilde{\sigma}^2/\sigma^2$, the ratio of the two variances. Comparison with the entropy loss (4) immediately shows that, the entropy loss is the same as the intrinsic loss for $\theta < 1$, *i.e.*, for $\tilde{\sigma}^2 < \sigma^2$, but rather different for $\theta > 1$ (*i.e.*, for $\tilde{\sigma}^2 > \sigma^2$).

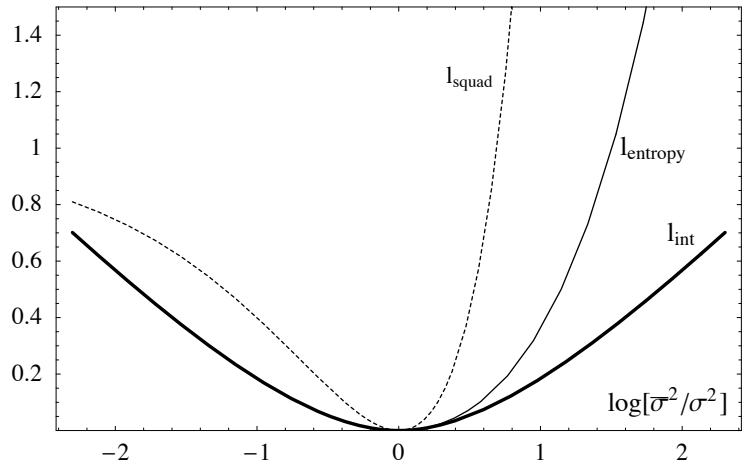

Figure 2. *Standardized quadratic loss* $\ell_{squad}(\tilde{\sigma}^2, \sigma^2)$*, entropy loss* $\ell_{entropy}(\tilde{\sigma}^2, \sigma^2)$ *and intrinsic loss,* $\ell_{int}(\tilde{\sigma}^2, \sigma^2)$ *as a function of* $\log[\tilde{\sigma}^2/\sigma^2]$.

Thus the intrinsic loss for this problem, a loss function which treats symmetrically the case $\tilde{\sigma}^2 < \sigma^2$ and the case $\tilde{\sigma}^2 > \sigma^2$, turns out to be a symmetrized version of the entropy loss which Stein (1964) suggested for this problem. This is better seen in a logarithmic scale; Figure 2 represents the standardized quadratic loss (2), the entropy loss (4) and the intrinsic loss (20) as a function of $\log[\tilde{\sigma}^2/\sigma^2]$. The dramatic overpenalization of large estimates by the standardized quadratic loss is immediately apparent; the entropy loss behaves better, but it still clearly overpenalizes large estimates.

Since the normal is a location-scale model, the reference prior when σ (or any one-to-one transformation of σ) is the parameter of interest is the universally recommended (improper) prior $\pi(\mu, \sigma) = \sigma^{-1}$. The corresponding reference posterior distribution of $\theta = \tilde{\sigma}^2/\sigma^2$, after a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of size $n \geq 2$ has been observed, is the (always proper) gamma density

$$\pi(\theta \,|\, \boldsymbol{x}) = \pi(\theta \,|\, n, s^2, \tilde{\sigma}^2) = \mathrm{Ga}\left(\theta \,\middle|\, \frac{n-1}{2}\,,\, \frac{ns^2}{2\tilde{\sigma}^2}\right), \qquad n \geq 2, \tag{21}$$

where, again, s^2 is the MLE of σ^2. Hence, the intrinsic estimator of the normal variance is that value σ^2_{int} of $\tilde{\sigma}^2$ which minimizes the expected posterior loss, *i.e.*,

$$\begin{aligned} \sigma^2_{int} &= \arg\min_{\tilde{\sigma}^2} \; d(\tilde{\sigma}^2 \,|\, n, s^2), \\ d(\tilde{\sigma}^2 \,|\, n, s^2) &= \int_0^\infty \delta(\theta)\, \pi(\theta \,|\, n, s^2, \tilde{\sigma}^2)\, \mathrm{d}\theta \end{aligned}, \tag{22}$$

where $\delta(\theta)$ is given by (20) and $\pi(\theta \,|\, n, s^2, \tilde{\sigma}^2)$ is the gamma density (21). Since (21) implies that the reference posterior distribution of $\tau = ns^2/\sigma^2$ is a central χ^2_{n-1}, the expected posterior loss from using $\tilde{\sigma}^2$ may further be written as

$$d(\tilde{\sigma}^2 \,|\, n, s^2) = d(a \,|\, n) = \int_0^\infty \delta(a\,\tau)\, \chi^2(\tau \,|\, n-1)\, \mathrm{d}\tau, \qquad a = \frac{\tilde{\sigma}^2}{ns^2}\,\cdot \tag{23}$$

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

$$\sigma^2_{int}(n, s^2) = c(n)\, s^2, \qquad c(n) = n\, a^*_n, \tag{24}$$

where a^*_n is the value of a which minimises $d(a \,|\, n)$ in (23). The exact value of a^*_n, and hence that of $c(n) = n\, a^*_n$, may be numerically found by one-dimensional numerical integration followed by numerical optimization. This is tabulated in Table 1, and represented in Figure 3. It may be appreciated that $c(2) \approx 5$ and that, for $n > 4$, $c(n)$ is reasonably well approximated by its asymptotic limit, $n/(n-2)$.

Table 1. *Exact value and approximation of the intrinsic estimator of the normal variance* $\sigma^2_{int} = c(n)\, s^2$.

n	$c(n)$	$n/(n-2)$
2	4.982	—
3	2.357	3.000
4	1.803	2.000
5	1.569	1.667
6	1.440	1.500
7	1.359	1.400
8	1.303	1.333
9	1.262	1.286
10	1.231	1.250
20	1.106	1.111
30	1.069	1.071
40	1.051	1.053
50	1.041	1.042
60	1.034	1.034
70	1.029	1.029
80	1.025	1.026
90	1.022	1.023
100	1.020	1.020

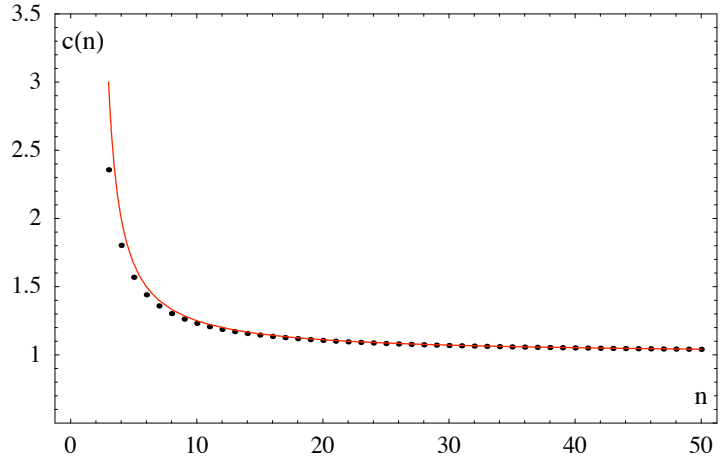

Figure 3. *The intrinsic estimator of the normal variance is* $\sigma^2_{int} = c(n)\, s^2$. *The exact values of* $c(n)$, $n > 2$, *are represented by dots and its approximation* $n/(n-2)$ *by a continuous line.*

Summarizing, $\sigma^2_{int}(2, s^2) \approx 5\, s^2$, $\sigma^2_{int}(3, s^2) \approx 2.4\, s^2$, $\sigma^2_{int}(4, s^2) \approx 1.8\, s^2$ and, for moderate sample sizes,

$$\sigma^2_{int}(n, s^2) \approx \frac{n\, s^2}{n-2} = \tilde{\sigma}^2_{n-2} \tag{25}$$

which is larger than both the MLE (which divides by n the sum of squares) and that the conventional unbiased estimate (which divides the sum of squares by $n-1$). Since intrinsic estimation is consistent under one-to-one reparametrizations, the intrinsic estimator of the standard deviation is just σ_{int}, the squared root of σ^2_{int}, and the intrinsic estimator of, say, $\log \sigma$ is simply $\log \sigma_{int}$.

4. DISCUSSION

As mentioned before, the behaviour of each possible estimator is best described by its reference posterior expected intrinsic loss,

$$d(\tilde{\sigma}^2_i \,|\, n, s^2) = \int_0^\infty \delta\{\tilde{\sigma}^2_i, \sigma^2\}\, \pi(\sigma^2 \,|\, n, s^2)\, \mathrm{d}\sigma^2, \tag{26}$$

which precisely measures, as a function of the relevant observed data (n, s^2), the expected loss to be suffered if $\tilde{\sigma}^2_i = \tilde{\sigma}^2_i(n, s^2)$ where used as a proxy for σ^2. Notice that since the intrinsic loss function δ in invariant under reparametrization, the value of the expected loss (26) is independent of the particular parametrization chosen; for instance, $d(\tilde{\sigma}_i \,|\, n, s) = d(\tilde{\sigma}^2_i \,|\, n, s)$.

It immediately follows form (23) that the expected intrinsic loss $d(\tilde{\sigma}^2_i \,|\, n, s^2)$ of any affine equivariant estimator $\tilde{\sigma}^2_i = k_n\, s^2$ is actually independent of s^2 and only depends on the sample size n. Table 2 provides, for different n values, the expected posterior intrinsic loss which correspond to the estimators σ^2_{int}, $\tilde{\sigma}^2_{n-1}$ and $\tilde{\sigma}^2_{n+1}$ which are respectively the Bayes estimators under the intrinsic loss, the entropy loss, and standardized quadratic loss.

It may be appreciated that with $n = 2$ the posterior expected loss of the intrinsic estimator σ^2_{int} is only 57% of that of the "straw man" $\tilde{\sigma}^2_{n+1}$ and 86% of the conventional estimator $\tilde{\sigma}^2_{n-1}$. As one would expect, those very large relative gains decrease with the sample size, since all those estimators converge to each other as $n \to \infty$.

Table 2. *Reference expected posterior intrinsic losses and intrinsic risks associated to the use of the intrinsic loss, the Stein loss and the quadratic loss Bayes estimators*

.

n	$d(\tilde{\sigma}^2_{int} \mid n, s^2)$ $r(\sigma^2_{int} \mid n, \sigma^2)$	$d(\tilde{\sigma}^2_{n-1} \mid n, s^2)$ $r(\tilde{\sigma}^2_{n-1} \mid n, \sigma^2))$	$d(\tilde{\sigma}^2_{n+1} \mid n, s^2)$ $r(\tilde{\sigma}^2_{n+1} \mid n, \sigma^2)$
2	0.486	0.562	0.848
3	0.225	0.250	0.381
4	0.146	0.159	0.236
5	0.109	0.116	0.168
10	0.048	0.050	0.065
20	0.023	0.024	0.028
30	0.016	0.016	0.018
40	0.012	0.012	0.013
50	0.009	0.009	0.010

The frequentist behaviour of each possible estimator is conventionally described by its the expected loss under sampling, *i.e.*, its *risk* function

$$r(\tilde{\sigma}^2_i \mid n, \sigma^2) = \int_0^\infty \delta\{\tilde{\sigma}^2_i, \sigma^2\}\, p(s^2 \mid n, \sigma^2)\, \mathrm{d}s^2. \tag{27}$$

In this problem σ^2 and s^2 play dual roles from, respectively, a Bayesian and a frequentist perspective. This follows from the interesting fact that the reference posterior distribution of $\tau = n\, s^2/\sigma^2$ is precisely the same as the sampling distribution of $t = n\, s^2/\sigma^2$, namely a χ^2_{n-1} distribution. Since the intrinsic loss is symmetric, so that $\delta\{\tilde{\sigma}^2_i, \sigma^2\} = \delta\{\sigma^2, \tilde{\sigma}^2_i\}$, this implies that the intrinsic risk of any affine equivariant estimator (the expected intrinsic loss under repeated sampling) is precisely equal to its expected reference posterior loss, that is, for all affine equivariant estimators $\tilde{\sigma}^2_i = k_n\, s^2$,

$$r(\tilde{\sigma}^2_i \mid n, \sigma^2) = d(\tilde{\sigma}^2_i \mid n, s^2), \qquad n \geq 2.$$

Hence, their risks for different sample sizes are also given by Table 2. This implies that, under intrinsic loss, the intrinsic estimator dominates all affine equivariant estimators.

It follows form the preceding discussion that, it the arguments given for the general use of the intrinsic discrepancy loss are accepted, then the *optimal* estimator of the normal variance is the intrinsic estimator $\tilde{\sigma}^2_{int}$ given by (24), quite well approximated by $\tilde{\sigma}^2_{n-2}$ for all but very small samples. Moreover, since intrinsic estimation is an invariant procedure, the optimal estimate of any one-to-one function $\psi[\sigma^2]$ of the normal variance is precisely $\tilde{\psi}_{int} = \psi[\tilde{\sigma}^2_{int}]$. Finally, the intrinsic estimator is also the best affine equivariant estimator under the frequentist decision-theoretic criterion of minimizing the (intrinsic) risk.

REFERENCES

Barnard, G. A. (1952). The frequency justification of certain sequential tests. *Biometrika* **39**, 155–150.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **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. (1999). Nested hypothesis testing: The Bayesian reference criterion. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 101–130 (with discussion).

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

Bernardo, J. M. and Juarez, M. (2003). Intrinsic estimation. *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press, 465–476.

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *Internat. Statist. Rev.* **70**, 351–372.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

Brewster, J. F. and Zidek, J. V. (1974). Improving on equivariant estimators. *Ann. Statist.* **2**, 21–38.

Brown, L. (1968). Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters. *Ann. Math. Statist.* **39**, 24–48.

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

Csörgö, S. and Faraway, J. J. (1996). On the estimation of a normal variance. *Statistics and Decisions* **14**, 23–34.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

George, E. I. (1990). Comment on Maata and Casella (1990).

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp.* **1** (J. Neyman and E. L. Scott, eds.) Berkeley: Univ. California Press, 361–380.

Jeffreys, H. (1961). *Theory of Probability*. (3rd ed.) Oxford: Oxford University Press.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation* (2nd ed.) Berlin: Springer

Lin, J.-J. and Pal, N. (2005). Comparison of normal variance estimators under multiple criteria and towards a compromise estimator. *J. Statist. Computation and Simulation* **75**, 645–666.

Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Volume 2: Inference. Cambridge: University Press.

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

Pal, N., Ling, C, and Lin, J.-J. (1998). Estimation of a normal variance—a critical review. *Statistical Papers* **39**, 389-404.

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

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Ann. Inst. Statist. Math.* **16**, 155–160.