BAYESIAN STATISTICS 7, pp. 465–476
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, 2003

Intrinsic Estimation

JOSÉ M. BERNARDO and MIGUEL A. JUÁREZ
Universitat de València, Spain
jose.m.bernardo@uv.es   miguel.juarez@uv.es

SUMMARY

In this paper, the problem of parametric point estimation is addressed from an objective Bayesian viewpoint. Arguing that pure statistical estimation may be appropriately described as a precise decision problem, where the loss function is a measure of the divergence between the assumed model and the estimated model, the information-based *intrinsic discrepancy* is proposed as an appropriate loss function. The *intrinsic estimator* is then defined as that minimizing the expected loss with respect to the appropriate *reference* posterior distribution. The resulting estimators are shown to have attractive *invariance* properties. As demonstrated with illustrative examples, the proposed theory either leads to new, arguably better estimators, or provides a new perspective on well-established solutions.

Keywords: INTRINSIC DISCREPANCY; INTRINSIC LOSS; LOGARITHMIC KL-DIVERGENCE; POINT ESTIMATION; REFERENCE ANALYSIS; REFERENCE PRIORS.

1. INTRODUCTION

It is well known that, from a Bayesian viewpoint, the final result of *any* problem of statistical inference is the posterior distribution of the quantity of interest. However, in more than two dimensions, the description (either graphical or analytical) of the posterior distribution is difficult and some "location" measure is often required for descriptive purposes. Moreover, there are many situations where a point estimate of the quantity of interest is specifically needed (and often even legally required) as part of the statistical report; simple examples include quoting the optimal dose of a drug per kilogram of body weight, or estimating the net weight of a canned food.

The typical Bayesian approach to point estimation formulates the problem as a decision problem, where the action space is the set of possible values of the quantity of interest. For each loss function and prior distribution on the model parameters, the *Bayes estimator* is obtained as that which minimizes the corresponding posterior expected loss. It is well known that the solution may dramatically depend both on the choice of the loss function and on the choice of the prior distribution.

In practice, in most situations where point estimation is of interest, an *objective* point estimate of the quantity of interest is actually required: objective in the very precise sense of exclusively depending on the assumed probability model (*i.e.*, on the conditional distribution of the data given the parameters) and the available data. Moreover, in purely inferential settings (where interest focuses on the actual mechanism which governs the data), this estimate is typically required to be *invariant* under one-to-one transformations of either the data or the parameter space. In this paper, an information-theory based loss function is combined with reference analysis to propose an objective Bayesian approach to point estimation which satisfies these desiderata.

In Section 2, the standard Bayesian formulation of point estimation as a decision problem is recalled and its conventional "automatic" answers are briefly discussed. Section 3 presents the proposed methodology. In Section 4, a number of illustrative examples are discussed. Section 5 contains some final remarks and suggests areas for additional research.

2. THE FORMAL DECISION PROBLEM

Let $\{p(\boldsymbol{x} \mid \boldsymbol{\theta}), \boldsymbol{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta\}$ be a *probability model* assumed to describe the probabilistic behavior of the observable data $\boldsymbol{x}$, and suppose that a point estimator $\boldsymbol{\theta}^e = \boldsymbol{\theta}^e(\boldsymbol{x})$ of the parameter $\boldsymbol{\theta}$ is required. It is well known that this can be formulated as a decision problem under uncertainty, where the action space is the class $\mathcal{A} = \{\boldsymbol{\theta}^e \in \Theta\}$ of possible parameter values. In a purely inferential setting, the optimal estimate $\boldsymbol{\theta}^*$ is supposed to identify the best proxy, $p(\boldsymbol{x} \mid \boldsymbol{\theta}^*)$, to the unknown probability model, $p(\boldsymbol{x} \mid \boldsymbol{\theta}^a)$, where $\boldsymbol{\theta}^a$ stands for the *actual* (unknown) value of the parameter.

Let $l(\boldsymbol{\theta}^e, \boldsymbol{\theta}^a)$ be a *loss function* measuring the consequences of estimating $\boldsymbol{\theta}^a$ by $\boldsymbol{\theta}^e$. In a purely inferential context, $l(\boldsymbol{\theta}^e, \boldsymbol{\theta}^a)$ should measure the consequences of using the model $p(\boldsymbol{x} \mid \boldsymbol{\theta}^e)$ instead of the true, unknown model $p(\boldsymbol{x} \mid \boldsymbol{\theta}^a)$. For any loss function $l(\boldsymbol{\theta}^e, \boldsymbol{\theta}^a)$ and (possibly improper) prior $p(\boldsymbol{\theta})$, the *Bayes estimator* $\boldsymbol{\theta}^b = \boldsymbol{\theta}^b(\boldsymbol{x})$ of the parameter $\boldsymbol{\theta}$ is that minimizing the corresponding posterior loss, so that

$$\boldsymbol{\theta}^b(\boldsymbol{x}) = \arg \min_{\boldsymbol{\theta}^e \in \Theta} \int_\Theta l\,(\boldsymbol{\theta}^e, \boldsymbol{\theta})\, p\,(\boldsymbol{\theta} \mid \boldsymbol{x})\; d\boldsymbol{\theta}, \tag{1}$$

where $p\,(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$ is the posterior distribution of the parameter vector $\boldsymbol{\theta}$.

A number of conventional loss functions have been proposed in the literature, and their associated Bayes estimators are frequently quoted in Bayesian analysis:

Squared error loss. If the loss function is quadratic, of the form $(\boldsymbol{\theta}^e - \boldsymbol{\theta}^a)^t W(\boldsymbol{\theta}^e - \boldsymbol{\theta}^a)$, where W is a (known) symmetric positive definite matrix, then the Bayes estimator is the *posterior mean* $\mathrm{E}[\boldsymbol{\theta} \mid \boldsymbol{x}]$, provided it exists.

Zero-one loss. If the loss function takes the value zero if $\boldsymbol{\theta}^e$ belongs to a ball of radius ϵ centered at $\boldsymbol{\theta}^a$, and the value one otherwise, then the Bayes estimator tends towards the *posterior mode* $\mathrm{Mo}[\boldsymbol{\theta} \mid \boldsymbol{x}]$ as $\epsilon \to 0$, provided the mode exists and is unique.

Absolute error loss. If θ is *one-dimensional*, and the loss function is of the form $c\,|\theta^e - \theta^a|$, for some $c > 0$, then the Bayes estimator is the *posterior median* $\mathrm{Me}[\theta \mid \boldsymbol{x}]$.

Neither the posterior mean nor the posterior mode are invariant under one-to-one transformations of the parameter of interest. Thus, $\boldsymbol{\theta}^e$ can be declared to be the best estimator of $\boldsymbol{\theta}^a$, while $\boldsymbol{\phi}(\boldsymbol{\theta}^e)$ is declared *not* to be the best estimator for $\boldsymbol{\phi}^a = \boldsymbol{\phi}(\boldsymbol{\theta}^a)$; this is unattractive in a scientific, purely inferential context, where interest is explicitly focused on identifying the actual probability model $p(\boldsymbol{x} \mid \boldsymbol{\theta}^a) = p(\boldsymbol{x} \mid \boldsymbol{\phi}^a)$. The one-dimensional posterior median *is* invariant, but is not easily generalizable to more than one dimension. Naturally, if besides the likelihood and the prior, the functional form of the loss function is consistently transformed, invariance would be achieved, but this is rarely done in purely inferential problems. Hence, it is suggested that *invariant loss functions* should be used. More precisely it is argued that, in a purely inferential context, the loss function $l(\boldsymbol{\theta}^e, \boldsymbol{\theta})$ should *not* be chosen to measure the discrepancy between $\boldsymbol{\theta}^e$ and $\boldsymbol{\theta}^a$, but to directly measure the discrepancy between the models $p(\boldsymbol{x} \mid \boldsymbol{\theta}^e)$ and $p(\boldsymbol{x} \mid \boldsymbol{\theta}^a)$ which they label. This type of *intrinsic* loss is typically invariant under reparametrization, and therefore produces invariant estimators.

An appropriate choice of the loss function is, however, only part of the solution. To obtain an objective Bayes estimator, an *objective prior* must be used. It is argued that reference analysis may successfully be used to provide an adequate prior specification.

3. INTRINSIC ESTIMATION

3.1. *The Loss Function*

Conventional loss functions typically depend on the particular metric used to index the model, being defined as a measure of the distance between the parameter and its estimate. We claim that, in a purely inferential context, one should rather be interested in the discrepancy between the models labelled by the true value of the parameter and its estimate. A loss function of the form $l\,(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = l\,\{p\,(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1)\,, p\,(\boldsymbol{x}\,|\,\boldsymbol{\theta}_2)\}$ is called an *intrinsic loss* (Robert, 1996).

Bernardo (1979a) and Bernardo and Smith (1994, Ch. 3) argue that scientific inference is well described as a formal decision problem, where the terminal loss function is a proper scoring rule. One of the most extensively studied of these is the *directed logarithmic divergence* (Gibbs, 1902; Shannon, 1948; Jeffreys, 1948; Good, 1950; Kullback and Leibler, 1951; Chernoff, 1952; Savage, 1954; Huzurbazar, 1955; Kullback, 1959; Jaynes, 1983). If $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1)$ and $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_2)$ are probability densities with the same support $\mathcal{X}$, the directed logarithmic divergence of $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_2)$ from $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1)$ is defined as

$$k_{\mathcal{X}}(\boldsymbol{\theta}_2\,|\,\boldsymbol{\theta}_1) = \int_{\mathcal{X}} p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1)\;\log\frac{p\,(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1)}{p\,(\boldsymbol{x}\,|\,\boldsymbol{\theta}_2)}\,d\boldsymbol{x}. \tag{2}$$

The directed logarithmic divergence (often referred to as Kullback–Leibler divergence) is *non-negative*, and it is *invariant* under one-to-one transformations of either $\boldsymbol{x}$ or $\boldsymbol{\theta}$. It is also *additive* in the sense that, if $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$ are conditionally independent given $\boldsymbol{\theta}$, then the divergence $k_{\mathcal{X},\mathcal{Y}}(\boldsymbol{\theta}_2\,|\,\boldsymbol{\theta}_1)$ of $p(\boldsymbol{x}, \boldsymbol{y}\,|\,\boldsymbol{\theta}_2)$ from $p(\boldsymbol{x}, \boldsymbol{y}\,|\,\boldsymbol{\theta}_1)$ is simply $k_{\mathcal{X}}(\boldsymbol{\theta}_2\,|\,\boldsymbol{\theta}_1) + k_{\mathcal{Y}}(\boldsymbol{\theta}_2\,|\,\boldsymbol{\theta}_1)$; in particular, if data $\boldsymbol{x}$ are assumed to be a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from $p(x\,|\,\boldsymbol{\theta})$, then the divergence of $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_2)$ from $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1)$ is simply n times the divergence of $p(x\,|\,\boldsymbol{\theta}_2)$ from $p(x\,|\,\boldsymbol{\theta}_1)$. Under appropriate regularity conditions, there are many connections between the logarithmic divergence and Fisher's information (see, *e.g.*, Stone, 1959; Bernardo and Smith, 1994, Ch. 5; Schervish, 1995, p. 118). Furthermore, $k_{\mathcal{X}}(\boldsymbol{\theta}_2\,|\,\boldsymbol{\theta}_1)$ has an attractive interpretation in information-theoretical terms: it is the expected amount of information (in natural units, *nits*) necessary to recover $p\,(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1)$ from $p\,(\boldsymbol{x}\,|\,\boldsymbol{\theta}_2)$.

However, the Kullback–Leibler divergence is not symmetric and it diverges if the support of $p\,(\boldsymbol{x}\,|\,\boldsymbol{\theta}_2)$ is a strict subset of the support of $p\,(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1)$. To simultaneously address these two unwelcome features we propose to use the symmetric *intrinsic discrepancy* $\delta_{\mathcal{X}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, introduced in Bernardo and Rueda (2002), and defined as $\delta_{\mathcal{X}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \min\{k_{\mathcal{X}}(\boldsymbol{\theta}_1\,|\,\boldsymbol{\theta}_2), k_{\mathcal{X}}(\boldsymbol{\theta}_2\,|\,\boldsymbol{\theta}_1)\}$. To simplify the notation, the subindex $\mathcal{X}$ will be dropped from both $\delta_{\mathcal{X}}(\boldsymbol{\theta}_2\,|\,\boldsymbol{\theta}_1)$ and $k_{\mathcal{X}}(\boldsymbol{\theta}_2\,|\,\boldsymbol{\theta}_1)$ whenever there is no danger of confusion.

Definition 1. (*Intrinsic Discrepancy Loss*). Let $\{p(\boldsymbol{x}\,|\,\boldsymbol{\theta}), \boldsymbol{x} \in \mathcal{X}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ be a family of probability models for $\boldsymbol{x} \in \mathcal{X}(\boldsymbol{\theta})$, where the sample space may depend on the parameter value. The intrinsic discrepancy, $\delta_{\mathcal{X}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, between $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1)$ and $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_2)$ is defined as

$$\min\left\{\int_{\mathcal{X}(\boldsymbol{\theta}_1)} p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1)\log\left[\frac{p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1)}{p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_2)}\right]d\boldsymbol{x}, \int_{\mathcal{X}(\boldsymbol{\theta}_2)} p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_2)\log\left[\frac{p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_2)}{p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1)}\right]d\boldsymbol{x}\right\}$$

provided one of the two integrals is finite.

The intrinsic discrepancy inherits a number of attractive properties from the directed logarithmic divergence. Indeed, it is non-negative and vanishes if, and only if, $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1) = p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_2)$

almost everywhere; it is invariant under one-to-one transformations of either $\boldsymbol{x}$ or $\boldsymbol{\theta}$; if the available data $\boldsymbol{x}$ consist of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from $p(x \,|\, \boldsymbol{\theta})$, then the intrinsic divergence between $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1)$ and $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_2)$ is simply n times the intrinsic divergence between $p(x \,|\, \boldsymbol{\theta}_1)$ and $p(x \,|\, \boldsymbol{\theta}_2)$. However, in contrast with the directed logarithmic divergence, the intrinsic discrepancy is *symmetric* and, if $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1)$ and $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_2)$ have nested supports, so that $p\,(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1) > 0$ iff $\boldsymbol{x} \in \mathcal{X}(\boldsymbol{\theta}_1)$, $p\,(\boldsymbol{x} \,|\, \boldsymbol{\theta}_2) > 0$ iff $\boldsymbol{x} \in \mathcal{X}(\boldsymbol{\theta}_2)$, and either $\mathcal{X}(\boldsymbol{\theta}_1) \subset \mathcal{X}(\boldsymbol{\theta}_2)$ or $\mathcal{X}(\boldsymbol{\theta}_2) \subset \mathcal{X}(\boldsymbol{\theta}_1)$, then the intrinsic discrepancy is typically finite, and reduces to a directed logarithmic divergence. More specifically, $\delta\,(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = k\,(\boldsymbol{\theta}_1 \,|\, \boldsymbol{\theta}_2)$ when $\mathcal{X}(\boldsymbol{\theta}_2) \subset \mathcal{X}(\boldsymbol{\theta}_1)$, and $\delta\,(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = k\,(\boldsymbol{\theta}_2 \,|\, \boldsymbol{\theta}_1)$ when $\mathcal{X}(\boldsymbol{\theta}_1) \subset \mathcal{X}(\boldsymbol{\theta}_2)$.

3.2. *The Prior Function*

Under the Bayesian paradigm, the outcome of any inference problem (the posterior distribution of the quantity of interest) combines the information provided by the data with relevant available prior information. In many situations, however, either the available prior information on the quantity of interest is too vague to warrant the effort required to have it formalized in the form of a probability distribution, or it is too subjective to be useful in scientific communication or public decision-making. It is, therefore, important to be able to identify the mathematical form of a "relatively uninformative" prior function, *i.e.*, a function (not necessarily a probability distribution) that, when formally used as a prior distribution in Bayes theorem, would have a minimal effect, relative to the data, on the posterior inference. More formally, suppose that the probability mechanism which has generated the available data $\boldsymbol{x}$ is assumed to be $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, for some $\boldsymbol{\theta} \in \Theta$, and that the quantity of interest is some real-valued function $\phi = \phi(\boldsymbol{\theta})$ of the model parameter $\boldsymbol{\theta}$. Without loss of generality, it may be assumed that the probability model is of the form $p(\boldsymbol{x} \,|\, \phi, \boldsymbol{\lambda})$, $\phi \in \Phi$, $\boldsymbol{\lambda} \in \Lambda$, where $\boldsymbol{\lambda}$ is some appropriately chosen nuisance parameter vector. What is then required is to identify that joint prior function $\pi_\phi(\phi, \boldsymbol{\lambda})$ which would have a *minimal effect* on the corresponding marginal posterior distribution of the quantity of interest ϕ,

$$\pi(\phi \,|\, \boldsymbol{x}) \propto \int_\Lambda p(\boldsymbol{x} \,|\, \phi, \boldsymbol{\lambda})\, \pi_\phi(\phi, \boldsymbol{\lambda})\, d\boldsymbol{\lambda},$$

a prior which, to use a conventional expression, "would let the data speak for themselves" about the likely values of ϕ. Note that, within a given probability model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, the prior which could be described as "relatively uninformative" about the value of $\phi = \phi(\boldsymbol{\theta})$ will typically depend on the particular quantity of interest, $\phi = \phi(\boldsymbol{\theta})$.

Much work has been done to formulate priors which make the idea described above mathematically precise. Using an information-theoretic approach, Bernardo (1979b) introduced an algorithm to derive *reference* distributions; this is possibly the most general approach available. The reference prior $\pi_\phi(\boldsymbol{\theta})$ identifies a *possible* prior for $\boldsymbol{\theta}$, namely that describing a situation were relevant knowledge about the quantity of interest $\phi = \phi(\boldsymbol{\theta})$ (beyond that universally accepted) may be held to be negligible compared to the information about that quantity which repeated experimentation from a particular data generating mechanism $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$ might possibly provide. More recent work containing many refinements to the original formulation include Berger and Bernardo (1989, 1992), Bernardo and Smith (1994, Ch. 5) and Bernardo (1997). Bernardo and Ramón (1998) offers a simple introduction to reference analysis.

Any statistical analysis obviously contains a fair number of subjective elements; these include (among others) the data selected, the model assumptions and the choice of the quantities of interest. Reference analysis may be argued to provide "objective" Bayesian inferences in precisely the same sense that conventional statistical methods claim to be "objective"; namely in the sense that the solutions provided depend exclusively on the model assumptions and on the observed data.

In any decision problem, the quantity of interest is that function of the parameters which enters the loss function. Formally, in a decision problem with uncertainty about $\boldsymbol{\theta}$, actions $\{a \in \mathcal{A}\}$, and loss function $l(a, \phi(\boldsymbol{\theta}))$, the quantity of interest is $\phi = \phi(\boldsymbol{\theta})$. We have argued that, in point estimation, an appropriate loss function is the intrinsic discrepancy $l(\boldsymbol{\theta}^e, \boldsymbol{\theta}) = \delta(\boldsymbol{\theta}^e, \boldsymbol{\theta})$. It follows that, to obtain an objective (reference) intrinsic estimator, one should minimize the expected intrinsic loss with respect to the reference posterior distribution $\pi_\delta(\boldsymbol{\theta} \,|\, \boldsymbol{x})$, derived from the reference prior $\pi_\delta(\boldsymbol{\theta})$ obtained when the quantity of interest is the intrinsic discrepancy $\delta = \delta(\boldsymbol{\theta}^e, \boldsymbol{\theta})$; thus, one should minimize

$$d(\boldsymbol{\theta}^e \,|\, \boldsymbol{x}) = \mathrm{E}[\delta \,|\, \boldsymbol{x}] = \int_\Theta \delta(\boldsymbol{\theta}^e, \boldsymbol{\theta})\, \pi_\delta(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}. \tag{3}$$

Definition 2. (*Intrinsic Estimator*). Let $\{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}), \boldsymbol{x} \in \mathcal{X}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ be a family of probability models for some observable data $\boldsymbol{x}$, where the sample space may possibly depend on the parameter value. The intrinsic estimator,

$$\boldsymbol{\theta}^*(\boldsymbol{x}) = \arg\min_{\boldsymbol{\theta}^e \in \Theta} d(\boldsymbol{\theta}^e \,|\, \boldsymbol{x}) = \arg\min_{\boldsymbol{\theta}^e \in \Theta} \int_\Theta \delta\,(\boldsymbol{\theta}^e, \boldsymbol{\theta})\; \pi_\delta\,(\boldsymbol{\theta} \,|\, \boldsymbol{x})\; d\boldsymbol{\theta}$$

is that minimizing the *reference* posterior expectation of the intrinsic discrepancy.

Reference distributions are known to be invariant under piecewise invertible transformations of the parameter (Datta and Ghosh, 1996) in the sense that, for any such transformation $\boldsymbol{\omega}(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$, the reference posterior of $\boldsymbol{\omega}$, $\pi(\boldsymbol{\omega} \,|\, \boldsymbol{x})$ is that obtained from $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ by standard probability calculus. Since the intrinsic discrepancy is itself invariant, it follows that (for any dimensionality) the intrinsic estimator is *invariant* under piecewise invertible transformations; thus, for any such transformation $\boldsymbol{\omega}(\boldsymbol{\theta})$ of the parameter vector, one has $\boldsymbol{\omega}^*(\boldsymbol{x}) = \boldsymbol{\omega}(\boldsymbol{\theta}^*(\boldsymbol{x}))$.

3.1. *A Simple Example: Bernoulli Data*

Let data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ consist of n conditionally independent Bernoulli observations with parameter θ, so that $p(x \,|\, \theta) = \theta^x (1-\theta)^{1-x}$, $x \in \{0, 1\}$. It is easily verified that the directed logarithmic divergence of $p(x \,|\, \theta_2)$ from $p(x \,|\, \theta_1)$ is

$$k(\theta_2 \,|\, \theta_1) = \theta_1 \log[\theta_1/\theta_2] + (1-\theta_1) \log[(1-\theta_1)/(1-\theta_2)]$$

Moreover, it is easily shown that $k(\theta_2 \,|\, \theta_1) < k(\theta_1 \,|\, \theta_2)$ iff $\theta_1 < \theta_2 < 1 - \theta_1$; thus, the intrinsic discrepancy between $p(\boldsymbol{x} \,|\, \theta^e)$ and $p(\boldsymbol{x} \,|\, \theta)$, represented in the left pane of Figure 1, is

$$\delta(\theta^e, \theta) = n \begin{cases} k(\theta \,|\, \theta^e) & \theta \in (\theta^e, 1-\theta^e), \\ k(\theta^e \,|\, \theta) & \text{otherwise.} \end{cases}$$

Since $\delta(\theta^e, \theta)$ is a piecewise invertible function of θ, the δ-reference prior is just the θ-reference prior and, since Bernoulli is a regular model, this is Jeffreys prior, $\pi(\theta) = \mathrm{Be}(\theta \,|\, \frac{1}{2}\,, \frac{1}{2}\,)$. The corresponding reference posterior is the Beta distribution $\pi(\theta \,|\, \boldsymbol{x}) = \mathrm{Be}(\theta \,|\, r + \frac{1}{2}\,, n - r + \frac{1}{2}\,)$, with $r = \sum x_i$, and the reference expected posterior intrinsic discrepancy is the concave function

$$d\,(\theta^e, \boldsymbol{x}) = \int_0^1 \delta\,(\theta^e, \theta)\;\mathrm{Be}(\theta \,|\, r + \tfrac{1}{2}, n - r + \tfrac{1}{2})\, d\theta.$$

The intrinsic estimator is its unique minimum $\theta^*(\boldsymbol{x}) = \arg\min_{\theta^e \in (0,1)} d\,(\theta^e, \boldsymbol{x})$, which is easily computed by one-dimensional numerical integration. A very good approximation is given by

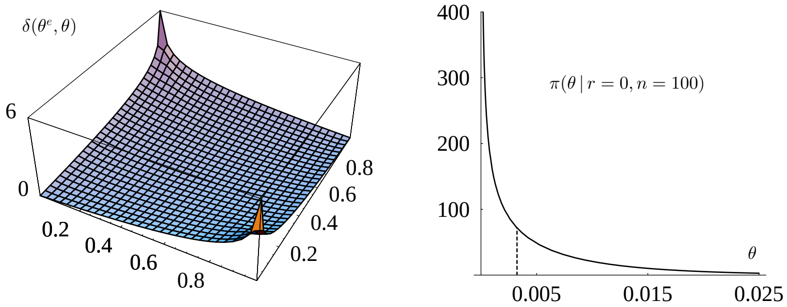

Figure 1. *Intrinsic discrepancy and reference posterior density for a Bernoulli parameter.*

the arithmetic average of the Bayes estimators which corresponds to using $k(\theta \mid \theta^e)$ and $k(\theta^e \mid \theta)$ as loss functions,

$$\theta^*(\boldsymbol{x}) \approx \frac{1}{2}\left(\frac{r+1/2}{n+1} + \frac{\exp[\psi(r+1/2)]}{\exp[\psi(r+1/2)] + \exp[\psi(n-r+1/2)]}\right), \tag{4}$$

where $\psi(.)$ is the digamma function.

As a numerical illustration, suppose that, to investigate the prevalence of a rare disease, a random sample of size $n = 100$ has been drawn and that no affected person has been found, so that $r = 0$. The reference posterior is $\text{Be}(\theta \mid 0.5, 100.5)$ (shown in the right pane of Figure 1), and the exact intrinsic estimator (shown with a dashed line) is $\theta^*(\boldsymbol{x}) = 0.00324$. The approximation (4) yields $\theta^*(\boldsymbol{x}) \approx 0.00318$. The posterior median is 0.00227. We note in passing that the MLE estimator, $\hat{\theta} = r/n = 0$, is obviously misleading in this case.

4. FURTHER EXAMPLES

To illustrate the above methodology and to compare the resulting estimators with those derived by conventional methods, a few more examples will be discussed.

4.1. *Uniform model,* $\text{Un}(x \mid 0, \theta)$

Consider first a simple non-regular example. Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, be a random sample from the uniform distribution $\text{Un}(x \mid 0, \theta) = \theta^{-1}$, $0 < x < \theta$. It is immediately verified that $t = \max\{x_1, \ldots, x_n\}$ is a sufficient statistic. The directed logarithmic divergence of $\text{Un}(x \mid 0, \theta_2)$ from $\text{Un}(x \mid 0, \theta_1)$ is

$$k(\theta_1 \mid \theta_2) = n \begin{cases} \log(\theta_1/\theta_2) & \theta_1 \geq \theta_2, \\ \infty & \theta_1 < \theta_2; \end{cases}$$

thus the intrinsic discrepancy between $p(\boldsymbol{x} \mid \theta^e)$ and $p(\boldsymbol{x} \mid \theta)$ is

$$\delta(\theta^e, \theta) = n \begin{cases} \log(\theta^e/\theta) & \theta \leq \theta^e, \\ \log(\theta/\theta^e) & \theta \geq \theta^e, \end{cases}$$

shown in the left pane of Figure 2. Since the intrinsic discrepancy $\delta(\theta^e, \theta)$ is a piecewise invertible function of θ, the δ-reference prior is also the θ-reference prior. Since the sample space

$\mathcal{X}(\theta) = (0, \theta)$ depends on the parameter θ, this is not a regular problem and, hence, Jeffreys prior is not defined. The general expression for the reference prior in one-dimensional continuous problems with an asymptotically sufficient, consistent estimator $\tilde{\theta} = \tilde{\theta}(\boldsymbol{x})$ is (Bernardo and Smith, 1994, p. 312)

$$\pi(\theta) \propto p^*(\theta \,|\, \tilde{\theta})\Big|_{\tilde{\theta}=\theta} \tag{5}$$

where $p^*(\theta \,|\, \tilde{\theta})$ is any asymptotic approximation to the posterior distribution of θ (an expression that reduces to Jeffreys prior in regular problems).

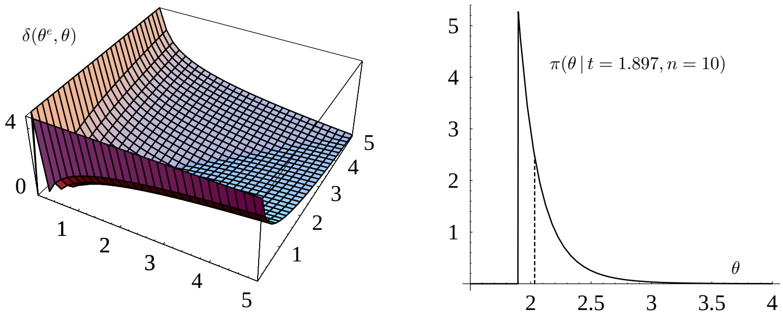

Figure 2. *Intrinsic discrepancy and reference density for the parameter θ of a uniform* $\mathrm{Un}(\cdot \,|\, 0, \theta)$ *model.*

In this problem, the likelihood function is $L(\theta \,|\, \boldsymbol{x}) = \theta^{-n}$, if $\theta > t$, and zero otherwise, where $t = \max\{x_1, \ldots, x_n\}$. Hence, an asymptotic posterior is $p^*(\theta \,|\, t, n) \propto \theta^{-n}, \theta > t$. Computing the missing proportionality constant yields $p^*(\theta \,|\, t, n) = (n-1)t^{n-1}\theta^{-n}$. Since t is a sufficient, consistent estimator of θ, Eq. (5) may be used to obtain the θ-reference prior as $\pi(\theta) \propto t^{n-1}\theta^{-n}|_{t=\theta} = \theta^{-1}$. The corresponding posterior is the Pareto distribution $\pi(\theta \,|\, \boldsymbol{x}) = \mathrm{Pa}(\theta \,|\, n, t) = n\, t^n\, \theta^{-(n+1)}$, $\theta > t$. The reference expected posterior intrinsic discrepancy is then easily found to be

$$d\,(\theta^e, \boldsymbol{x}) = \int_t^\infty \delta(\theta^e, \theta)\ \mathrm{Pa}(\theta \,|\, n, t)\, d\theta = 2\Big(\frac{t}{\theta^e}\Big)^n - n\, \log\Big(\frac{t}{\theta^e}\Big) - 1,$$

which is minimized at $\theta^e = 2^{1/n}\ t$. Hence, the intrinsic estimator is $\theta^*(\boldsymbol{x}) = 2^{1/n}\ t$, which is actually the median of the reference posterior.

As an illustration, a random sample of size $n = 10$ was simulated from a Uniform distribution $\mathrm{Un}(x \,|\, 0, \theta)$ with $\theta = 2$, yielding a maximum $t = 1.897$. The corresponding reference posterior, $\mathrm{Pa}(\theta \,|\, 10, 1.897)$, is shown in the right pane of Figure 2. The intrinsic estimator, $\theta^*(\boldsymbol{x}) = 2.033$ is indicated with a dashed line.

4.2. *Normal Variance with Known Mean*

Combining the invariance properties of the intrinsic discrepancy with a judicious choice of the parametrization often simplifies the required computations. As an illustration, consider estimation of the normal variance.

Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from a Normal $\mathrm{N}(x \,|\, \mu, \sigma^2)$ distribution with known mean μ, and let s_x^2 be the corresponding sample variance, so that $ns_x^2 = \sum_j (x_j - \mu)^2$. The required intrinsic discrepancy is $n\,\delta\{N(x \,|\, \mu, \sigma^2), N(x \,|\, \mu, \sigma_e^2)\}$; letting $y = (x - \mu)/\sigma_e$ and using the fact that the intrinsic discrepancy is invariant under one-to-one transformations of the data, this may be written as $n\,\delta\{N(y\,|\,0,\sigma^2/\sigma_e^2), N(y\,|\,0,1)\}$. The last discrepancy has a simple expression in terms of $\theta = \theta(\sigma_e) = \log(\sigma/\sigma_e)$; specifically,

$$\delta\{N(x\,|\,\mu,\sigma^2), N(x\,|\,\mu,\sigma_e^2)\} = \delta(\theta) = 2|\theta| + 2\exp(-2|\theta|) - 1, \tag{6}$$

a symmetric function around zero, which is represented in the left pane of Figure 3. Moreover, since δ is a piecewise invertible function of θ (and hence of σ) and this is a regular problem, the reference prior for δ is the same as the reference prior for σ, the corresponding Jeffreys prior, $\pi(\sigma) = \sigma^{-1}$; in terms of θ, this transforms into the uniform prior $\pi_\delta(\theta) = 1$. The corresponding reference posterior is easily found to be

$$\pi(\theta\,|\,\boldsymbol{x},\sigma_e) = 2e^{-2\theta}\,\mathrm{Ga}\Big(\lambda \;\Big|\; \frac{n}{s}, \frac{ns_y^2}{2}\Big)\Big|_{\lambda=e^{-2\theta}}, \qquad ns_y^2 = \frac{ns_x^2}{\sigma_e^2}.$$

The intrinsic estimator $\sigma^*(\boldsymbol{x})$ is that value of σ_e which minimizes the expected reference posterior discrepancy:

$$\sigma^*(\boldsymbol{x}) = \arg\min_{\sigma_e>0}\; d(\sigma_e\,|\,\boldsymbol{x}) = \arg\min_{\sigma_e>0} \int_{\Re} \delta(\theta)\,\pi(\theta\,|\,\boldsymbol{x},\sigma_e)\,d\theta, \tag{7}$$

which may easily found by numerical methods. A simple, extremely good approximation to $\sigma^*(\boldsymbol{x})$ is easily derived. Indeed, expanding $\delta(\theta)$ around zero shows that $\delta(\theta)$ behaves as θ^2 near the origin; thus,

$$d(\sigma_e\,|\,\boldsymbol{x}) = \int_{\Re} \delta(\theta)\,\pi(\theta\,|\,\boldsymbol{x},\sigma_e)\,d\theta \approx \int_0^\infty (\log\sigma - \log\sigma_e)^2\,\pi(\sigma\,|\,\boldsymbol{x})\,d\sigma,$$

and the last integral is minimized by σ_e^* such that

$$\log\sigma_e^* = \mathrm{E}[\log\sigma\,|\,\boldsymbol{x}] = \frac{1}{2}\Big[\log\frac{ns_x^2}{2} - \psi\Big(\frac{n}{2}\Big)\Big],$$

where $\psi(.)$ is the digamma function which, for moderate values of x, is well approximated by $\log x - (2x)^{-1}$. Since intrinsic estimation is an invariant procedure under reparametrization, this provides an approximation to the intrinsic estimator of the standard deviation given by

$$\sigma^*(\boldsymbol{x}) \approx s_x\,\frac{n+1/2}{n}; \qquad s_x = \sqrt{\frac{\sum_{j=1}^n (x_j-\mu)^2}{n}}\,, \tag{8}$$

which is slightly larger than the MLE estimator $\hat\sigma = s_x$.

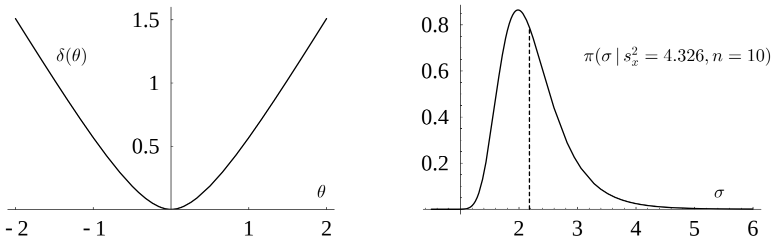

Figure 3. *Intrinsic discrepancy between $N(x\,|\,\mu,\sigma^2)$ and $N(x\,|\,\mu,\sigma_e^2)$ in terms of $\theta = \log(\sigma/\sigma_e)$, and marginal reference posterior of the standard deviation σ, given a simulated random sample of size $n = 10$ from a Normal $\mathrm{N}(x\,|\,0,2^2)$ distribution.*

As a numerical illustration, a sample of size $n = 10$ was simulated from a Normal distribution, $\mathrm{N}(x \mid 0,\, 2^2)$, yielding $s_x^2 = 4.326$. The corresponding reference posterior, the square root of an inverted gamma with parameters 5 and 21.63, is shown in the right pane of Figure 3. The exact intrinsic estimator, obtained from (7), is $\sigma^*(\boldsymbol{x}) = 2.178$, indicated with a dashed line. The approximation (8) yields $\sigma^*(\boldsymbol{x}) \approx 2.184$. The MLE is $\hat{\sigma} = s_x = 2.080$.

4.3. *Multivariate Mean Vector*

Let data consist of the mean vector $\overline{\boldsymbol{x}}$ from k-variate normal $\mathrm{N}_k(\overline{\boldsymbol{x}} \mid \boldsymbol{\mu}, n^{-1}\boldsymbol{I})$. The directed logarithmic divergence of $p(\overline{\boldsymbol{x}} \mid \boldsymbol{\mu}^e)$ from $p(\overline{\boldsymbol{x}} \mid \boldsymbol{\mu})$ is symmetric in this case, and hence equal to the intrinsic discrepancy

$$\delta(\boldsymbol{\mu}^e, \boldsymbol{\mu}) = \frac{n}{2}(\boldsymbol{\mu}^e - \boldsymbol{\mu})^t(\boldsymbol{\mu}^e - \boldsymbol{\mu}) = \frac{n}{2}\,\phi,$$

where $\phi = (\boldsymbol{\mu}^e - \boldsymbol{\mu})^t(\boldsymbol{\mu}^e - \boldsymbol{\mu})$. Thus, in this problem, the intrinsic discrepancy loss is a quadratic loss in terms of the parameter vector $\boldsymbol{\mu}$.

The intrinsic discrepancy is a linear function of $\phi = \|\boldsymbol{\mu}^e - \boldsymbol{\mu}\|$. Changing to centered generalized polar coordinates, it is found (Bernardo, 1979b; Ferrándiz, 1985; Berger *et al.*, 1998) that the reference posterior density for ϕ is

$$\pi(\phi \mid \overline{\boldsymbol{x}}) = \pi(\phi \mid t) \propto p(t \mid \phi)\, \pi(\phi) \propto \chi^2(nt \mid k, n\phi)\, \phi^{-1/2},$$

where $t = (\boldsymbol{\mu}^e - \overline{\boldsymbol{x}})^t(\boldsymbol{\mu}^e - \overline{\boldsymbol{x}})$. Note that this is *very different* from the posterior for ϕ which corresponds to the usual uniform prior for $\boldsymbol{\mu}$, known to lead to Stein's (1959) paradox. The expected reference posterior intrinsic loss may then be expressed in terms of the hypergeometric ${}_1F_1$ function as

$$d(\boldsymbol{\mu}^e, \overline{\boldsymbol{x}}) = \frac{n}{2}\,\mathrm{E}[\phi \mid \overline{\boldsymbol{x}}] = \frac{1}{2}\frac{{}_1F_1(3/2,\, k/2,\, nt/2)}{{}_1F_1(1/2,\, k/2,\, nt/2)} = d(nt, k),$$

which only depends on the data through

$$t = t(\boldsymbol{\mu}^e, \overline{\boldsymbol{x}}) = \|\boldsymbol{\mu}^e - \overline{\boldsymbol{x}}\|\,.$$

The expected intrinsic loss $d(nt, k)$ increases with nt for any dimension k and attains its minimum at $t = 0$, that is, when $\boldsymbol{\mu}^e = \overline{\boldsymbol{x}}$. The behavior of $d(nt, k)$ as a function of nt is shown in the left pane of Figure 4 for different values of k. It follows that, if the model is multivariate normal and there is *no further assumption on exchangeability* of the μ_j's, then the intrinsic estimator $\boldsymbol{\mu}^*$ is simply the sample mean $\overline{\boldsymbol{x}}$. The expected intrinsic loss of the Bayes estimator, namely $\boldsymbol{\mu}^* = \overline{\boldsymbol{x}}$, is

$$d(\boldsymbol{\mu}^*, \overline{\boldsymbol{x}}) = d(0, k) = \tfrac{1}{2}\,\cdot$$

Shrinking towards the overall mean $\boldsymbol{x}_0$, leading to ridge-type estimates of the general form

$$\tilde{\boldsymbol{\mu}}(\alpha) = \alpha\, \boldsymbol{x}_0 + (1 - \alpha)\overline{\boldsymbol{x}}, \qquad 0 < \alpha < 1,$$

will only increase the expected loss. Indeed,

$$d(\tilde{\boldsymbol{\mu}}(\alpha), \overline{\boldsymbol{x}}) = \frac{1}{2}\,\frac{{}_1F_1(3/2,\, k/2,\, nr_\alpha/2)}{{}_1F_1(1/2,\, k/2,\, nr_\alpha/2)}\,, \qquad r_\alpha = \frac{\alpha^2}{k}\sum_{i \neq j}(\mu_i - \mu_j)^2,$$

is an increasing function of nr_α and, hence, an increasing function of α. It follows that, with respect to the reference posterior, all ridge estimators have a *larger* expected loss than the sample mean. Similarly, the James–Stein estimator (James and Stein, 1961),

$$\tilde{\boldsymbol{\mu}}_{js} = (1 - (k-2)||\overline{\boldsymbol{x}}||^{-1})\overline{\boldsymbol{x}}, \quad k > 2,$$

which shrinks towards the origin rather than towards the overall mean, corresponds to $r_\alpha = 1$ and, hence, also has a larger expected loss than the sample mean. The expected intrinsic losses (quadratic in this case) of these estimators, for $k = 3$ and the particular random sample $\overline{\boldsymbol{x}} = \{0.72, -0.71, 1.67\}$, simulated from $\mathrm{N}_3(\overline{\boldsymbol{x}} \,|\, 0, \boldsymbol{I}_3)$, are compared in the right pane of Figure 4.

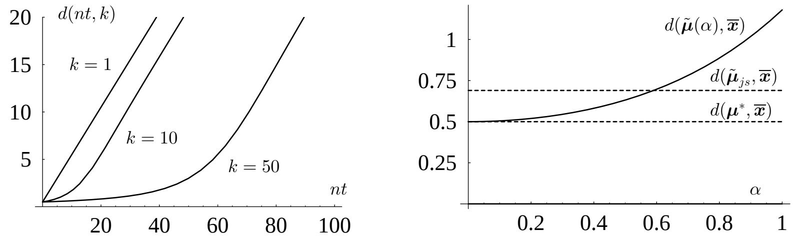

Figure 4. *Reference expected posterior losses in estimating a multivariate normal mean.*

The preceding analysis suggests that the frequent practice of shrinking towards either the origin or the overall mean may be inappropriate, unless there is information which justifies an exchangeability assumption for the μ_i's; in this case, a hierarchical model should be developed, and the intrinsic estimator will indeed be a ridge-type estimator. However, with a plain multivariate normal assumption, shrinking will only increase the (reference) expected loss. Thus, do not shrink without a good reason!

5. FINAL REMARKS

The intrinsic discrepancy, based on the theory of information, introduced in Bernardo and Rueda (2002) for densities which either have the same or nested supports, and further explored in this paper, has been shown to have many attractive properties. It is *symmetric*; it is *invariant*; it is typically finite for *non-regular problems*, and it is *calibrated* in natural information units. Indeed, the intrinsic divergence may be used to define a new type of *convergence* which is natural to consider in Bayesian statistics:

Definition 3. (*Intrinsic Convergence*). The sequence of probability densities $\{p_i\}_{i=1}^{\infty}$ converges *intrinsically* to the probability density p if, and only if, $\lim_{i\to\infty} \delta\{p_i, p\} = 0$.

Exploring the properties of this new definition of convergence will be the subject of future research. Further work is also needed to extend this definition to situations in which the densities are defined over arbitrary supports.

Intrinsic estimators were obtained by minimizing the reference posterior expectation of the intrinsic loss,

$$d(\boldsymbol{\theta}^e \,|\, \boldsymbol{x}) = \int_\Theta \delta(\boldsymbol{\theta}^e, \boldsymbol{\theta})\, \pi_\delta(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}.$$

Conditional on the assumed model, the positive statistic $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{x})$ is a natural measure of the *compatibility* of any $\boldsymbol{\theta}_0 \in \Theta$ with the observed data $\boldsymbol{x}$. Consequently, the intrinsic statistic

$d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{x})$ is a natural test statistic which finds immediate applications in *precise hypothesis testing*, leading to BRC, the Bayesian reference criterion (Bernardo, 1999; Bernardo and Rueda, 2002).

We have focused on the use of the intrinsic discrepancy in reference problems, where no prior information is assumed on the parameter values. However, because of its nice properties, the intrinsic discrepancy is an eminently reasonable loss function to consider in problems where prior information (possibly in the form of a hierarchical model) is, in fact, available.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the comments received from Professor Dennis Lindley and from an anonymous referee. The research of José M. Bernardo was partially funded with grants GV01-7 of the Generalitat Valenciana, and BMF2001-2889 of the Ministerio de Ciencia y Tecnología, Spain. The research of Miguel Juárez was supported by CONACyT, Mexico.

REFERENCES

Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Am. Statist. Ass.* **84**, 200–207.

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: Oxford University Press, 35–60 (with discussion).

Berger, J. O., Philippe, A. and Robert, C. P. (1998). Estimation of quadratic functions: Uninformative priors for non-centrality parameters. *Statist. Sinica* **8**, 359–376.

Bernardo, J. M. (1979a). Expected information as expected utility. *Ann. Statist.* **7**, 686–690.

Bernardo, J. M. (1979b). Reference posterior distributions for Bayesian inference. *J. R. 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. (1997). Non-informative priors do not exist. *J. Statist. Plann . Inference* **65**, 159–189 (with discussion).

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: Oxford University Press, 101–130 (with discussion).

Bernardo, J. M. and Ramón, J. M. (1998). An introduction to Bayesian reference analysis: Inference on the ratio of multinomial parameters. *J. R. Statist. Soc. D* **47**, 101–135.

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.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* **23**, 493–507.

Datta, G. S. and Ghosh, M. (1996). On the invariance of uninformative priors. *Ann. Statist.* **24**, 141–159.

Ferrándiz, J. R. (1985). Bayesian inference on Mahalanobis distance: an alternative approach to Bayesian model testing. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds), Amsterdam: North-Holland, 645–654.

Gibbs, J. W. (1902). *Elementary Principles of Statistical Mechanics*. Reprinted, 1981. Woodbridge, CT: Ox Bow Press.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. New York: Hafner Press.

Huzurbazar, V. S. (1955). Exact forms of some invariants for distributions admitting sufficient statistics. *Biometrika* **42**, 533–537.

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.

Jaynes, E. T. (1983). *Papers on Probability, Statistics and Statistical Physics* (R. D. Rosenkrantz, ed). Dordrecht: Reidel.

Jeffreys, H. (1948). *Theory of Probability* (3rd edn, 1961). Oxford: Oxford University Press.

Kullback, S. (1959). *Information Theory and Statistics*, 2nd edn, 1968. New York: Dover.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.

Robert, C. P. (1996). Intrinsic losses. *Theory and Decision* **40**, 191–214.

Savage, L. J. (1954). *The Foundations of Statistics.* New York: Dover.

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

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Tech. J.* **27**, 379–423 and 623–656. Reprinted in *The Mathematical Theory of Communication* (Shannon, C. E. and Weaver, W., 1949). Urbana: University of Illinois Press.

Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.* **30**, 877–880.

Stone, M. (1959). Application of a measure of information to the design and comparison of experiments. *Ann. Math. Statist.* **30**, 55–70.