

BAYESIAN STATISTICS 9
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, 2010

Integrated Objective Bayesian Estimation and Hypothesis Testing

JOSÉ M. BERNARDO
Universitat de València, Spain
jose.m.bernardo@uv.es

SUMMARY

The complete final product of Bayesian inference is the posterior distribution of the quantity of interest. Important inference summaries include point estimation, region estimation, and precise hypotheses testing. Those summaries may appropriately be described as the solution to specific decision problems which depend on the particular loss function chosen. The use of a continuous loss function leads to an integrated set of solutions where the same prior distribution may be used throughout. Objective Bayesian methods are those which use a prior distribution which only depends on the assumed model and the quantity of interest. As a consequence, objective Bayesian methods produce results which only depend on the assumed model and the data obtained. The combined use of the intrinsic discrepancy, an invariant information-based loss function, and appropriately defined reference priors, provides an integrated objective Bayesian solution to both estimation and hypothesis testing problems. The ideas are illustrated with a large collection of non-trivial examples.

Keywords and Phrases: FOUNDATIONS; DECISION THEORY; KULLBACK-LEIBLER DIVERGENCE; INTRINSIC DISCREPANCY; REFERENCE ANALYSIS; REFERENCE PRIORS; POINT ESTIMATION; INTERVAL ESTIMATION; REGION ESTIMATION; PRECISE HYPOTHESIS TESTING; HARDY-WEINBERG EQUILIBRIUM; CONTINGENCY TABLES.

1. INTRODUCTION

From a Bayesian viewpoint, the final outcome of any problem of inference is the posterior distribution of the vector of interest. Thus, given a probability model $\mathcal{M}_{\mathbf{z}} = \{p(\mathbf{z} | \boldsymbol{\omega}), \mathbf{z} \in \mathcal{Z}, \boldsymbol{\omega} \in \Omega\}$ which is assumed to describe the mechanism which has generated the available data \mathbf{z} , all that can be said about any function $\boldsymbol{\theta}(\boldsymbol{\omega}) \in \Theta$ of the parameter vector $\boldsymbol{\omega}$ is contained in its posterior distribution $p(\boldsymbol{\theta} | \mathbf{z})$. This is deduced from standard probability theory arguments via the posterior distribution $p(\boldsymbol{\omega} | \mathbf{z}) \propto p(\mathbf{z} | \boldsymbol{\omega})p(\boldsymbol{\omega})$ which is based on the assumed prior $p(\boldsymbol{\omega})$. To facilitate the assimilation of the inferential contents of $p(\boldsymbol{\theta} | \mathbf{z})$, one often tries to *summarize* the information contained in this posterior by (i) providing $\boldsymbol{\theta}$ values which, in the light of the data, are likely to be close to its true value (*estimation*) and by (ii)

measuring the compatibility of the data with hypothetical values $\theta_0 \in \Theta_0 \subset \Theta$ of the vector of interest which might have been suggested by the research context (*hypothesis testing*). One would expect that the *same* prior $p(\omega)$, whatever its basis, could be used to provide both types of summaries. However, since the pioneering book by Jeffreys (1961), Bayesian methods have often made use of two *radically different* types of priors, some for estimation and some for hypothesis testing. We argue that this is certainly not necessary, and probably not convenient, and describe a particular form of doing this within the framework of Bayesian decision theory. Many of the ideas described below have already appeared in the literature over the past few years. Thus, this is mainly an up-to-date review paper, which unifies notation, definitions and available results. However, it also contains some previously unpublished material.

Section 2 formalizes the decision theoretic formulation of point estimation, region estimation and precise hypothesis testing, and emphasizes that the results are highly dependent on the choices of both the loss function and the prior distribution. Section 3 reviews a set of desiderata for loss functions to be used in stylized non problem-specific theoretical inference, and defines the *intrinsic discrepancy*, an invariant information-based loss function, which is suggested for general use in those circumstances. Section 4 describes objective Bayesian methods as those using a prior distribution which only depends on the assumed model, and reviews some basic concepts behind *reference priors*, a particular form of objective prior functions which is proposed for general use. In multiparameter problems, reference priors are known to depend on the quantity of interest; a criterion is proposed to select joint priors which could safely be used for a set of different quantities of interest. In Section 5, the combined use of the intrinsic discrepancy and appropriately chosen reference priors is proposed as an integrated objective Bayesian solution to both estimation and hypothesis testing problems. The theory is illustrated via many examples.

2. BAYESIAN INFERENCE SUMMARIES

Let \mathbf{z} be the available data which are assumed to have been generated as one random observation from model $\mathcal{M}_{\mathbf{z}} = \{p(\mathbf{z} | \omega), \mathbf{z} \in \mathcal{Z}, \omega \in \Omega\}$. Often, but not always, data will consist of a random sample $\mathbf{z} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ from some distribution $q(\mathbf{x} | \omega)$, with $\mathbf{x} \in \mathcal{X}$; in this case $p(\mathbf{z} | \omega) = \prod_{i=1}^n q(\mathbf{x}_i | \omega)$ and $\mathcal{Z} = \mathcal{X}^n$. Let $\theta(\omega)$ be the vector of interest. Without loss of generality, the model may explicitly be expressed in terms of θ so that $\mathcal{M}_{\mathbf{z}} = \{p(\mathbf{z} | \theta, \lambda), \mathbf{z} \in \mathcal{Z}, \theta \in \Theta, \lambda \in \Lambda\}$, where λ is some appropriately chosen nuisance parameter vector. Let $p(\theta, \lambda) = p(\lambda | \theta) p(\theta)$ be the assumed prior, and let $p(\theta | \mathbf{x})$ be the corresponding marginal posterior distribution of θ . Appreciation of the inferential contents of $p(\theta | \mathbf{z})$ may be enhanced by providing both point and region estimates of the vector of interest θ , and by declaring whether or not some context suggested specific value θ_0 (or maybe a set of values Θ_0), is (are) compatible with the observed data \mathbf{z} . A large number of Bayesian estimation and hypothesis testing procedures have been proposed in the literature. We argue that their choice is better made in decision theoretical terms. Although it has been argued that the use of loss functions may not be directly relevant for inference problems, it is generally accepted that better inference procedures may often be obtained with the aid of decision-theoretic machinery; this is certainly our point of view.

Let $\ell\{\theta_0, (\theta, \lambda)\}$ describe, as a function of the (unknown) parameter values (θ, λ) which have generated the available data, *the loss* to be suffered if, working with

model $\mathcal{M}_{\mathbf{z}}$, the value θ_0 were used as a proxy for the unknown value of θ . As summarized below, point estimation, region estimation and hypothesis testing may all be appropriately described as specific decision problems using a common prior distribution and a common loss function. The results, which are obviously all conditional on the assumed model $\mathcal{M}_{\mathbf{z}}$, may dramatically depend on the particular choices made for both the prior and the loss functions but, given the available data \mathbf{z} , they all only depend on those through the corresponding posterior expected loss,

$$\bar{\ell}(\theta_0 | \mathbf{z}) = \int_{\Theta} \int_{\Lambda} \ell\{\theta_0, (\theta, \lambda)\} p(\theta, \lambda | \mathbf{z}) d\theta d\lambda.$$

As a function of $\theta_0 \in \Theta$, the expected loss $\bar{\ell}(\theta_0 | \mathbf{z})$ provides a direct measure of the relative unacceptability of all possible values of the quantity of interest in the light of the information provided by the data. As will later be illustrated, plotting $\bar{\ell}(\theta_0 | \mathbf{z})$ as a function of θ_0 when θ is one-dimensional, or producing a contour plot of $\bar{\ell}(\theta_0 | \mathbf{z})$ when θ is two-dimensional, may be a very useful addition to the conventional presentation of inferential results.

2.1. Point Estimation

To choose a point estimate for θ may be seen as a decision problem where the action space is the class Θ of all possible θ values. Foundations of decision theory dictate that the best estimator is that which minimizes the expected loss.

Definition 1 *The Bayes estimator $\theta^*(\mathbf{z}) = \arg \inf_{\theta_0 \in \Theta} \bar{\ell}(\theta_0 | \mathbf{z})$ is that which minimizes the posterior expected loss.*

Conventional examples of loss functions include the ubiquitous quadratic loss $\ell\{\theta_0, (\theta, \lambda)\} = (\theta_0 - \theta)^t(\theta_0 - \theta)$, which yields the posterior expectation as the Bayes estimator, and the zero-one loss on a neighborhood of the true value, which yields the posterior mode as a limiting result.

Bayes estimators are usually *not* invariant under one to one transformations. Thus, the Bayes estimator under quadratic loss of a variance (its posterior expectation), is not the square of the Bayes estimator of the standard deviation. This is rather difficult to explain when, as it is the case in theoretical inference, one merely wishes to report an estimate of some quantity of interest. Invariant Bayes estimators may easily be obtained by using invariant loss functions (see Section 3), rather than the conventional (non-invariant) loss functions mentioned above.

2.2. Region Estimation

Bayesian region estimation is easily achieved by quoting posterior credible regions. To choose a q -credible region for θ may be seen as a decision problem where the action space is the class of subsets of Θ with posterior probability q . Foundations dictate that the best region is that which contains those θ values with minimum expected loss.

Definition 2 *A Bayes q -credible region $\Theta_q^*(\mathbf{z}) \subset \Theta$ is a q -credible region where any value within the region has a smaller posterior expected loss than any value outside the region, so that $\forall \theta_i \in \Theta_q^*(\mathbf{z}), \forall \theta_j \notin \Theta_q^*(\mathbf{z}), \bar{\ell}(\theta_i | \mathbf{z}) \leq \bar{\ell}(\theta_j | \mathbf{z})$.*

The quadratic loss function yields credible regions which contain those values of θ closest to the posterior expectation in the Euclidean distance sense. A zero-one loss function leads to highest posterior density (HPD) credible regions. Again, Bayes credible regions are generally *not* invariant under one to one transformations. Thus, HPD regions in one parameterization will not transform to HPD regions in another. Invariant Bayes credible regions may however be obtained by using invariant loss functions. The concept of a Bayes credible region was introduced in Bernardo (2005b) under the name of *lower posterior loss* (LPL) credible regions; the paper, and its ensuing discussion, includes the analysis of many examples.

2.3. Precise Hypothesis Testing

Consider a value θ_0 of the vector of interest which deserves special consideration, either because assuming $\theta = \theta_0$ would noticeably simplify the model, or because there are additional context specific arguments suggesting that $\theta = \theta_0$. Intuitively, the value θ_0 should be judged to be *compatible* with the observed data \mathbf{z} if its posterior density $p(\theta_0 | \mathbf{z})$ is relatively high. However, a more precise form of conclusion is typically required.

Formally, testing the hypothesis $H_0 \equiv \{\theta = \theta_0\}$ may be described as a decision problem where the action space $\mathcal{A} = \{a_0, a_1\}$ contains only two elements: to accept (a_0) or to reject (a_1) the hypothesis under scrutiny. Foundations require to specify a loss function $\ell_h\{a_i, (\theta, \lambda)\}$ measuring the consequences of accepting or rejecting H_0 as a function of the actual parameter values. It is important to be explicit about what is precisely meant by accepting or rejecting H_0 . By assumption, a_0 means to *act as if* H_0 were true, that is to work with the model $\mathcal{M}_0 = \{p(\mathbf{z} | \theta_0, \lambda_0), \mathbf{z} \in \mathcal{Z}, \lambda_0 \in \Lambda\}$, while a_1 means to reject this simplification and to keep working with model $\mathcal{M}_z = \{p(\mathbf{z} | \theta, \lambda), \mathbf{z} \in \mathcal{Z}, \theta \in \Theta, \lambda \in \Lambda\}$. Alternatively, an already established model \mathcal{M}_0 may have been embedded into a more general model \mathcal{M}_z , constructed to include promising departures from $\theta = \theta_0$, and it is required to verify whether presently available data \mathbf{z} are still compatible with $\theta = \theta_0$, or whether the extension to $\theta \in \Theta$ is really necessary. Given the available data \mathbf{z} , the optimal action will be to reject the hypothesis considered if (and only if) the expected posterior loss of accepting (a_0) is larger than that of rejecting (a_1), so that

$$\int_{\Theta} \int_{\Lambda} [\ell_h\{a_0, (\theta, \lambda)\} - \ell_h\{a_1, (\theta, \lambda)\}] p(\theta, \lambda | \mathbf{z}) d\theta d\lambda > 0.$$

Hence, only the loss difference $\Delta\ell_h\{\theta_0, (\theta, \lambda)\} = \ell_h\{a_0, (\theta, \lambda)\} - \ell_h\{a_1, (\theta, \lambda)\}$, which measures the *advantage* of rejecting $H_0 \equiv \{\theta = \theta_0\}$ as a function of the parameter values, must be specified. The hypothesis H_0 should be rejected whenever the expected advantage of rejecting is positive. Without loss of generality, the function $\Delta\ell_h$ may be written in the form

$$\Delta\ell_h\{\theta_0, (\theta, \lambda)\} = \ell\{\theta_0, (\theta, \lambda)\} - \ell_0$$

where (precisely as in estimation), $\ell\{\theta_0, (\theta, \lambda)\}$ describes, as a function of the parameter values which have generated the data, the non-negative loss to be suffered if θ_0 were used as a proxy for θ . Since $\ell\{\theta_0, (\theta_0, \lambda)\} = 0$, so that $\Delta\ell_h\{\theta_0, (\theta_0, \lambda)\} = -\ell_0$, the constant $\ell_0 > 0$ describes (in the same loss units) the context-dependent non-negative advantage of accepting $\theta = \theta_0$ when it is true. With this formulation, the

optimal action is to reject $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ whenever the expected value of $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - \ell_0$ is positive, *i.e.*, whenever $\bar{\ell}(\boldsymbol{\theta}_0 | \mathbf{z})$, the posterior expectation of $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, is larger than ℓ_0 . Thus the solution to the hypothesis testing decision problem posed is found in terms of the *same* expected loss function that was needed for estimation.

Definition 3 *The Bayes test criterion to decide on the compatibility of $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ with available data \mathbf{z} is to reject $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ if (and only if), $\bar{\ell}(\boldsymbol{\theta}_0 | \mathbf{z}) > \ell_0$, where ℓ_0 is a context dependent positive constant.*

The compound case may be analyzed by separately considering each of the values which make part of the compound hypothesis to test. Thus, depending on the context, a compound hypothesis of the form $H_0 \equiv \{\boldsymbol{\theta}_0 \in \Theta_0\}$ could be accepted when *at least one* of its elements would be accepted, so that $\inf_{\boldsymbol{\theta}_0 \in \Theta_0} \bar{\ell}(\boldsymbol{\theta}_0 | \mathbf{z}) < \ell_0$, or when *all* its elements would be accepted, so that $\sup_{\boldsymbol{\theta}_0 \in \Theta_0} \bar{\ell}(\boldsymbol{\theta}_0 | \mathbf{z}) < \ell_0$.

Using the zero-one loss function, $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = 0$ if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = 1$ otherwise, so that the loss advantage of rejecting $\boldsymbol{\theta}_0$ is a constant whenever $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ and zero otherwise, leads to rejecting H_0 if (and only if) $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0 | \mathbf{z}) < p_0$ for some context-dependent p_0 . Notice that, using this particular loss function, if one is to avoid a systematic rejection of H_0 (whatever the data), the prior probability $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0)$ must be *strictly positive*. If $\boldsymbol{\theta}$ is a continuous parameter this forces the use of a non-regular “sharp” prior, concentrating a positive probability mass at $\boldsymbol{\theta}_0$. With no mention to the (rather naïve) loss structure which is implicit behind, this type of solution was early advocated by Jeffreys (1961). Notice however, that this formulation implies the use of radically different (and often polemic) priors for hypothesis testing than those used for estimation. Moreover, this formulation is also known to lead to the difficulties associated to Lindley’s paradox (Lindley, 1957; Bartlett, 1957; Robert, 1993). For an illustration of the possible consequences of Lindley’s paradox, see Example 7 in Section 5.

Using the quadratic loss function leads to rejecting a $\boldsymbol{\theta}_0$ value whenever its Euclidean distance to $E[\boldsymbol{\theta} | \mathbf{z}]$, the posterior expectation of $\boldsymbol{\theta}$, is sufficiently large. Observe that the use of continuous loss functions (such as the quadratic loss) permits the use in hypothesis testing of precisely the same priors that are used in estimation. In general, the Bayes test criterion is not invariant under one-to-one transformations. Thus, if $\boldsymbol{\phi}(\boldsymbol{\theta})$ is a one-to-one transformation of $\boldsymbol{\theta}$, rejecting $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ does not generally imply rejecting $\boldsymbol{\phi}(\boldsymbol{\theta}) = \boldsymbol{\phi}(\boldsymbol{\theta}_0)$. Once more, invariant Bayes test procedures are available by using invariant loss functions.

The threshold constant ℓ_0 , which is used to decide whether or not an expected loss is too large, is part of the specification of the decision problem, and should be context-dependent. However, as demonstrated below, a judicious choice of the loss function leads to calibrated expected losses, where the relevant threshold constant has an immediate, operational interpretation.

3. LOSS FUNCTIONS

The methods described above are completely general. Indeed, for a given loss function and a given prior distribution, they describe essentially unique procedures to perform both estimation and hypothesis testing; they are the only procedures which are compatible with the foundations-based decision-theoretic attitude which is at the heart of Bayesian methods. However, the results will be extremely dependent on the particular choices made of both the loss function and the prior distribution.

In this section the choice of the loss function is analyzed. Section 4 considers the choice of the prior.

Conditional on model $\mathcal{M}_{\mathbf{z}} = \{p(\mathbf{z} | \boldsymbol{\theta}, \boldsymbol{\lambda}), \mathbf{z} \in \mathcal{Z}, \boldsymbol{\theta} \in \Theta, \boldsymbol{\lambda} \in \Lambda\}$, the required loss function $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ should describe, in terms of the unknown parameter values $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ which have generated the available data, the loss to be suffered if, working with model $\mathcal{M}_{\mathbf{z}}$, the value $\boldsymbol{\theta}_0$ were used as a proxy for $\boldsymbol{\theta}$. It may naïvely appear that what is needed is just some measure of the discrepancy between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$. However, since all parameterizations are arbitrary, what is really required is some measure of the discrepancy between the *models* labelled by $\boldsymbol{\theta}$ and by $\boldsymbol{\theta}_0$. By construction, such a discrepancy measure will be independent of the particular parameterization used. Robert (1996) coined the word *intrinsic* to refer to those model-based loss functions. They are always invariant under one-to-one reparameterizations.

Any reasonable measure of the dissimilarity $\delta\{p_{\mathbf{z}}, q_{\mathbf{z}}\}$ between two probability densities $p(\mathbf{z})$ and $q(\mathbf{z})$ for a random vector $\mathbf{z} \in \mathcal{Z}$ should surely be non-negative, and zero if (and only if), $p(\mathbf{z}) = q(\mathbf{z})$ almost everywhere. Moreover it should be invariant under one-to-one transformations of \mathbf{z} ; indeed, if $\mathbf{y} = \mathbf{y}(\mathbf{z})$ is such a transformation and J is the appropriate Jacobian, $p_{\mathbf{y}} = p_{\mathbf{z}}/|J|$, and $q_{\mathbf{y}} = q_{\mathbf{z}}/|J|$ are expressions of precisely the same uncertainties and, therefore, one should certainly have $\delta\{p_{\mathbf{z}}, q_{\mathbf{z}}\} = \delta\{p_{\mathbf{y}}, q_{\mathbf{y}}\}$. To avoid undesirable asymmetries (see Example 2 below), one would also like δ to be a symmetric functional, so that $\delta\{p_{\mathbf{z}}, q_{\mathbf{z}}\} = \delta\{q_{\mathbf{z}}, p_{\mathbf{z}}\}$. Finally, it should also be possible to use δ to compare densities with strictly nested supports, since many approximations are precisely obtained by restricting the original support to some strict subspace.

3.1. The Intrinsic Loss Function

Not many divergence measures in functional analysis satisfy the desiderata mentioned above, but they are all satisfied by the *intrinsic discrepancy*, a divergence measure introduced in Bernardo and Rueda (2002), which has both an information theoretical justification, and a simple operational interpretation in terms of average log-density ratios.

Definition 4 The intrinsic discrepancy $\delta\{p_1, p_2\}$ between two probability distributions for the random vector \mathbf{z} with densities $p_1(\mathbf{z})$, $\mathbf{z} \in \mathcal{Z}_1$, and $p_2(\mathbf{z})$, $\mathbf{z} \in \mathcal{Z}_2$, is

$$\delta\{p_1, p_2\} = \min [\kappa\{p_1 | p_2\}, \kappa\{p_2 | p_1\}]$$

where $\kappa\{p_j | p_i\} = \int_{\mathcal{Z}_i} p_i(\mathbf{z}) \log[p_i(\mathbf{z})/p_j(\mathbf{z})] d\mathbf{z}$ is the Kullback-Leibler (KL) directed logarithmic divergence of p_j from p_i . The intrinsic discrepancy between a probability distribution p and a family of distributions $\mathcal{F} = \{q_i, i \in I\}$ is the intrinsic discrepancy between p and the closest of them,

$$\delta\{p, \mathcal{F}\} = \inf_{q \in \mathcal{F}} \delta\{p, q\}.$$

It is easily verified that $\delta\{p_1, p_2\}$ is non-negative and it is zero if (and only if), $p_1 = p_2$ almost everywhere. It is invariant under one-to-one transformations of \mathbf{z} , and it is obviously symmetric.

Notice that this definition allows for the possibility that one of the supports is strictly nested into the other one; if this is the case, one of the integrals diverges and the discrepancy is measured by the other. If both p_1 and p_2 have the same

support, the minimum is reached when integration is done with respect to the more concentrated density; indeed this may be used to *define* an order relation among probability distributions with the same support: p_1 is *more concentrated* than p_2 if $\kappa\{p_2 | p_1\} < \kappa\{p_1 | p_2\}$.

The intrinsic discrepancy $\delta\{p_1, p_2\}$ is the minimum expected log density ratio of one density over the other, and has an operative interpretation as the minimum amount of information (in natural information units or *nits*) expected to be required to discriminate between p_1 and p_2 . The intrinsic discrepancy may be used to define an appropriate loss function for all the decision problems considered in this paper.

The intrinsic loss is defined as the intrinsic discrepancy between the model, labelled by $(\boldsymbol{\theta}, \boldsymbol{\lambda})$, and the family \mathcal{M}_0 of models with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and arbitrary $\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}$.

Definition 5 Consider $\mathcal{M}_z = \{p(z | \boldsymbol{\theta}, \boldsymbol{\lambda}), z \in \mathcal{Z}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$. The intrinsic loss of using $\boldsymbol{\theta}_0$ as a proxy for $\boldsymbol{\theta}$ is the intrinsic discrepancy between the true model and the class of models with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\mathcal{M}_0 = \{p(z | \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0), z \in \mathcal{Z}, \boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}\}$,

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) | \mathcal{M}_z\} = \delta\{p_z(\cdot | \boldsymbol{\theta}, \boldsymbol{\lambda}), \mathcal{M}_0\} = \inf_{\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}} \delta\{p_z(\cdot | \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0), p_z(\cdot | \boldsymbol{\theta}, \boldsymbol{\lambda})\}.$$

Notice the complete generality of Definition 5; this may be used with either discrete or continuous data models (in the discrete case, the integrals in Definition 4 will obviously be sums), and with either discrete or continuous parameter spaces of any dimensionality.

The intrinsic loss has many attractive invariance properties. For any one-to-one reparameterization of the form $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ and $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta}, \boldsymbol{\lambda})$,

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) | \mathcal{M}_z\} = \ell_\delta\{\boldsymbol{\phi}_0, (\boldsymbol{\phi}, \boldsymbol{\psi}) | \mathcal{M}_z\},$$

so that the use of this loss function will lead to estimation and hypothesis testing procedures which are *invariant* under those transformations. Moreover, if $\boldsymbol{t} = \boldsymbol{t}(z)$ is a sufficient statistic for model \mathcal{M}_z , one may equivalently work with the marginal model $\mathcal{M}_t = \{p(\boldsymbol{t} | \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{t} \in \mathcal{T}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ since, in that case,

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) | \mathcal{M}_z\} = \ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) | \mathcal{M}_t\}.$$

Computations are often simplified by using the additive property of the intrinsic loss: if data consist of a random sample $\boldsymbol{z} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ from some underlying model \mathcal{M}_x , so that $\mathcal{Z} = \mathcal{X}^n$, and $p(\boldsymbol{z} | \boldsymbol{\theta}, \boldsymbol{\lambda}) = \prod_{i=1}^n p(\boldsymbol{x}_i | \boldsymbol{\theta}, \boldsymbol{\lambda})$, then

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) | \mathcal{M}_z\} = n \ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) | \mathcal{M}_x\}.$$

An interesting interpretation of the intrinsic loss follows directly from Definitions 4 and 5. Indeed, $\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) | \mathcal{M}_z\}$ is just the minimum log-likelihood ratio which may be expected under repeated sampling between the true model, identified by $(\boldsymbol{\theta}, \boldsymbol{\lambda})$, and the class of models which have $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Thus, *the intrinsic loss formalizes the use of the minimum average log-likelihood ratio under sampling as a general loss function*.

In particular, a suggested value $\boldsymbol{\theta}_0$ for the vector of interest should be judged to be incompatible with the observed data \boldsymbol{z} if $\bar{\ell}_\delta(\boldsymbol{\theta}_0 | \boldsymbol{z})$, the posterior expectation of $\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) | \mathcal{M}_z\}$, is larger than a suitably chosen constant ℓ_0 . For instance, if for some arbitrary k , $\ell_0 = \log[10^k]$, then $\boldsymbol{\theta}_0$ would be rejected whenever, given the observed data, the minimum sampling average likelihood ratio against $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, may be expected to be larger than about 10^k . Conventional choices for ℓ_0 are $\{\log 10, \log 100, \log 1000\} \approx \{2.3, 4.6, 6.9\}$.

Since the intrinsic divergence is also defined when the distributions to be compared have different supports, the intrinsic loss may easily deal with non-regular models:

Example 1 (Uniform model). Let $\mathbf{z} = \{x_1, \dots, x_n\}$ be a random sample of size n from a uniform distribution on $(0, \theta)$, so that $p(x|\theta) = \theta^{-1}$, if $0 < x < \theta$, and zero otherwise. Definition 5 immediately yields $\ell_\delta\{\boldsymbol{\theta}_0, \boldsymbol{\theta} | \mathcal{M}_\mathbf{z}\} = n \log(\theta_0/\theta)$, if $\theta_0 \geq \theta$, and $n \log(\theta/\theta_0)$ otherwise. The same answer is obtained using the sampling distribution of the sufficient statistic, $t = \max\{x_1, \dots, x_n\}$, the largest observation in the sample. Most known divergence functionals between distributions cannot deal with this simple example.

Under regularity conditions, the intrinsic loss has an alternative expression which is generally much simpler to compute (Juárez, 2004, Sec. 2.4):

Theorem 1 *If the support of $p(\mathbf{z} | \boldsymbol{\theta}, \boldsymbol{\lambda})$ is convex for all $(\boldsymbol{\theta}, \boldsymbol{\lambda})$, then the intrinsic loss may also be written as*

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) | \mathcal{M}_\mathbf{z}\} = \min \left[\inf_{\boldsymbol{\lambda}_0 \in \Lambda} \kappa\{\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0 | \boldsymbol{\theta}, \boldsymbol{\lambda}\}, \inf_{\boldsymbol{\lambda}_0 \in \Lambda} \kappa\{\boldsymbol{\theta}, \boldsymbol{\lambda} | \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0\} \right],$$

where $\kappa\{\boldsymbol{\theta}_j, \boldsymbol{\lambda}_j | \boldsymbol{\theta}_i, \boldsymbol{\lambda}_i\}$ is the KL-divergence of $p_{\mathbf{z}}(\cdot | \boldsymbol{\theta}_j, \boldsymbol{\lambda}_j)$ from $p_{\mathbf{z}}(\cdot | \boldsymbol{\theta}_i, \boldsymbol{\lambda}_i)$.

When there is no danger of confusion, $\mathcal{M}_\mathbf{z}$ will be dropped from the notation, and $\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) | \mathcal{M}_\mathbf{z}\}$ will simply be written $\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, but the dependence on the model of the intrinsic loss should always be kept in mind.

Example 2 (Univariate normal model). Consider a random sample $\mathbf{z} = \{x_1, \dots, x_n\}$ from a normal $N(x | \mu, \sigma)$ distribution, and suppose that μ is the quantity of interest. It may be verified that

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

If simultaneous inferences about μ and σ are required, the relevant intrinsic loss function is $\ell_\delta\{(\mu_0, \sigma_0), (\mu, \sigma)\} = \min[\kappa\{\mu, \sigma | \mu_0, \sigma_0\}, \kappa\{\mu_0, \sigma_0 | \mu, \sigma\}]$.

Suppose however that μ is the parameter of interest. Since $\inf_{\sigma_0 > 0} \kappa\{\mu_0, \sigma_0 | \mu, \sigma\} = (1/2) \log[1 + (\mu - \mu_0)^2/\sigma^2]$, and $\inf_{\sigma_0 > 0} \kappa\{\mu, \sigma | \mu_0, \sigma_0\} = (1/2)(\mu - \mu_0)^2/\sigma^2$, use of the fact that $x \geq \log(1 + x)$, Theorem 1, and the additive property of the intrinsic loss, yields

$$\ell_\delta\{\mu_0, (\mu, \sigma) | \mathcal{M}_\mathbf{z}\} = \frac{n}{2} \log \left[1 + \frac{(\mu - \mu_0)^2}{\sigma^2} \right] = \frac{n}{2} \log \left[1 + \frac{\theta^2}{n} \right],$$

a function of the standardized distance $\theta = (\mu - \mu_0)/(\sigma/\sqrt{n})$ between μ and μ_0 , which converges to $\theta^2/2$ as $n \rightarrow \infty$. It may be noticed that for $|\theta| \geq \sqrt{n}$ the intrinsic discrepancy loss is concave, showing an attractive (but not often seen) decreasing marginal loss.

Similarly, if the parameter of interest is σ (or, since the intrinsic loss is invariant, any one-to-one transformation of σ), one has $\inf_{\mu_0 > 0} \kappa\{\mu_0, \sigma_0 | \mu, \sigma\} = (1/2)g(\sigma^2/\sigma_0^2)$ and $\inf_{\mu_0 > 0} \kappa\{\mu, \sigma | \mu_0, \sigma_0\} = (1/2)g(\sigma_0^2/\sigma^2)$ where $g(x) = (t - 1) - \log t$, $t > 0$. Using the fact that $g(t) < g(1/t)$ if, and only if $t < 1$, now yields

$$\ell_\delta\{\sigma_0, (\mu, \sigma) | \mathcal{M}_\mathbf{z}\} = \ell_\delta\{\sigma_0, \sigma | \mathcal{M}_\mathbf{z}\} = \begin{cases} (n/2) [(\phi - 1) - \log \phi] & \text{if } \phi < 1 \\ (n/2) [(\phi^{-1} - 1) - \log \phi^{-1}] & \text{if } \phi > 1, \end{cases}$$

a function of the variance ratio $\phi = \sigma_0^2/\sigma^2$, which does not depend on μ .

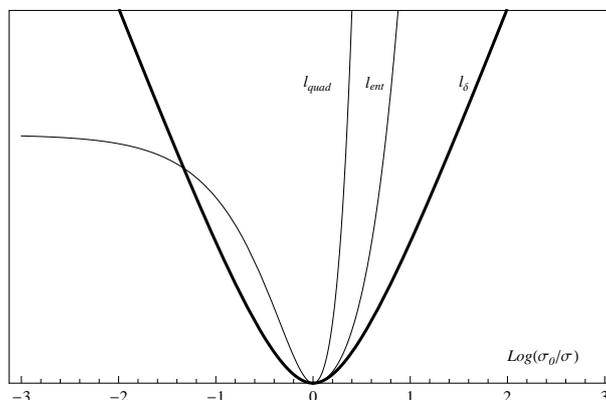


Figure 1: Invariant loss functions for estimating the variance of a normal model.

Figure 1 shows this intrinsic loss (for $n = 1$) as a function of $\log(\sigma_0/\sigma)$ (solid line), together with Stein entropy loss (James and Stein, 1961) $\ell_{ent}(\sigma_0, \sigma) = (1/2)g(\phi)$, and with the standardized quadratic loss, $\ell_{quad}(\sigma_0, \sigma) = (\phi - 1)^2$. It may be appreciated that both the entropy loss and the standardized quadratic loss penalize far more severely overestimation than underestimation, and therefore will lead to choosing too small estimates for the variance. For further details, see Bernardo (2006).

In the simple, important case of a multivariate normal model with known covariance matrix, the intrinsic loss is proportional to the Mahalanobis distance:

Example 3 (Multivariate normal model). Let $\mathbf{z} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be a random sample from a k -variate normal distribution $N(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with known covariance matrix $\boldsymbol{\Sigma}$. The KL divergence of $N(\mathbf{x} | \boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ from $N(\mathbf{x} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ is $\kappa\{\boldsymbol{\mu}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}\} = \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$. Since this is symmetric, and the intrinsic discrepancy is additive,

$$\delta\{\boldsymbol{\mu}_0, \boldsymbol{\mu} | \boldsymbol{\Sigma}\} = \frac{n}{2}(\boldsymbol{\mu}_0 - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}),$$

which is $n/2$ times the Mahalanobis distance between $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}$.

3.2. Approximations

Under regularity conditions, the result of Example 3 may be combined with conventional asymptotic results to obtain large sample approximations to intrinsic losses.

Theorem 2 Let data $\mathbf{z} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ consist of a random sample from $p(\mathbf{x} | \boldsymbol{\theta}, \boldsymbol{\lambda})$, let $F(\boldsymbol{\theta}, \boldsymbol{\lambda})$ be the corresponding Fisher matrix, and let $V(\boldsymbol{\theta}, \boldsymbol{\lambda}) = F^{-1}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ be its inverse. Then, for large n and under conditions for asymptotic normality,

$$\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) | \mathcal{M}_{\mathbf{z}}\} \approx \frac{n}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^t V_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}, \boldsymbol{\lambda})(\boldsymbol{\theta} - \boldsymbol{\theta}_0),$$

where $V_{\boldsymbol{\theta}\boldsymbol{\theta}}$ is the submatrix of $V(\boldsymbol{\theta}, \boldsymbol{\lambda})$ which corresponds to the vector of interest $\boldsymbol{\theta}$.

Proof. Under regularity conditions, the MLE's $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\lambda}})$ will be jointly sufficient and asymptotically normal with mean $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ and precision matrix $nF(\boldsymbol{\theta}, \boldsymbol{\lambda})$. Since the intrinsic discrepancy is invariant under reduction to sufficient statistics, the result in Example 3 yields

$$\delta\{(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0), (\boldsymbol{\theta}, \boldsymbol{\lambda}) | \mathcal{M}_{\mathbf{z}}\} \approx \frac{n}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0, \boldsymbol{\lambda} - \boldsymbol{\lambda}_0)^t F(\boldsymbol{\theta}, \boldsymbol{\lambda})(\boldsymbol{\theta} - \boldsymbol{\theta}_0, \boldsymbol{\lambda} - \boldsymbol{\lambda}_0).$$

Moreover, it may be verified (after some algebra) that, for fixed θ_0 and (θ, λ) , the KL-divergence $\delta\{(\theta_0, \lambda_0), (\theta, \lambda) | \mathcal{M}_z\}$ reaches its minimum, as a function of the nuisance vector λ_0 when, in terms of the corresponding submatrices of Fisher matrix $F(\theta, \lambda)$, λ_0 takes the value $\lambda + F_{\theta\lambda}F_{\lambda\lambda}^{-1}F_{\lambda\theta}(\theta - \theta_0)$. Substitution then yields

$$\ell\{\theta_0, (\theta, \lambda)\} = \inf_{\lambda_0 \in \Lambda_0} \delta\{(\theta_0, \lambda_0), (\theta, \lambda) | \mathcal{M}_z\} \approx \frac{n}{2}(\theta - \theta_0)^t V_{\theta\theta}^{-1}(\theta, \lambda)(\theta - \theta_0). \quad \square$$

The invariance of the intrinsic loss under reparameterization may be exploited to improve the approximation above, by simply choosing a parameterization where the asymptotic convergence to normality is faster. The following result (Bernardo, 2005b) is a one-parameter example of this technique, which makes use of the variance stabilization transformation.

Theorem 3 *Let $\mathbf{z} = \{x_1, \dots, x_n\}$ be a random sample of size n from model $p(x | \theta)$, and let $\tilde{\theta}_n = \tilde{\theta}_n(\mathbf{z})$ be an asymptotically sufficient consistent estimator of θ , whose sampling distribution is asymptotically normal with standard deviation $s(\theta)/\sqrt{n}$. Define $\phi(\theta) = \int^\theta s(y)^{-1} dy$. Then, for large values of n ,*

$$\ell\{\theta_0, \theta | \mathcal{M}_z\} \approx (n/2)[\phi(\theta_0) - \phi(\theta)]^2.$$

4. OBJECTIVE BAYESIAN METHODS

The decision-theoretic procedures described in Section 2 to derive Bayesian inference summaries are totally general, so that they may be used with any loss function and any prior distribution. We have argued above for the advantages of using the intrinsic loss function: it is invariant under both reparameterization and reduction to sufficient statistics, it easily deals with the presence of nuisance parameters, it may be used with non regular models, and it has a simple operational interpretation in terms of average log-likelihood ratios. The choice of the prior is considered now.

Foundations indicate that the prior distribution should describe available prior knowledge. In many situations however, either the available prior information on the quantity of interest is too vague or too complex to warrant the effort required to formalize it, or it is too subjective to be useful in scientific communication. An “objective” procedure is therefore often required, where the prior function is intended to describe a situation where there is no relevant information about the quantity of interest. Objectivity is an emotionally charged word, and it should be explicitly qualified whenever it is used. No statistical analysis is really objective, since both the experimental design and the model assumed have very strong subjective inputs. However, conventional procedures are often branded as “objective” just because their conclusions are only conditional on the model assumed and the data obtained. Bayesian methods where the prior function is directly derived from the assumed model are objective in this limited, but precise sense. For lively discussions of this, and related issues, see Bernardo (1997), Berger (2006), and ensuing discussions.

4.1. Development of Objective Priors

There is a vast literature devoted to the formulation of objective priors; relevant pointers are included in Bernardo and Smith (1994, Sec. 5.6), Kass and Wasserman (1996), Datta and Mukerjee (2004), Bernardo (2005a), Berger (2006), Ghosh,

Delampady and Samanta (2006), and references therein. Reference analysis, introduced by Bernardo (1979) and further developed by Berger and Bernardo (1989, 1992a,b,c), Sun and Berger (1998) and Berger, Bernardo and Sun (2009, 2011a,b), has been one of the most popular approaches for developing objective priors.

We will not repeat here arguments for reference analysis, but it may be worth emphasizing some basic issues and briefly reviewing some recent developments.

We first note that the same mathematical concepts which lie behind the definition of the intrinsic discrepancy provide an intuitive basis for the definition of reference priors. Indeed, for the one parameter model $\mathcal{M} = \{p(\mathbf{z} | \theta), \mathbf{z} \in \mathcal{Z}, \theta \in \Theta\}$, the intrinsic discrepancy $I\{p_\theta | \mathcal{M}\} = \delta\{p(\mathbf{z}, \theta), p(\mathbf{z})p(\theta)\}$ between the joint prior $p(\mathbf{z}, \theta)$ and the product of their marginals $p(\mathbf{z})p(\theta)$ is a functional of the prior $p(\theta)$ which measures the association between the data and the parameter and hence, the amount of information that, given prior $p(\theta)$, data \mathbf{z} may be expected to provide about θ . If one considers k independent observations from \mathcal{M} then, as k increases, $I\{p_\theta | \mathcal{M}^k\}$ will approach the *missing information* about θ which repeated sampling from \mathcal{M} could provide. If $\pi_k(\theta)$ denotes the prior which maximizes $I\{p_\theta | \mathcal{M}^k\}$, the sequence $\{\pi_k(\theta)\}_{k=1}^\infty$ will converge to that prior function which maximizes the missing information about θ , and this is defined to be the reference prior $\pi(\theta | \mathcal{M})$.

Theorem 4 Let $\mathbf{z}^{(k)} = \{\mathbf{z}_1, \dots, \mathbf{z}_k\}$ denote k conditionally independent observations from $\mathcal{M}_\mathbf{z}$. Then, for sufficiently large k

$$\pi_k(\theta) \propto \exp \left\{ \mathbb{E}_{\mathbf{z}^{(k)} | \theta} [\log p_h(\theta | \mathbf{z}^{(k)})] \right\}$$

where $p_h(\theta | \mathbf{z}^{(k)}) \propto \prod_{i=1}^k p(\mathbf{z}_i | \theta) h(\theta)$ is the posterior which corresponds to any arbitrarily chosen prior function $h(\theta)$ which makes the posterior proper for any $\mathbf{z}^{(k)}$.

For precise conditions and a formal proof of this very general result see Berger, Bernardo and Sun (2009).

Consider a set $\mathbf{z} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of n values $\mathbf{x}_i \in \mathcal{X}$; for any real valued function g with dominion \mathcal{X} the g -average of \mathbf{z} is defined to be $g^{-1}\{n^{-1} \sum_{i=1}^n g(x_i)\}$. For instance, the harmonic mean is the g -average with corresponds to the reciprocal function $g(x) = 1/x$. Theorem 4 implies that the reference prior at a particular point θ is proportional to the *logarithmic average* of the posterior density which this point would have under repeated sampling, if this θ value were the true parameter value. The parameter values which could be expected to get relatively large asymptotic posterior densities if they were true, will then precisely be those with relatively large reference prior densities.

The result in Theorem 4 makes very simple the numerical derivation of a one-parameter reference prior. One first chooses some formal prior $h(\theta)$, maybe one for which exact or approximate posterior computation is easy, and a relatively large number of replications k . For each particular θ value whose reference prior is desired, one generates a collection $\{\mathbf{z}_1^{(k)}, \dots, \mathbf{z}_s^{(k)}\}$ of s replications $\mathbf{z}_i^{(k)} = \{\mathbf{z}_{i1}, \dots, \mathbf{z}_{ik}\}$ of size k from the original model $p(\mathbf{z} | \theta)$, computes the corresponding s posterior densities at θ , $\{p_h(\theta | \mathbf{z}_j^{(k)})\}_{j=1}^s$, and approximate the reference prior at this point by its logarithmic average,

$$\pi(\theta) \approx \exp \left\{ \frac{1}{s} \sum_{j=1}^s \log p_h(\theta | \mathbf{z}_j^{(k)}) \right\}.$$

Under regularity conditions explicit formulae for the reference priors are readily available. In particular, if the posterior distribution of θ given a random sample of size n from $p(\mathbf{x}|\theta)$ is asymptotically normal with standard deviation $s(\hat{\theta}_n)/\sqrt{n}$, where $\hat{\theta}_n$ is a consistent estimator of θ , then the reference prior is $\pi(\theta) = s(\theta)^{-1}$. This includes Jeffreys-Perks prior (Jeffreys, 1946; Perks, 1947)

$$\pi(\theta) \propto i(\theta)^{1/2}, \quad i(\theta) = E_{\mathbf{x}|\theta}[-\partial^2 \log p(\mathbf{z}|\theta)/\partial\theta^2],$$

as a particular case. Similarly, if $p(\mathbf{x}|\theta)$ is a non regular model with support $S(\theta) = \{x; a_1(\theta) < x < a_2(\theta)\}$, where the $a_i(\theta)$'s are monotone functions of θ and $S(\theta)$ is either increasing or decreasing then, under regularity conditions (Ghosal and Samanta, 1997), the reference prior is

$$\pi(\theta) \propto E_{\mathbf{x}|\theta}[|\partial \log p(\mathbf{z}|\theta)/\partial\theta|].$$

In multiparameter problems, reference priors depend of the quantity of interest, a necessary feature in the construction of objective priors, if one is to prevent unacceptable behaviour in the posterior, such as marginalization paradoxes (Dawid, Stone and Zidek, 1973) or strong inconsistencies (Stone, 1976).

The joint reference prior is derived sequentially. Thus, if the model is $p(\mathbf{z}|\theta, \lambda)$ and θ is the quantity of interest, one works conditionally on θ and uses the one-parameter algorithm to derive the *conditional reference prior* $\pi(\lambda|\theta)$. If this is proper, it is used to obtain the *integrated model* $p(\mathbf{z}|\theta) = \int_{\Lambda} p(\mathbf{z}|\theta, \lambda) \pi(\lambda|\theta) d\lambda$, to which the one-parameter algorithm is applied again to obtain the *marginal reference prior* $\pi(\theta)$. The *joint reference prior* to compute the reference posterior for θ is then defined to be $\pi(\lambda|\theta) \pi(\theta)$. If $\pi(\lambda|\theta)$ is not proper, one proceeds similarly within a compact approximation to the parameter space (where all reference priors will be proper) and then derive the corresponding limiting result.

In general, reference priors are sequentially derived with respect to an ordered parameterization. Thus, given a model $\mathcal{M}_{\mathbf{z}} = \{p(\mathbf{z}|\boldsymbol{\omega}), \mathbf{z} \in \mathcal{Z}, \boldsymbol{\omega} \in \Omega\}$ with m parameters, the reference prior with respect to a particular ordered parameterization $\boldsymbol{\phi}(\boldsymbol{\omega}) = \{\phi_1, \dots, \phi_m\}$ (where the ϕ_i 's are ordered by inferential importance) is sequentially obtained as $\pi(\boldsymbol{\phi}) = \pi(\phi_m | \phi_{m-1}, \dots, \phi_1) \times \dots \times \pi(\phi_2 | \phi_1) \pi(\phi_1)$. Unless all reference priors turn out to be proper, the model must be endowed with an appropriate compact approximation to the parameter space $\{\Omega_j\}_{j=1}^{\infty} \subset \Omega$, which should remain the same for all reference priors obtained within the same model. Berger and Bernardo (1992c) describe the relevant algorithm for regular multiparameter models where asymptotic normality may be established. In typical applications, $\theta = \phi_1$ will be the quantity of interest, and the joint reference prior $\pi(\boldsymbol{\phi})$, which is often denoted $\pi_{\theta}(\boldsymbol{\phi})$ to emphasize the role of θ , is a just a technical device to produce the desired one-dimensional marginal reference posterior $\pi(\theta|\mathbf{z})$ of the quantity of interest.

4.2. Approximate Reference Priors

There are many situations where one may be simultaneously interested in all the parameters of the model or, more realistically, in several functions of them. Given a model $\mathcal{M}_{\mathbf{z}} = \{p(\mathbf{z}|\boldsymbol{\omega}), \mathbf{z} \in \mathcal{Z}, \boldsymbol{\omega} \in \Omega \subset \mathbb{R}^m\}$ with m parameters, consider a set $\boldsymbol{\theta}(\boldsymbol{\omega}) = \{\theta_1(\boldsymbol{\omega}), \dots, \theta_r(\boldsymbol{\omega})\}$ of $r > 1$ functions of interest, where r may be larger, smaller or equal to the number of parameters m . Berger, Bernardo and Sun (2011b) suggest a procedure to select a joint prior $\pi_{\boldsymbol{\theta}}(\boldsymbol{\omega})$ whose corresponding marginal

posteriors $\{\pi_{\theta}(\theta_i | \mathbf{z})\}_{i=1}^r$ could be expected to be close, for all possible data sets $\mathbf{z} \in \mathcal{Z}$, to the set of reference posteriors $\{\pi(\theta_i | \mathbf{z})\}_{i=1}^r$ yielded by the set of reference priors $\{\pi_{\theta_i}(\boldsymbol{\omega})\}_{i=1}^r$ which may be derived under the assumption that each of the θ_i 's is of interest.

If one is able to find a single joint prior $\pi_{\theta}(\boldsymbol{\omega})$ whose corresponding marginal posteriors are precisely equal to the reference posteriors for each of the θ_i 's so that, for all \mathbf{z} values, $\pi_{\theta}(\theta_i | \mathbf{z}) = \pi(\theta_i | \mathbf{z})$, then it is natural to argue that this should be a solution. Notice, however, that there may be many other priors which satisfy this condition. If the joint reference priors for the θ_i are all equal, then $\pi_{\theta}(\boldsymbol{\omega}) = \pi_{\theta_i}(\boldsymbol{\omega})$ will obviously satisfy the required condition, and it will be argued that this is *the* solution to the problem posed. Notice that this apparently naïve suggestion may have far reaching consequences. For instance, in the univariate normal model, this implies that $\pi(\mu, \sigma) = \sigma^{-1}$, which is the reference prior when either μ or σ are of interest, should also be used to make joint inferences for (μ, σ) , or to obtain a reference predictive distribution.

Since one will not generally be able to find a single joint prior $\pi_{\theta}(\boldsymbol{\omega})$ which would yield marginal posteriors for each of the θ_i 's which are all equal to the corresponding reference posteriors, an approximate solution must be found. This is easily implemented using intrinsic discrepancies:

Definition 6 Consider $\mathcal{M}_{\mathbf{z}} = \{p(\mathbf{z} | \boldsymbol{\omega}), \mathbf{z} \in \mathcal{Z}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ and let $\{\theta_1(\boldsymbol{\omega}), \dots, \theta_r(\boldsymbol{\omega})\}$ be $r > 1$ functions of interest. Let $\{\pi_{\theta_i}(\boldsymbol{\omega})\}_{i=1}^r$ be the relevant reference priors, and let $\{\pi_{\theta_i}(\mathbf{z})\}_{i=1}^r$ and $\{\pi(\theta_i | \mathbf{z})\}_{i=1}^r$ respectively be the corresponding prior predictives and reference posteriors. Let $\mathcal{F} = \{\pi(\boldsymbol{\omega} | \mathbf{a}), \mathbf{a} \in \mathcal{A}\}$ be a family of prior functions. For each $\boldsymbol{\omega} \in \boldsymbol{\Omega}$, the best approximate joint reference prior within \mathcal{F} is that which minimizes the average expected intrinsic loss

$$d(\mathbf{a}) = \frac{1}{r} \sum_{i=1}^r \int_{\mathcal{Z}} \delta\{\pi_{\theta_i}(\cdot | \mathbf{z}), p_{\theta_i}(\cdot | \mathbf{z}, \mathbf{a})\} \pi_{\theta_i}(\mathbf{z}) d\mathbf{z}, \quad \mathbf{a} \in \mathcal{A},$$

where $p(\theta_i | \mathbf{z}, \mathbf{a})$ is the marginal posterior of θ_i which corresponds to $\pi(\boldsymbol{\omega} | \mathbf{a})$.

The idea behind Definition 6 is to select some mathematically tractable family of prior distributions for $\boldsymbol{\omega}$, and to choose that element within the family which minimizes the average expected intrinsic discrepancy between the marginal posteriors for the θ_i 's obtained from that prior and the corresponding reference posteriors.

Example 4 (Multinomial model). Consider a multinomial model with m categories and parameters $\{\theta_1, \dots, \theta_{m-1}\}$, define $\theta_m = 1 - \sum_{i=1}^{m-1} \theta_i$, and suppose that the functions of interest are the m probabilities $\{\theta_1, \dots, \theta_m\}$. Let $\mathbf{z} = \{n_1, \dots, n_m\}$ be the results observed from a random sample of size n . Berger and Bernardo (1992a) show that the reference prior for θ_i depends on i , and that the reference posterior of θ_i is the beta distribution $\pi(\theta_i | \mathbf{z}) = \pi(\theta_i | n_i, n) = \text{Be}(\theta_i | n_i + 1/2, n - n_i + 1/2)$, which, as one would hope, only depends on the number of observations n_i which fall in category i and on the total number n of observations (therefore avoiding the partition paradox which occurs when the posterior for θ_i depends on the total number m of categories considered). Consider the family of (proper) Dirichlet priors of the form $p(\boldsymbol{\theta} | a) \propto \prod_{i=1}^m \theta_i^{a-1}$, with $a > 0$. The corresponding marginal posterior distribution of θ_i is $\text{Be}(\theta_i | n_i + a, n - n_i + (m-1)a)$ (notice the dependence on the number m of categories). The intrinsic discrepancy between this distribution and the corresponding reference prior is $\delta_i\{a | n_i, m, n\} = \delta_{\beta}\{n_i + 1/2, n - n_i + 1/2, n_i + a, n - n_i + (m-1)a\}$, where $\delta_{\beta}\{\alpha_1, \beta_1, \alpha_2, \beta_2\} = \min[\kappa_{\beta}\{\alpha_2, \beta_2 | \alpha_1, \beta_1\}, \kappa_{\beta}\{\alpha_1, \beta_1 | \alpha_2, \beta_2\}]$ and κ_{β} is the KL divergence between two beta densities with parameters (α_1, β_1) and (α_2, β_2) , given by

$$\begin{aligned} \kappa_{\beta}\{\alpha_2, \beta_2 \mid \alpha_1, \beta_1\} &= \log \left[\frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_2 + \beta_2)} \frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)} \frac{\Gamma(\beta_2)}{\Gamma(\beta_1)} \right] \\ &+ (\alpha_1 - \alpha_2) \psi(\alpha_1) + (\beta_1 - \beta_2) \psi(\beta_1) - ((\alpha_1 + \beta_1) - (\alpha_2 + \beta_2)) \psi(\alpha_1 + \beta_1), \end{aligned}$$

where $\psi(\cdot) = d \log[\Gamma(x)]/dx$ is the digamma function. The discrepancy $\delta_i\{a \mid n_i, m, n\}$ between the two posteriors of θ_i depends on the data only through n_i and n , and the corresponding reference predictive for n_i is

$$\pi(n_i \mid n) = \int_0^1 \text{Bi}(n_i \mid n, \theta_i) \text{Be}(\theta_i \mid 1/2, 1/2) d\theta_i = \frac{1}{\pi} \frac{\Gamma(n_i + 1/2) \Gamma(n - n_i + 1/2)}{\Gamma(n_i + 1) \Gamma(n - n_i + 1)}.$$

Hence, using Definition 6, the average expected intrinsic loss of using a joint Dirichlet prior with parameter a with a sample of size n is $d(a \mid m, n) = \sum_{n_i=0}^n \delta\{a \mid n_i, m, n\} \pi(n_i \mid n)$ since, by the symmetry of the problem, the m parameters $\{\theta_1, \dots, \theta_m\}$ yield all the same expected loss.

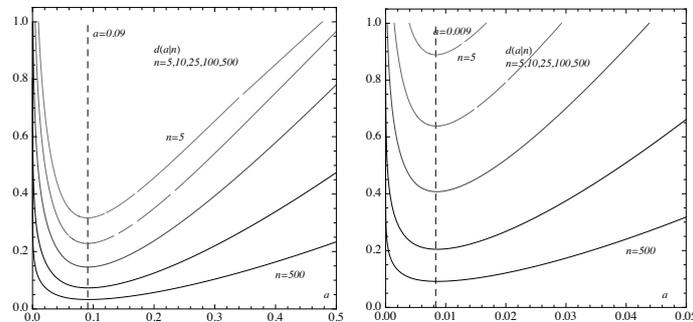


Figure 2: Expected intrinsic losses, of using a Dirichlet prior with parameter a in a multinomial model with m cells, for sample sizes 5, 10, 25, 100 and 500. Left panel, $m = 10$; right panel, $m = 100$. In both cases, the optimal value for all sample sizes is $a^* \approx 1/m$.

The function $d(a \mid m = 10, n)$ is represented in the left panel of Figure 2 for several values of n . The expected loss decreases with n and, for any n , the function $d(a \mid m, n)$ is concave, with a unique minimum numerically found to be at $a^* \approx 1/m$. Similarly, the function $d(a \mid m = 100, n)$ is represented in the right panel of Figure 2 for the same values of n and with the same vertical scale, yielding qualitatively similar results although, as one may expect, the expected losses are now larger than those obtained with $m = 10$ for the same sample size. Once more, the function $d(a \mid m, n)$ is concave, with a unique minimum numerically found to be at $a^* \approx 1/m$.

It follows that, for practical purposes, the best global Dirichlet prior when one is interested in all the cells of a multinomial model (and therefore in all the cells of a contingency table) is that with parameter $a = 1/m$, yielding an approximate marginal reference posterior $\text{Be}(\theta_i \mid n_i + 1/m, n - n_i + (m - 1)/m)$, with mean $(n_i + 1/m)/(n + 1)$. This is an important result for an objective Bayesian analysis of sparse frequency and contingency tables.

5. INTEGRATED REFERENCE ANALYSIS

With the loss function chosen to be the intrinsic loss, all that is required to implement the construction of the Bayesian reference summaries described in Section 2 is to specify a prior distribution. It will not come as a surprise that we recommend the use of a reference prior. The corresponding Bayes point estimators, Bayes credible

regions and Bayes test criteria will respectively be referred to as *reference intrinsic* estimators, credible regions or test criteria. The basic ideas were respectively introduced in Bernardo and Juárez (2003), Bernardo (2005), and Bernardo and Rueda (2002). All inference summaries depend on the data only through the expected reference intrinsic loss, $d(\theta_0 | \mathbf{z})$, the expectation of intrinsic loss with respect to the appropriate joint reference posterior

$$d(\theta_0 | \mathbf{z}) = \int_{\Theta} \int_{\Lambda} \ell_{\delta}\{\theta_0, (\theta, \lambda) | \mathcal{M}_{\mathbf{z}}\} \pi(\theta, \lambda | \mathbf{z}) d\theta d\lambda.$$

In one parameter problems, the reference prior is unique and the solution is therefore conceptually immediate. The following example is intended to illustrate the general procedure:

Example 5 (Uniform model, continued). Consider again the uniform model on $(0, \theta)$ of Example 1, where the intrinsic loss was found to be $\ell_{\delta}\{\theta_0, \theta | \mathcal{M}_{\mathbf{z}}\} = n \log(\theta_0/\theta)$, if $\theta_0 \geq \theta$, and $n \log(\theta/\theta_0)$ otherwise. The reference prior for this model is $\pi(\theta) = \theta^{-1}$. This leads to the Pareto reference posterior $\pi(\theta | \mathbf{z}) = \pi(\theta | t, n) = n t^n \theta^{-(n+1)}$ with support on (t, ∞) , where $t = \max\{x_1, \dots, x_n\}$ is a sufficient statistic. The q -posterior quantile is $\theta_q = t(1-q)^{-1/n}$; in particular the reference posterior median is $t 2^{1/n}$. Changing variables in $\pi(\theta | t, n)$, the *posterior* distribution of $(t/\theta)^n$ as a function of θ is found to be uniform on $(0, 1)$; on the other hand the sampling distribution of t is the inverted Pareto $p(t | \theta, n) = n t^{n-1} \theta^{-n}$ with support on $(0, \theta)$ and, therefore, the *sampling* distribution of $(t/\theta)^n$ as a function of t is also uniform on $(0, 1)$. Thus, the reference posterior has exact probability matching: all reference posterior q -credible intervals are also frequentist confidence intervals of level q .

The reference posterior expected intrinsic loss of using θ_0 as a proxy for θ (with $\theta_0 > t$ since, given the data, θ is known to be larger than t) is $\bar{\ell}_{\delta}(\theta_0 | t, n) = 2r - \log r - 1$, where $r = (t/\theta_0)^n$. This is a positive convex function of r with support on $(0, 1)$ which tends to ∞ as $r \rightarrow 0$, has unique minimum at $r = 1/2$ and takes the value 1 at $r = 1$. As a function of θ_0 , this is minimized at $\theta^* = t 2^{1/n}$, which is therefore the reference intrinsic estimator (and, as mentioned above, it is also the reference posterior median).

An intrinsic q -credible region will consist of the set of θ values with lower expected loss which have a posterior probability equal to q . It follows from the shape of $\bar{\ell}_{\delta}(\theta_0 | t, n)$ that, for sufficiently large q , these regions will be of the form $R_q = (t, \theta_q)$, where $\theta_q = t(1-q)^{-1/n}$ is the q -quantile of $\pi(\theta | t, n)$.

It may easily be shown that the sampling distribution of $r = (t/\theta_0)^n$ is uniform in $(0, (\theta/\theta_0)^n)$ and therefore, the expected value of $\bar{\ell}_{\delta}(\theta_0 | t, n)$ under repeated sampling is

$$E[\bar{\ell}_{\delta}(\theta_0 | t, n) | \theta] = (\theta/\theta_0)^n - n \log(\theta/\theta_0),$$

which is precisely equal to one if $\theta = \theta_0$, and increases with n otherwise. Hence, under repeated sampling, one would expect to obtain $\bar{\ell}_{\delta}$ values around 1 when $\theta = \theta_0$, and one would always reject a false θ_0 value for sufficiently large n . The procedure is therefore consistent.

A particular θ_0 value should be judged to be incompatible with the observed data (t, n) if $\bar{\ell}_{\delta}(\theta_0 | t, n) > \ell_0$, for suitably chosen ℓ_0 . This precisely means that, given available data, the minimum expected value under sampling of the log-likelihood ratio between the true model and the model identified by θ_0 may be expected to be larger than ℓ_0 . Thus, if $\ell_0 = \log[1000] \approx 6.9$, then θ_0 would be rejected whenever, given (t, n) , the average likelihood ratio against $\theta = \theta_0$ may be expected to be larger than about 1000.

To illustrate the type of results obtained, a sample of size $n = 10$ was simulated from a uniform distribution on $(0, 2)$, and this had a maximum value $t = 1.71$. The corresponding

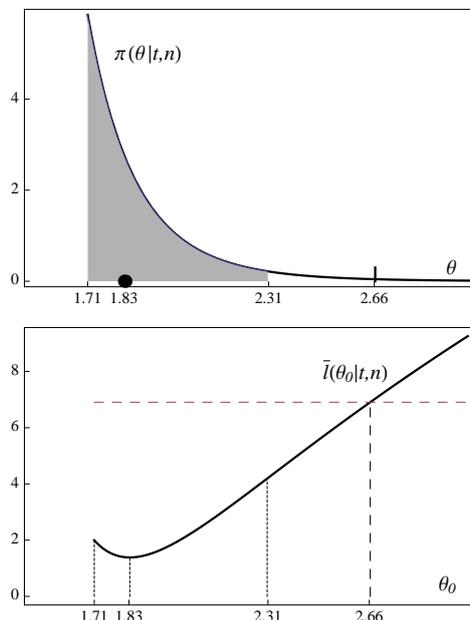


Figure 3: *Posterior reference analysis of the parameter of a uniform model.*

reference posterior density is plotted in the upper panel of Figure 3. The figure includes the intrinsic estimator $\theta^* = 1.83$ (indicated by a solid dot) and the intrinsic 0.95-credible region $(t, 2.31)$ (indicated as a shaded area). The expected intrinsic loss $\bar{\ell}_\delta(\theta_0 | t, n)$ is plotted in lower panel of Figure 3. It may be appreciated that the intrinsic estimator corresponds to the minimum value of the expected loss, and that all values within the intrinsic credible region have smaller expected loss than all values outside the region. The dashed horizontal line corresponds to $\ell_0 = \log 1000$, and this intersects the expected loss function at $\theta_0 = 2.66$. Thus, if in this application one wants to reject any value θ_0 with an expected average log-likelihood ratio against it larger than $\log 1000$, one should reject whenever $\theta_0 > 2.66$.

Figure 3 provides a rather complete, intuitive, easily interpretable, impressionist summary of the posterior reference analysis of the problem under scrutiny. Indeed, we argue that systematic use of this type of representation for any one-dimensional quantity of interest would greatly enhance the comprehension by the user of the inferential conclusions which, given the assumed model, could reasonably be drawn from the data.

Inference on the parameters of a univariate normal model is surely one of the oldest problems in mathematical statistics and yet, there is no consensus about its more appropriate solution. We review below the intrinsic reference results for this problem. Further details may be found in Bernardo (2005b, 2007).

Example 6 (Normal model, continued). Let \mathbf{z} be a random sample of size n from a normal $N(x | \mu, \sigma)$ and let (\bar{x}, s) be the jointly sufficient MLE estimators of its parameters. The reference prior when either μ or σ are the parameters of interest is $\pi(\mu, \sigma) = \sigma^{-1}$, and the corresponding marginal posteriors are the Student $\pi(\mu | \mathbf{z}) = \text{St}(\mu | \bar{x}, s/\sqrt{n-1}, n-1)$

and the squared root inverted gamma density $\pi(\sigma | \mathbf{z}) = \text{Ga}^{-1/2}(\sigma | (n-1)/2, ns^2/2)$ (so that the reference posterior of the precision $\tau = 1/\sigma^2$ is gamma distribution with the same parameters). Intrinsic estimation of the mean produces mainly conventional results; the intrinsic point estimator is $\mu^* = \bar{x}$ and the intrinsic credible intervals are the HPD intervals in $\pi(\mu | \mathbf{z})$. The relevant reference expected intrinsic loss is

$$d(\mu_0 | \mathbf{z}) = d(t, n) \approx \frac{n}{2} \log \left[1 + \frac{1}{n+1} (1+t^2) \right],$$

a one-to-one function of the conventional test statistic $t = \sqrt{n-1}(\bar{x} - \mu_0)/s$. As $n \rightarrow \infty$, the function $d(t, n)$ converges to $(1+t^2)/2$; thus, for large samples (but only for large samples), there will be a one-to-one correspondence between the intrinsic test and any test based of the value of t . The implementation of the intrinsic test is however radically different: rather than relying on the sampling distribution of t , one simply checks whether or not $d(\mu_0 | t, n)$ indicates too large expected log-likelihood ratios against μ_0 . For instance, with $n = 10$ and $t = 2.262$ so that the p -value is 0.05, $d(t, n) = 2.387 = \log[10.9]$, so the average likelihood ratio against the null is expected to be about 11, hardly strong evidence for rejection.

Intrinsic estimation of σ (or of any one-to-one function of σ , since the methodology is invariant under those transformations) produces however new results. Thus, the intrinsic point estimator of σ is

$$\sigma_n^* \approx \frac{n}{n-1} s, \quad n > 2,$$

with $\sigma_2^* \approx (\sqrt{5}/2) |x_2 - x_1|$ when $n = 2$. As Figure 1 already suggested, the intrinsic estimator is larger than most conventional estimators (see Bernardo, 2007, for the exact, complicated expression). The differences are very noticeable for small sample sizes.

The exact form of intrinsic q -credible intervals for σ is complicated (see Bernardo, 2007, for details), but for moderate sample sizes they are approximately of the form

$$R_q^* = (\sigma^* e^{-a_q/\sqrt{2(n-1)}}, \sigma^* e^{a_q/\sqrt{2(n-1)}}),$$

with a_q chosen to have posterior probability q . As n increases, a_q converges to the $(q+1)/2$ quantile of the standard normal.

Using the intrinsic loss $\ell_\delta\{\sigma_0, \sigma | \mathcal{M}_\mathbf{z}\}$ derived in Example 2, the reference expected intrinsic loss is for using σ_0 as a proxy for σ is

$$d(\sigma_0 | s, n) = \int_0^\infty \ell_\delta\{\sigma_0, \sigma | \mathcal{M}_\mathbf{z}\} \pi(\sigma | s, n) d\sigma,$$

and testing the compatibility of the data with a particular σ_0 value reduces to evaluating $d(\sigma_0 | s, n)$. For instance, with $n = 10$ and $s = 1$, testing $\sigma_0 = 1.924$ (which is the 0.975-quantile of the reference posterior of σ) yields $d(\sigma_0 | s, n) = 2.367 = \log[10.7]$; thus the average likelihood ratio against σ_0 is expected to be about 11 which, again, is hardly strong evidence for rejection.

This is a general feature: frequentist rejection with a 0.05 p -value typically corresponds to an expected average likelihood ratio against the null of about 11, far from conclusive evidence for rejection.

Joint inferences about μ and σ are easily computed in terms of the expected intrinsic loss $d(\mu_0, \sigma_0 | \mathbf{z})$, the expectation of the intrinsic loss $\ell_\delta\{\mu_0, \sigma_0, (\mu, \sigma)\}$, derived in Example 2, with respect to the joint posterior which corresponds to the reference prior $\pi(\mu, \sigma) = \sigma^{-1}$.

Figure 4 is a contour plot of the expected intrinsic loss which corresponds to $n = 25$ observations, simulated from $N(x|0, 1)$, which yielded $\bar{x} = 0.024$ and $s = 1.077$. The resulting surface has a unique minimum at $(\mu^*, \sigma^*) = (0.024, 1.133)$, which is the intrinsic joint estimate, represented by a solid dot; notice that $\mu^* = \bar{x}$, and $\sigma^* \approx sn/(n-1) = 1.122$. The three contours shown describe the intrinsic q -credible regions which correspond to $q = 0.50, 0.95$ and 0.99 . For instance, the 0.95-credible region (middle contour in the figure)

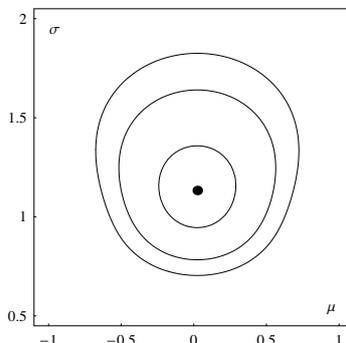


Figure 4: Joint reference analysis of the parameters of a univariate normal model.

is the set of $\{\mu_0, \sigma_0\}$ points whose intrinsic expected loss is not larger than $3.35 = \log[28]$. Testing a joint value $\{\mu_0, \sigma_0\}$ reduces to checking whether or not this point belongs to the intrinsic region defined by $d(\mu_0, \sigma_0 | \mathbf{z}) = \ell_0$, where ℓ_0 is the minimum average log-likelihood ratio against the null which is required for rejection.

In one-parameter problems, Theorem 3 may be used to obtain useful large sample approximations to the quantities required for intrinsic estimation and hypothesis testing. For details and proofs, see Bernardo (2005b).

Theorem 5 Let data $\mathbf{z} = \{x_1, \dots, x_n\}$ be a random sample from $p(x | \theta)$, and let $\tilde{\theta}_n$ be an asymptotically sufficient consistent estimator of θ with asymptotically normal sampling distribution of standard deviation $s(\theta)/\sqrt{n}$. Define $\phi(\theta) = \int^\theta s(y)^{-1} dy$. Then, for large n , $\ell\{\theta_0 | \mathbf{z}\} \approx (1/2) + (n/2)[E[\phi | \mathbf{z}] - \phi(\theta_0)]^2$, where $E[\phi | \mathbf{z}]$ is the expected posterior of $\phi(\theta)$. The intrinsic estimator of θ is $\theta^{-1}(\phi^*) \approx \theta^{-1}\{E[\phi | \mathbf{z}]\}$, and the intrinsic q -credible interval of θ is $R_q^* \approx \theta^{-1}\{E[\phi | \mathbf{z}] \pm n_q/\sqrt{n}\}$, where n_q is the $(q + 1)/2$ quantile of a standard normal distribution.

The following example, taken from the extra sensory power (ESP) testing literature, illustrates the radically different answers which the two alternative types of priors commonly used in Bayesian hypothesis testing may produce with the same data.

Example 7 (Binomial parameter: ESP testing). Let $\mathbf{z} = \{x_1, \dots, x_n\}$ be a random sample of size n from $p(x | \theta) = \theta^x (1 - \theta)^{1-x}$, with $x \in \{0, 1\}$, and let $r = \sum x_i$. The reference prior is the (proper) Beta $\pi(\theta) = \text{Be}(\theta | 1/2, 1/2)$, and the reference posterior is $\pi(\theta | r, n) = \text{Be}(\theta | r + 1/2, n - r + 1/2)$. The intrinsic loss function is $\ell_\delta(\theta_0, \theta) = n \kappa(\theta/\theta_0)$, if $\theta_0 < \min\{\theta, 1 - \theta\}$ or $\theta_0 > \max\{\theta, 1 - \theta\}$, and $n \kappa(\theta_0/\theta)$ otherwise, where

$$\kappa(\theta_j | \theta_i) = \theta_i \log \frac{\theta_i}{\theta_j} + (1 - \theta_i) \log \frac{1 - \theta_i}{1 - \theta_j}.$$

The expected intrinsic loss is the concave function $d(\theta_0 | r, n) = \int_0^1 \ell_\delta(\theta_0, \theta) \pi(\theta | r, n) d\theta$. For large sample sizes, Theorem 5 yields $d(\theta_0 | r, n) \approx (1/2) + (n/2)[E[\phi | \mathbf{z}] - \phi(\theta_0)]^2$, where $\phi(\theta) = 2 \arcsin \sqrt{\theta}$. Hence, the intrinsic estimator of $\phi(\theta)$ is $\phi^* \approx E[\phi | \mathbf{z}]$ and, by invariance, $\theta^* = \theta^{-1}(\phi^*)$. This yields $\theta^* \approx (r + 1/4)/(n + 1/2)$, which is close to the median of the reference posterior.

As an illustration, consider the results reported by Jahn, Dunne and Nelson (1987) using a random event generator based in a radioactive source, and arranged so that one gets a random sequence of 0's and 1's with theoretically equal probability for each outcome. A subject then attempted to mentally "influence" the results so that, if successful, data would show a proportion of 1's significantly different from 0.5. There were $n = 104,490,000$ trials resulting in $r = 52,263,471$ successes, about 0.018% over chance expectation. The huge sample size means that one may safely use asymptotic approximations. Using conventional testing procedures, the authors reject the hypothesis that $\theta_0 = 1/2$ on the grounds of the very low p -value they derive. Jefferys (1990) reanalyzed the data from a Bayesian perspective, using a prior which placed probability p_0 on $\theta_0 = 1/2$ and continuously spread the rest over the $(0, 1)$ interval, and obtained a posterior probability $\Pr[\theta_0 = 1/2 | r, n]$ larger than p_0 . Hence, this particular Bayesian analysis seems to support $\theta_0 = 1/2$ despite of the strong rejection by the classical test. This is a remarkable example of Lindley's paradox. To use the methods advocated here one simply computes the expected intrinsic loss to obtain $d(\theta_0 = 1/2 | r, n) = 7.24 = \log[1400]$ (we have used the reference prior, but given the huge sample size, any continuous prior will give essentially the same result). Thus, the expected minimum likelihood ratio against θ_0 is about 1400 and, we argue, the hypothesis that $\theta_0 = 1/2$ should really be rejected. Of course, this does not necessarily mean that the subject had extra sensory powers. Indeed, a more likely explanation is that the random event generator had some small bias. However, the argument establishes that, under the accepted assumptions, the *precise* value $\theta_0 = 1/2$ is rather incompatible with the data.

The following example illustrates the use of the methods described to derive novel solutions to paradigmatic problems.

Example 8 (Equality of Normal means). Let $\mathbf{z} = \{\mathbf{x}, \mathbf{y}\}$ be two independent random samples, $\mathbf{x} = \{x_1, \dots, x_n\}$ from $N(x | \mu_x, \sigma_x)$, and $\mathbf{y} = \{y_1, \dots, y_m\}$ from $N(x | \mu_y, \sigma_y)$, and suppose that one is interested in comparing the two means. In particular, one may be interested in testing whether or not the precise hypothesis $H_0 \equiv \{\mu_x = \mu_y\}$ is compatible with available data \mathbf{z} . Consider first the case where it may be assumed that $\sigma_x = \sigma_y$. Using the additive property of the intrinsic loss and the first result in Example 2, to derive the logarithmic divergence of $p(\mathbf{z} | \mu_0, \mu_0, \sigma_0)$ from $p(\mathbf{z} | \mu_x, \mu_y, \sigma)$, and then minimizing over both μ_0 and σ_0 yields $\inf_{\mu_0 \in \mathbb{R}, \sigma_0 > 0} \kappa\{\mu_0, \mu_0, \sigma_0 | \mu_x, \mu_y, \sigma\} = k_{nm} \theta^2$, where $k_{nm} = 2nm/(m+n)$ is the harmonic mean of the two sample sizes, and $\theta = (\mu_x - \mu_y)/\sigma$ is the standardized difference between the two means. On the other hand, $\inf_{\mu_0 \in \mathbb{R}, \sigma_0 > 0} \kappa\{\mu_x, \mu_y, \sigma | \mu_0, \mu_0, \sigma_0\}$ yields $[(m+n)/2] \log[1 + (k_{nm}/(2(m+n)))\theta^2]$, which is always smaller. Hence, the intrinsic loss of accepting H_0 is

$$\ell_\delta\{H_0, (\mu_x, \mu_y, \sigma)\} = \ell_\delta\{H_0, \theta | \mathcal{M}\} = \frac{n+m}{2} \log \left[1 + \frac{k_{nm}}{2(n+m)} \theta^2 \right],$$

which reduces to $n \log[1 + \theta^2/4]$ when $n = m$. Here, the parameter of interest is θ . Bernardo and Pérez (2007) find that the marginal reference posterior of θ only depends on the data through the sample sizes and $t = t(\mathbf{z}) = (\bar{x} - \bar{y})/(s/\sqrt{2/k_{nm}})$, where s is the MLE of σ . Therefore, the required marginal reference posterior of θ is $\pi(\theta | \mathbf{z}) = \pi(\theta | t, m, n) \propto p(t | \theta) \pi(\theta)$ where $p(t | \theta)$ is the noncentral Student sampling distribution of t , and $\pi(\theta) = (1 + (k_{nm}/(4(m+n)))\theta^2)^{-1/2}$ is the marginal reference prior for θ . The posterior $\pi(\theta | t, m, n)$ may be used to provide point and interval estimates of θ , the standardized difference between the two means, and hence inferential statements about their relative positions.

The expected intrinsic loss $d(H_0 | t, n, m) = \int_{-\infty}^{\infty} \ell_\delta\{H_0, \theta | \mathcal{M}\} \pi(\theta | t, n, m) d\theta$ may be used to test H_0 . This has no simple analytical expression, but its value may easily be obtained by one-dimensional numerical integration. A good large sample approximation is

$$d(H_0 | t, n, m) \approx \frac{n+m}{2} \log \left[1 + \frac{1}{n+m} (1 + t^2) \right].$$

The sampling distribution of $d(H_0 | t, n, m)$ is asymptotically $(1/2)[1 + \chi_1^2(\lambda)]$, where $\chi_1^2(\lambda)$ is a non-central chi-squared distribution with one degree of freedom and non-centrality parameter $\lambda = k_{nm}\theta^2/2$. It follows that the expected value under sampling of $d(H_0 | t, n, m)$ is equal to one when $\mu_x = \mu_y$, and increases linearly with the harmonic mean of the samples when this is not true. Thus, the testing procedure is consistent.

In the general case, when the two variances σ_x and σ_y are allowed to be different, the intrinsic loss function is

$$\ell_\delta\{H_0, (\mu_x, \mu_y, \sigma_x, \sigma_y | \mathcal{M})\} \approx \frac{n}{2} \log \left[1 + \frac{\theta_1^2}{(1 + \tau^2)^2} \right] + \frac{m}{2} \log \left[1 + \frac{\theta_2^2}{(1 + \tau^{-2})^2} \right],$$

where $\theta_1 = (\mu_x - \mu_y)/\sigma_x$ and $\theta_2 = (\mu_x - \mu_y)/\sigma_y$ are the two standardized differences of the means, and $\tau = (n\sigma_y)/(m\sigma_x)$ is a measure of the design balance. Derivation of the exact form of the joint reference prior $\pi_\phi(\mu_x, \mu_y, \sigma_x, \sigma_y)$ when the quantity of interest is $\phi = \ell_\delta\{H_0, (\mu_x, \mu_y, \sigma_x, \sigma_y | \mathcal{M})\}$ is daunting, but the arguments in Subsection 4.2 may be invoked to use the joint reference prior $\pi(\mu_x, \mu_y, \sigma_x, \sigma_y) = \sigma_x^{-1}\sigma_y^{-1}$. Indeed, this prior gives the correct marginal reference posteriors for the four parameters, and may be therefore expected to provide a marginal posterior for ϕ not too different from its exact reference posterior.

It follows from Theorem 2 that, under regularity conditions for asymptotic normality, the two KL divergences whose minimum define the intrinsic discrepancy converge to a common, symmetric limit. Hence, for large samples one may just take whichever of those is easier to compute, which typically is the KL divergence of the null from the assumed model, and use $\ell\{\theta_0, (\theta, \lambda)\} \approx \inf_{\lambda_0 \in \Lambda} \kappa\{\theta_0, \lambda_0 | \theta, \lambda\}$. Moreover, the exact reference prior when the parameter of interest is taken to be $\phi = \ell\{\theta_0, (\theta, \lambda)\}$ may well be very difficult to derive, but one may use the arguments described in Subsection 4.2, and use instead the approximate joint reference prior whose marginal posteriors minimize the expected average intrinsic discrepancies from the exact reference posteriors for all the parameters involved.

We conclude this section by using these two approximations to obtain relatively simple solutions to a couple of important problems in precise hypothesis testing. We first consider a question in genetics which has become an important test case to compare alternative procedures for precise hypothesis testing.

Example 9 (Hardy-Weinberg Equilibrium). To determine whether or not a population mates randomly is an important problem in biology. At a single autosomal locus with two alleles, a diploid individual has three possible genotypes, typically denoted $\{AA, aa, Aa\}$, with (unknown) population frequencies $\{\alpha_1, \alpha_2, \alpha_3\}$, where $0 < \alpha_i < 1$ and $\sum_{i=1}^3 \alpha_i = 1$. The population is said to be in Hardy-Weinberg (HW) equilibrium (compatible with random mating) if there exists a probability $p = \Pr(A)$, $0 < p < 1$, such that

$$\{\alpha_1, \alpha_2, \alpha_3\} = \{p^2, (1-p)^2, 2p(1-p)\}.$$

Given a random sample of size n from the population, and observed $\mathbf{z} = \{n_1, n_2, n_3\}$ individuals (with $n = n_1 + n_2 + n_3$) from each of the three possible genotypes $\{AA, aa, Aa\}$, the question is whether or not these data support the hypothesis of HW equilibrium. This is a good example of *precise* hypothesis in the sciences, since HW equilibrium corresponds to a zero measure set within the original simplex parameter space.

The relevant statistical model is obviously trinomial $\text{Tr}(n_1, n_2 | \alpha_1, \alpha_2, n)$, where the parameter space is the simplex $\mathcal{A} = \{(\alpha_1, \alpha_2); 0 < \alpha_1 < 1, 0 < \alpha_2 < 1, 0 < \alpha_1 + \alpha_2 < 1\}$, while the hypothesis H_0 to test, the HW equilibrium, is the line with parametric equations $\{(\alpha_1, \alpha_2); \alpha_1 = p^2, \alpha_2 = (1-p)^2, 0 < p < 1\}$, so that $\sqrt{\alpha_1} + \sqrt{\alpha_2} = 1$.

The KL divergence of H_0 from the model is the minimum, for all p in $(0, 1)$, of $\kappa\{\text{Tr}(\cdot | p^2, (1-p)^2, n) | \text{Tr}(\cdot | \alpha_1, \alpha_2, n)\}$. This minimum is achieved at $p = (1 + \alpha_1 - \alpha_2)/2$, which would be the value of $\text{Pr}(A)$ if the population were really in HW equilibrium. Substitution yields the intrinsic loss,

$$\ell_\delta\{H_0, (\alpha_1, \alpha_2)\} \approx n [2 H\{\omega, 1 - \omega\} - H\{\alpha_1, \alpha_2, 1 - \alpha_1 - \alpha_2\} - (1 - \alpha_1 - \alpha_2) \log[2]],$$

where $H\{q_1, \dots, q_k\} = -\sum_{j=1}^k q_j \log q_j$ is the entropy of a discrete distribution (q_1, \dots, q_k) , and $\omega = (1 + \alpha_1 - \alpha_2)/2$. As explained above, this approximation assumes that the KL divergence of the model from the null, $\kappa\{\text{Tr}(\cdot | \alpha_1, \alpha_2, n) | \text{Tr}(\cdot | p^2, (1-p)^2, n)\}$, attains a similar minimum. It has been numerically verified that this is indeed the case, even for very moderate sample sizes.

The function $\ell_\delta\{H_0, (\alpha_1, \alpha_2) | \mathcal{M}_z\}$ is a measure on $[0, n \log 2]$ of the divergence between the null H_0 and the model identified by (α_1, α_2) .

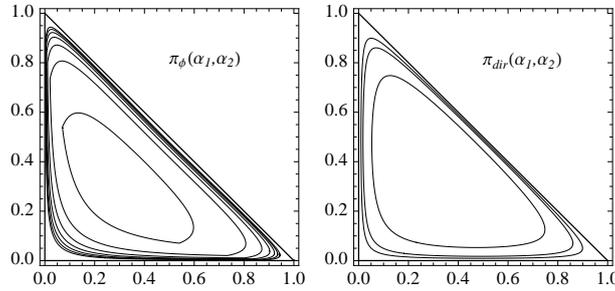


Figure 5: Exact and approximate reference priors for testing Hardy-Weinberg equilibrium.

The (proper) reference prior $\pi_\phi(\alpha_1, \alpha_2)$ when $\phi(\alpha_1, \alpha_2) = \ell_\delta\{H_0, (\alpha_1, \alpha_2)\}$ is the quantity of interest was obtained in Bernardo and Tomazella (2010), and it is rather complicated. Its contour plot is represented in the left panel of Figure 5. For comparison, the right panel shows the Dirichlet prior with parameter vector $(1/3, 1/3, 1/3)$, so that

$$\pi_{dir}(\alpha_1, \alpha_2) = \Gamma^{-3}[1/3] \alpha_1^{1/3-1} \alpha_2^{1/3-1} (1 - \alpha_1 - \alpha_2)^{1/3-1}$$

which, as described in Example 4, has been found to be the best global approximation for the trinomial model. It may be noticed that the two priors are not very different. To test H_0 given data $\{n_1, n_2, n_3\}$ one numerically computes

$$d(H_0 | n_1, n_2, n_3) = \int_{\mathcal{A}} \ell_\delta\{H_0, (\alpha_1, \alpha_2) | \mathcal{M}_z\} \pi(\alpha_1, \alpha_2 | n_1, n_2, n_3) d\alpha_1 d\alpha_2,$$

and reports the value obtained. With the posterior which corresponds to the reference prior this requires rather delicate numerical analysis. If the Dirichlet prior is used, the numerical integration is straightforward: one simply generates a large number of samples from the corresponding Dirichlet posterior, with parameter vector $(n_1 + 1/3, n_2 + 1/3, n_3 + 1/3)$, and computes the average of the corresponding $\ell_\delta\{H_0, (\alpha_1, \alpha_2) | \mathcal{M}_z\}$ values. As one would expect, the results obtained from both priors are qualitatively similar.

For instance, simulation of $n = 30$ observations from a trinomial with $\{\alpha_1, \alpha_2\} = \{0.45, 0.40\}$, so that the population is *not* in HW equilibrium (the actual value of the intrinsic discrepancy is $n\phi(0.45, 0.40) = n \cdot 0.269 = 8.08$), yielded $\{n_1, n_2, n_3\} = \{12, 12, 6\}$. The expected posterior intrinsic loss with the exact and the approximate reference priors were respectively $5.84 = \log 334$ and $5.82 = \log 336$, both clearly indicating rejection.

Similarly, simulation of another 30 observations from a population in HW equilibrium (with $p = \Pr[A] = 0.3$, so that $\{\alpha_1, \alpha_2\} = \{p^2, (1-p)^2\} = \{0.09, 0.49\}$), yielded $\{n_1, n_2, n_3\} = \{2, 15, 13\}$ and expected posterior intrinsic losses $0.321 = \log 1.38$ and $0.51 = \log 1.66$, both suggesting that those data are certainly compatible with the hypothesis of HW equilibrium.

Our final example provides a new Bayesian objective procedure to test independence in contingency tables.

Example 10 (Independence in contingency tables). Consider an $a \times b$ contingency table, with unknown parameters $0 \leq \theta_{ij} \leq 1$, $\sum_{i=1}^a \sum_{j=1}^b \theta_{ij} = 1$, and let $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_a\}$ and $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_b\}$ be the corresponding marginal distributions. Thus,

$$\alpha_i = \sum_{j=1}^b \theta_{ij}, \quad \sum_{i=1}^a \alpha_i = 1, \quad \beta_j = \sum_{i=1}^a \theta_{ij}, \quad \sum_{j=1}^b \beta_j = 1.$$

Given a random sample of size n from the population, and observed n_{ij} individuals in each of the $a \times b$ cells, so that $\mathbf{z} = \{\{n_{11}, \dots, n_{1b}\}, \dots, \{n_{a1}, \dots, n_{ab}\}\}$, with $0 \leq n_{ij} \leq n$ and $\sum_{i=1}^a \sum_{j=1}^b n_{ij} = n$, the question is whether or not these data support the hypothesis of independence, $H_0 \equiv \{\theta_{ij} = \alpha_i \beta_j, \forall i, \forall j\}$. This is another example of precise hypothesis testing since H_0 corresponds to a zero measure set within the original simplex parameter space.

The KL divergence of H_0 from the model is the minimum, for all α_0 and β_0 distributions, of the KL divergence $\kappa\{\alpha_{01}\beta_{01}, \dots, \alpha_{0a}\beta_{0b} \mid \theta_{11}, \dots, \theta_{ab}\}$ of a k -multinomial distribution with parameters $(\alpha_{01}\beta_{01}, \dots, \alpha_{0a}\beta_{0b})$ from one with parameters $(\theta_{11}, \dots, \theta_{ab})$, where $k = a \times b$ is the total number of cells. The minimum is achieved when when $\alpha_{0i} = \sum_{j=1}^b \theta_{ij}$ and $\beta_{0j} = \sum_{i=1}^a \theta_{ij}$, that is, when $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ take the values which $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ would have under independence. Substitution yields

$$\ell_\delta\{H_0, \boldsymbol{\theta} \mid \mathcal{M}_\mathbf{z}\} \approx n \sum_{i=1}^a \sum_{j=1}^b \theta_{ij} \log \left[\frac{\theta_{ij}}{\alpha_i \beta_j} \right] = n \phi(\boldsymbol{\theta}),$$

where $\phi(\boldsymbol{\theta}) = \sum_i \sum_j \theta_{ij} \log[\theta_{ij}/(\alpha_i \beta_j)]$ is the KL divergence of the discrete distribution on the k -dimensional simplex with probabilities $\alpha_i \beta_j$ from the discrete distribution on the same space with with probabilities θ_{ij} . The function $\phi(\boldsymbol{\theta})$ is a non-negative conditional measure of how far a contingency table with parameters θ_{ij} is from independence, and it is zero if (and only if) the independence condition is verified. Once again, the approximation sign refers to the fact that only the KL-divergence of H_0 from the model, which may be analytically found, has been considered. It has been numerically verified however, that the KL-divergence of the model from H_0 yields very similar values.

Derivation of the joint reference prior $\pi_\phi(\boldsymbol{\theta})$ when the parameter of interest is $\phi(\boldsymbol{\theta})$ does not seem to be analytically feasible. Thus, we invoke again the arguments in Subsection 4.2 and Example 4, and use instead the corresponding approximate joint reference prior which, in this case, is a $(k-1)$ -dimensional Dirichlet with parameter vector $\{1/k, \dots, 1/k\}$. This leads to an joint reference posterior $\pi(\boldsymbol{\theta} \mid \mathbf{z})$ which is a $(k-1)$ -dimensional Dirichlet with parameter vector $\{n_{11} + 1/k, \dots, n_{ab} + 1/k\}$, from which simulation is straightforward. The expected intrinsic loss,

$$d\{H_0 \mid \mathbf{z}\} \approx n \int_{\Theta} \phi(\boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{z}) d\boldsymbol{\theta},$$

where Θ is the $(k-1)$ dimensional simplex, may easily be computed by Monte Carlo. One simulates a large number s of $\boldsymbol{\theta}_j$ values from $\pi(\boldsymbol{\theta} \mid \mathbf{z}) = \text{Di}_{k-1}(\boldsymbol{\theta} \mid n_{11} + 1/k, \dots, n_{ab} + 1/k)$, and evaluates $d\{H_0 \mid \mathbf{z}\} \approx (n/s) \sum_{j=1}^s \phi(\boldsymbol{\theta}_j)$.

To illustrate the procedure, we describe the results obtained with data simulated from two different 2×2 contingency tables, one where independence holds, and another where independence does not hold. In the first case, $n = 100$ observations were simulated from

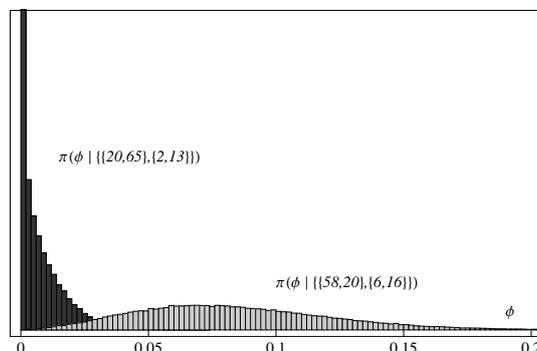


Figure 6: Posterior distributions of $\phi = \phi(\theta)$ in 2×2 contingency tables, under both independence (left density), and no independence (right density).

a contingency table with cell probabilities $\{\{0.24, 0.56\}, \{0.06, 0.14\}\}$, an independent contingency table (which therefore has $\phi(\theta) = 0$), with marginal probabilities $\{0.8, 0.2\}$ and $\{0.3, 0.7\}$. This yielded data $\mathbf{z} = \{\{20, 65\}, \{2, 13\}\}$. The marginal posterior distribution of ϕ , obtained from 100,000 simulations from the corresponding Dirichlet joint reference posterior is represented at the left side of Figure 6. This has an expected value of 0.0080. Thus, the expected intrinsic loss is $d\{H_0 | \mathbf{z}\} = n E[\phi | \mathbf{z}] = 0.80 = \log[2.23]$, suggesting that the observed data are indeed compatible with the independence hypothesis.

For the second case, $n = 100$ observations were simulated from a non independent contingency table with cell probabilities $\{\{0.60, 0.20\}, \{0.05, 0.15\}\}$, (where the true value of the quantity of interest is $\phi(\theta) = 0.0851$) and obtained data $\mathbf{z} = \{\{58, 20\}, \{6, 16\}\}$. The corresponding marginal posterior distribution of ϕ is represented at the right side of Figure 6. This has an expected value of 0.0835. Thus, the expected intrinsic loss is $d\{H_0 | \mathbf{z}\} = n E[\phi | \mathbf{z}] = 8.35 = \log[4266]$, clearly suggesting that the observed data are *not* compatible with the independence assumption.

REFERENCES

- Bartlett, M. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika* **44**, 533–534.
- Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis* **1**, 385–402 and 457–464 (with discussion).
- Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84**, 200–207.
- Berger, J. O. and Bernardo, J. M. (1992a). Ordered group reference priors with applications to a multinomial problem. *Biometrika* **79**, 25–37.
- Berger, J. O. and Bernardo, J. M. (1992b). Reference priors in a variance components problem. *Bayesian Analysis is Statistics and Econometrics* (P. K. Goel and N. S. Yyengar, eds.) Berlin: Springer, 323–340.
- Berger, J. O. and Bernardo, J. M. (1992c). 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).
- Berger, J. O., Bernardo, J. M. and Sun, D. (2009). The formal definition of reference priors. *Ann. Statist.* **37**, 905–938.
- Berger, J. O., Bernardo, J. M. and Sun, D. (2011a). Reference priors for discrete parameters. *J. Amer. Statist. Assoc.* (under revision).

- Berger, J. O., Bernardo, J. M. and Sun, D. (2011b). Overall reference priors. *Tech. Rep.*, Duke University, USA.
- 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. (1997). Noninformative priors do not exist. *J. Statist. Planning and Inference* **65**, 159–189 (with discussion).
- Bernardo, J. M. (2005a). Reference analysis. *Bayesian Thinking: Modeling and Computation, Handbook of Statistics* **25** (Dey, D. K. and Rao, C. R., 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 Pub, 110–121.
- Bernardo, J. M. (2007). Objective Bayesian point and region estimation in location-scale models. *Sort* **31**, 3–44, (with discussion).
- Bernardo, J. M. and Juárez, M. A. (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 Pérez, S. (2007). Comparing normal means: New methods for an old problem. *Bayesian Analysis* **2**, 45–58.
- 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.
- Bernardo, J. M. and Tomazella, V. (2010). Bayesian reference analysis of the Hardy-Weinberg equilibrium. *Frontiers of Statistical Decision Making and Bayesian Analysis. In Honor of James O. Berger* (M.-H. Chen, D. K. Dey, P. Müller, D. Sun and K. Ye, eds.) New York: Springer, 31–43.
- Datta, G. S. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*. Berlin: Springer.
- Dawid, A. P., Stone, M. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. B* **35**, 189–233 (with discussion).
- Ghosh, J. K., Delampady, M. and Samanta, T. (2006). *An Introduction to Bayesian Analysis: Theory and Methods*. Berlin: Springer.
- Ghosal, S. and Samanta, T. (1997). Expansion of Bayes risk for entropy loss and reference prior in nonregular cases. *Statistics and Decisions* **15**, 129–140.
- Jahn, R.G., Dunne, B.J. and Nelson, R.D. (1987). Engineering anomalies research. *J. Scientific Exploration* **1**, 21–50.
- 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.
- Jefferys, W. H. (1990). Bayesian Analysis of Random Event Generator Data. *J. Scientific Exploration* **4**, 153–169.
- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Royal Soc.* **186**, 453–461.
- Jeffreys, H. (1961). *Theory of Probability* (3rd edition). Oxford: University Press.
- Juárez, M. A. (2004). *Métodos Bayesianos Objetivos de Estimación y Contraste de Hipótesis*. Ph.D. Thesis, Universitat de València, Spain.
- Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91**, 1343–1370.
- Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.
- Perks, W. (1947). Some observations on inverse probability, including a new indifference rule. *J. Inst. Actuaries* **73**, 285–334 (with discussion).

- Robert, C. P. (1993). A note on Jeffreys-Lindley paradox. *Statistica Sinica* **3**, 601–608.
- Robert, C. P. (1996). Intrinsic loss functions. *Theory and Decision* **40**, 192–214.
- Stone, M. (1976). Strong inconsistency from uniform priors. *J. Amer. Statist. Assoc.* **71**, 114–125 (with discussion).
- Sun, D. and Berger, J. O. (1998). Reference priors under partial information. *Biometrika* **85**, 55–71.

DISCUSSION

LUIS PERICCHI (*Universidad de Puerto Rico en Río Piedras, Puerto Rico*)

The achievements of a unified objective Bayesian decision theory. I begin by listing some the achievements summarized by this historical paper, the last invited talk in a Valencia meeting:

- (i) Professor Bernardo proved that it can be done! To put point estimation, interval estimation and hypothesis testing in the same (objective Bayes) decision theory system. This is very ambitious and an achievement in itself.
- (ii) Reference priors have emerged over the years, starting with Bernardo (1979) read paper to the Royal Statistical Society, and followed by Berger and Bernardo (1992) and later papers with Sun, as the most successful and accepted method to develop objective priors for estimation. It can be argued that reference priors gave a key contribution to make Bayesian statistics “acceptable” to mainstream statistics, since it solved several “paradoxes” and criticisms exposed by eminent statisticians and philosophers.
- (iii) Bernardo’s version of “intrinsic” loss has interesting mathematical properties, and in the examples given lead to attractive distances between distributions.
- (iv) Less convincing (in my view) is Bernardo’s replacement of HPD Intervals, although admittedly invariance is a convenient property that his procedure enjoys. But recall :

“The authors feel that in general nonlinear transformations *ought* to change the relative credibility of any two parameter points and that invariance under nonlinear transformation is therefore not to be expected. Insistence on invariance for problems which ought not to be invariant serves only to guarantee inappropriate solutions” (Box and Tiao, 1973, p. 124).

- (v) The least convincing, my opinion, is this paper recipe for hypothesis testing. I now concentrate my discussion to this last point.

Does the Bayes factor need replacement? or rather the p-value ought to be replaced as a measure of evidence for better scientific practice?

Three of the main arguments in the paper, against the Bayes factor and posterior model probabilities are: (i) changes the prior used for estimation to a different one used for testing for the same parameters (ii) assumes a positive probability of the null hypothesis $\Pr(H_0) = p_0$, say, and (iii) the loss function should be an “intrinsic loss” function (which incidentally has nothing to do with “intrinsic priors”) that we call here “Bernardo’s loss”(since there are other “intrinsic” loss functions) . We revisit the interesting Example 7 on ESP testing, to illustrate that the three

criticisms above simply do not apply to this example. Furthermore, this example show a disturbing similarity of the testing procedure of the paper with p -values, at least for large samples.

Extra Sensorial Perception: ESP or no ESP? Here, $H_0 : p = 1/2$ vs. $p \neq 1/2$ and we have a huge sample, $n = 104,490,000$, with $s = 52,263,471$ successes and ratio: $s/n = 0.5001768$. The p -value against the null is minute, namely 0.0003, leading to a compelling rejection of H_0 . The expected Bernardo's loss, $7.03 = \log(1130)$, is bigger than the largest incompatibility in the author's table ($\log(1000)$) and also compellingly leads to the rejection of H_0 . But, we can calculate in this situation a Bayes factor (BF) with a reference prior, since the reference prior is proper. The reference (also Jeffreys) prior is

$$\pi^R(p) = \text{Be}(p | 1/2, 1/2) = \pi^{-1} p^{-1/2} (1-p)^{-1/2}.$$

Then the Bayes Factor is there! Without any extra assumptions. It is the *ratio of evidences* (as it is called by physicists),

$$\text{BF}_{01} = \frac{p(\text{data} | p = 1/2)}{\int p(\text{data} | p) \pi^R(p) dp} = \frac{\pi (1/2)^n}{B(s + 0.5, n - s + 0.5)} = 18.7,$$

where $B(a, b)$ above denotes the Beta function. Jefferys (1990), calculated this as 12 using a uniform prior. Thus with the same prior used for estimation, the data favours the null hypothesis and the ratio of evidences is close to 19, pointing at opposite direction than the expected Bernardo's loss. So it is *not* the change of the prior the reason for the discrepancy, and notice that we have not yet assumed anything about the probability of a hypothesis. It is interesting that the Bayes factor is still not compelling although substantially in favour of H_0 .

Now let us assume that the prior probability of the null hypothesis is *not* zero (if we assume it is zero, then for *any* data the posterior probability is zero, a case of pure dogmatism or a violation of Lindley's Cromwell's rule.) Assume then that the probability associated to H_0 is *not* exactly 0 (if it is, what is the reason to test?). If $\Pr(H_0) > \epsilon > 0$ then, by Bayes theorem,

$$\Pr(H_0 | \text{data}) \geq \left(1 + \frac{(1-\epsilon)}{\epsilon \text{BF}_{01}}\right)^{-1}.$$

If $\epsilon = 0.1$, then $\Pr(H_0 | \text{data}) \geq 0.67$, while if $\epsilon = 0.5$, then $\Pr(H_0 | \text{data}) \geq 0.95$. So, the null hypothesis is very likely, but not overwhelmingly so! Notice that, for whatever loss function, the posterior probability is a fundamental inferential quantity that summarizes the evidence.

But, is the reference prior sensible for the test? It is not, because it does not take into account the special status of the null point, $p = 1/2$ (which is objective information). Also, ironically, it is too favourable to the null, because the prior under the alternative, does not give high probability to alternatives close to the null.

General and amazing fact. To alleviate the divergence between Bayes Factors and p -values, in the direction of a p -value, it is necessary to put substantial (very subjective) prior probability around the null, so that the prior has an information content comparable to the likelihood. To illustrate this general point let us assume a simple class of priors:

$$p(p | a, a) = \text{Be}(p | a, a), \quad 1/2 \leq a \leq n/2,$$

a class of Beta densities centered in the null hypothesis, and letting the “equivalent sample size”, equal to $2a$, from one to 18 millions.

$2a$	1	10	50	100	200	500	9,000,000	18,000,000
B_{01}	18.7	4.83	2.12	1.49	1.055	0.67	0.009	0.01

Here I follow a principle first stated in Berger and Pericchi (1996) (see also Polson and Scott, 2011) “The best way to analyse a statistical procedure is to judge the prior that yields it.” I would argue that a prior more concentrated than an equivalent sample size of say $m = 2a = 500$, can be thought as of a very dogmatic prior. See Pericchi (2010) for an argument not to take equivalent prior samples larger than $n^{1/3}$, the cubic root of the sample size. In fact the minimum of BF_{01} over the whole range (up to $2a = n$) is for $n = 9$ million. It is for that dogmatic prior that the Bayes factor yields overwhelming support against H_0 , and still the overall minimum BF is 30 times bigger than the p -value and the maximum $1/BF_{01}$ is ten times smaller than Bernardo’s rate of 1130. But for the reference prior, and for a reasonable range of priors, the Bayes factors are not overwhelming but cautiously in favour of H_0 . I argue that this type of summary is better suited to science than an inordinately strong rejection of H_0 . It has long been argued that Bayes factors are sensitive to change in the priors. But it is better to have a large interval of posterior probabilities in the right direction than to decide robustly, in the wrong direction.

I will finish this subsection with two illuminating quotations, both about testing without posterior probabilities:

“Do you want to reject a hypothesis? Just take enough data!” (Wonnacott and Wonnacott in several of their writings).

“In real life, null hypothesis will always be rejected if enough data are taken because there will be inevitably be uncontrolled sources of bias”. (Berger and Delampady, 1987).

Posterior model probabilities may improve scientific practice. There is a growing dissatisfaction in the scientific community in the way evidence is weighted via significance testing. For example, recall

“*Law of initial results:* so often early promising results are followed by others that are less impressive. It is almost as if there is a law that states that first results are always spectacular, and subsequent ones are mediocre: the law of initial results.” (Ioannidis, 2005)

This is consistent with Jeffreys(1961) comments.

Consistency of the Bayes Factor with objective (intrinsic) priors, but not of the expected loss $d(H_0|t, n)$.

Mathematical consistency is a very relevant and a minimally necessary requirement for a procedure to be reasonable. To illustrate the inconsistency under the null of the expected Bernardo’s loss for decision in hypothesis testing, take Example 8, on the equality of Normal means, with $n = m$, and let $N = 2n$. Here the criterion of this paper is $\exp[-d(H_0|t, N)] = [1 + N^{-1}(1 + t^2)]^{-N/2}$, which under H_0 converges to $\exp(-\frac{1}{2}(1 + t^2))$ as $N \rightarrow \infty$, and hence is it is bounded under H_0 . Thus, there a positive probability of missing H_0 , even with infinite information sampled from it. Another way to check inconsistency under the null is from the

result in the paper in Example 8; indeed, $d(H_0|t, N)$ is distributed asymptotically as $\frac{1}{2}(1 + \chi_1^2(\lambda))$, with the noncentrality parameter $\lambda = n\theta^2/2$, which is zero under H_0 . So for all sample sizes N , no matter how large we choose the cut-point to decide against and in favour of H_0 it follows that the probability of wrong rejection is positive $P(\text{Reject } H_0|H_0) > 0$.

Quite differently, under the alternative hypothesis H_1 , Bernardo's procedure is consistent, since now the non-centrality parameter $\lambda \rightarrow \infty$ with N .

Of course I am not claiming that the proposed procedure is equivalent with significance testing for all sample sizes; in fact in Example 6 there is an instance of a difference with a sample of size 10. But for large samples, at least for the test of comparison of normal means it turns out that,

$$2 \times d(H_0|t, n) \simeq 1 - 2 \times \log(\text{Likelihood Ratio}01),$$

and so the inference has no growing penalty with the sample size for over-parameterization, and thus it is not consistent under H_0 . The same occurs with procedures based on Akaike's criterion or DIC. Indeed the expression above is equivalent with Akaike's but with smaller penalty for over-parametrization, multiplying by one the extra parameters, instead of two as in Akaike. None of these procedures, place the null and the alternative hypothesis on equal footing, and it can be claimed that are biased in favour of the more complex hypothesis. There is a clever resource in the paper, in that the scale chosen to reject the null is set rather high, like $\log(100)$ or $\log(1000)$. But the problem with the procedure is deep, since that scale should not be independent of the sample size, or the amount of information in the problem. (See above in the ESP example that $\log(1000)$ was not high enough, but in a problem with $n = 10$, $\log(10)$ may be too high, the same problem as with p -values.)

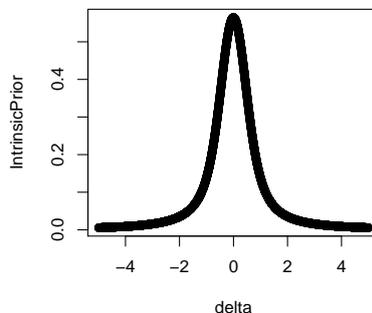


Figure 7: *Intrinsic prior for the difference of normal means, centred at the null hypothesis*

The problem of consistency and the right scale for rejection has now solutions, via objective Bayes factors, based for example on intrinsic priors, which is consistent both under the null and under the alternative hypothesis. To see this, in the example of comparison of means, let us use the intrinsic prior with a conceptual prior for simplicity of size 4 (minimal training sample is 3)

$$\delta^2 = \frac{(\mu_x - \mu_y)^2}{4}, \quad \text{IPrior } \pi^I(\delta) = \frac{\sigma}{4\sqrt{\pi}\delta^2} [1 - \exp(-4\delta^2/\sigma^2)],$$

$$BF_{01}^I \rightarrow \sqrt{n} \exp[-t_{2n-2}^2/2] \rightarrow \infty \text{ under } H_0, \text{ and } \rightarrow 0, \text{ under } H_1.$$

So it is consistent both under H_0 and H_1 . The advantage is that we now have methods for specifying objective priors for model comparison and testing, particularly the *intrinsic priors*, Berger and Pericchi (1996), that naturally are centred around the null hypothesis (see Figure 7). Objective priors centred around the null (a tradition that comes from Jeffreys, de Finetti, and Savage among others) are better for testing than reference priors.

The paradox about Lindley paradox.

“Paradox: a person or thing that combines contradictory features or qualities”
The Oxford English Dictionary.

Lindley’s paradox has been misunderstood in several places, including by myself in the distant past. It is unfortunate that opposite to Lindley’s written words, his “paradox” has been misunderstood as an “*illness*” of Bayes factors and posterior probabilities. To put the record straight I propose to rename it, and have the word *paradox* replaced by “disagreement”, “discrepancy” or “divergence”.

Let us recall some of Lindley’s comments and Bartlett’s replica, both warning about the problems of significance testing with fixed significance levels.

The phenomenon (conflict between significance test at fixed level and posterior probability of a hypothesis) is fairly general with significance tests and casts doubts on the meaning of a significance level in some circumstances” ... “5% in today’s small sample does not mean the same as 5% in tomorrow’s large one... The value of θ_0 is fundamentally different for any value of $\theta_0 \neq \theta$. (Lindley, 1957).

I would agree that he (Lindley) establishes the point that one must be cautious when using a fixed significance level for testing a null hypothesis irrespective of the size of sample one is taken. (Bartlett, 1957).

The above quotations from Lindley and the replica by Bartlett, establish in a crystal clear way, that it is significance testing, and equivalent to significance testing (even Bayesian) procedures, with fixed (with n) errors that deserve scepticism. I suggest changing the misleading name “Lindley paradox” (Jeffreys exposed earlier the phenomenon as recognized by Lindley), by Bayes/Fisher discrepancy (as suggested by I. J. Good) or by Bayes/NonBayes disagreement (as suggested by J. O. Berger) or by probabilities/ p -values divergence, stressing that they diverge as the information accumulates. The Oxford English Dictionary To resolve the divergence we have an advantage now: we have methods to assign objective priors for testing in some generality, such as intrinsic priors, improved BIC’s and many others.

Posterior probabilities answer a scientific question that p -values cannot answer.

“What is the probability of a hypothesis or a model given the data?”, is perhaps the most relevant question for a scientist. This is a Bayesian question. We should be able to answer it!

Epilogue: The significance of the Valencia meetings. As one of the statisticians of the generations deeply influenced by the Valencia meetings, it is fair to say: *Gracias José-Miguel* for the nine Valencia meetings!, —and for keeping yourself as active as ever— This gratitude is extensive to the Valencia Organizing Committee. You have made statistical history and deserve our thanks!

BERTRAND CLARKE (*University of Miami, USA*)

The role of integrated Bayesian analysis is best seen as an extension of reference priors to an entire reference analysis. That is, integrated Bayesian analysis is so clear and so precise that its most important use may well be as a standard with which other analyses may be compared. It's not that the integrated analysis necessarily generates the inferences one wants to report; rather, the inferences one reports may be better interpreted if compared with the integrated analysis.

Professor Bernardo has made an enormous number of deep intellectual contributions over several decades. He has also given all of us fully nine Valencia conferences which have dramatically accelerated the development and dissemination of Bayesian thinking – to say nothing of the great fun we have had! In short, my esteem and respect for Professor Bernardo is unbounded. Now, Bernardo has written an important paper where he proposes an integrated objective Bayesian analysis based on an intrinsic discrepancy and suitably defined objective priors that should be used for both estimation and testing. This is a natural, principled, and unified treatment of the central problem in Bayesian statistics.

Integrated analysis. Let me begin by restating Bernardo's proposal using the same notation as he uses. In Section 2 he proposes choosing a loss function $\ell\{\theta_0, (\theta, \lambda)\}$ with posterior expectation $\bar{\ell}(\theta_0 | \mathbf{z})$. Then point estimates are the familiar minimum posterior risk estimates and credible regions are from the level sets of $\bar{\ell}(\theta | \mathbf{z})$ as a function of θ . That is, the $(1 - \alpha)$ credible region is of the form

$$V_\alpha = \{\theta | \bar{\ell}(\theta | \mathbf{z}) \leq u_\alpha\} \quad \text{where} \quad \Pr(V_\alpha | \mathbf{z}) = 1 - \alpha,$$

a lowest posterior risk region (in the posterior density) rather than a highest posterior density region. Analogously, for testing $H_0 \equiv \{\theta = \theta_0\}$, the rule is to reject the null when $\bar{\ell}(\theta_0 | \mathbf{z})$ is larger than a threshold value ℓ_0 . In both the testing and estimation settings, the loss is to be intrinsic and the prior is to be objective. The sort of analysis he suggests can, in principle, always be done: It is enough to specify a prior, likelihood, and loss.

In his earlier work, Bernardo extensively studied prior selection and more recently has studied loss function selection. I assume he would say that the likelihood comes from modelling the experiment generating the data. However, a reference likelihood can also be identified essentially from a loss function and prior via the rate distortion function in the information theory context, see Yuan and Clarke (1999). An obvious extension of that work can be used to generate a reference prior and a reference likelihood. (Given a prior, find the reference likelihood. Given the reference likelihood, find the reference prior. Cycle until convergence is obtained.) Taken together this would provide a complete, mostly objective, specification of the key ingredients in a Bayesian analysis starting from a loss function.

Thus, I see this paper as the natural conceptual completion of Bernardo's seminal contribution, reference priors. Recall that the idea of a reference prior is to find an objective prior that encapsulates our lack of knowledge. Thus, it is not necessarily to be used for inference but rather to provide a standard analysis with which the analyses generated by other priors might be compared. In the same spirit, the integrated analysis starting from a loss function, *i.e.*, the estimators and tests generated by the analysis a given loss function provides by using it to find a reference likelihood and reference prior, can be regarded as benchmark inferences with which inferences from other choices of loss functions can be compared.

Beyond integrated analysis. Now, let me raise two caveats to Bernardo's proposal. First, there is no concept of model bias, *i.e.*, likelihood mis-specification, and, second, there is no concept of robustness. Note that these concerns are motivated by the sort of validation involved in the prequential approach, see Dawid (1982).

My first thought for examining model bias is to take the data \mathbf{z} and form a non-parametric estimate of the density, say $\hat{q}(\cdot)$. Then choose

$$\lambda^* = \arg \min_{\lambda} d\{\hat{q}, p(\cdot | \theta^*, \lambda)\},$$

where d is a measure of distance on densities. If $d\{\hat{q}(\cdot), p(\cdot | \theta^*, \lambda^*)\}$ is large we know that there is no value of the nuisance parameter that makes the likelihood evaluated at θ^* , treated as a conditional density for \mathbf{z} , mimic a nonparametric estimate of the density of \mathbf{z} .

A second idea, more in keeping with decision theory based on ℓ , is to convert the notion of model mis-specification to a parametric problem. Suppose the parameter θ is embedded in a larger parameter $\phi = (\theta, \nu)$ where ν consists of some extra dimensions that can be used to describe a density. Suppose also that ℓ is extended to ϕ and assume that the density indexed by θ_0 in the initial statement of the problem corresponds to $\phi_0 = (\theta_0, \nu_0)$. Then, model mis-specification can be evaluated by seeing if the posterior risk meaningfully decreases due to the inclusion of ν . That is, we may surmise model-mis-specification if

$$\bar{\ell}(\theta_0 | \mathbf{z}) \gg \bar{\ell}(\theta_0, \nu_0 | \mathbf{z}).$$

Both of these suggestions introduce new models. However, Dawid (1984) argues convincingly that such alternatives must be examined.

My first thought on how to examine robustness to the inferences is to use a sort of errors-in-variables technique. A recent example of this can be found in Wang *et al.* (2009). The basic idea is to perturb each x_i by i.i.d. noise U_i so that a collection of data sets of the form $W_i = X_i + U_i$ for $i = 1, \dots, n$ is generated. These new data sets can be analysed in the same way as the original data set and a collection of estimates of θ can be found – one for each perturbation of the data. If the histogram of these estimates is too spread out relative to, say, $\text{Var}(U_i)$ then we have reason to doubt the usefulness of θ^* .

A second idea, more in keeping with the decision theoretic structure in Professor Bernardo's paper, is to call for stability of inferences under small changes to the loss function. This is most important because Professor Bernardo's theory essentially rests on correct specification of the loss. However, it is very hard, too.

A third sort of robustness of the inferences is the following. Let \mathbf{z}_i be the original data \mathbf{z} with x_i removed, *i.e.*, $\mathbf{z}_i = \{x_1, \dots, x_n\} \setminus \{x_i\}$. Then using the existing decision theoretic structure we can find n estimates of θ : $\theta^*(\mathbf{z}_1), \dots, \theta^*(\mathbf{z}_n)$. Thus we have n densities

$$\int p(\cdot | \theta^*(\mathbf{z}_i), \lambda) p(\lambda | \theta^*(\mathbf{z}_i)) d\lambda. \quad (1)$$

Let us generate new data $\mathbf{z}^* = (\mathbf{z}_1^*, \dots, \mathbf{z}_n^*)$ where \mathbf{z}_i^* is a random draw from (1). Now we can form a new estimate of θ , $\theta(\mathbf{z}^*)$ and ask how different it is from θ^* . That is we can calculate

$$d(\theta^*, \theta(\mathbf{z}^*)) = \int \ell\{\theta^*, (\theta(\mathbf{z}^*), \lambda)\} p(\lambda | \theta^*) d\lambda \quad (2)$$

to see how well the predictions \mathbf{z}_i^* replicate the inferential behaviour of the original data. Obviously, we do not expect (2) to be zero; it's size will be bounded below in terms of n and the spread of the likelihood.

Integrating integrated analysis into practice. The caveats above are merely that, caveats. So, I am confident there will be a goodly number of problems where Bernardo's integrated analysis can be used as is to get good results. However, I suspect there will be even more cases where his integrated analysis will serve as a useful benchmark for interpreting the results of another analysis that is actually advocated. Finally, I suggest that Bernardo's integrated analysis can be extended in ways that do not do much damage to his intent so as to provide a benchmark analysis for the frustratingly complex problems that most statisticians face today.

SUBHASHIS GHOSAL (*North Carolina State University, Raleigh, USA*)

First, let me congratulate the author for a very lucid description of his recent work on a unified method of prior selection for various inference problems. I fully agree with Bernardo's motivating idea that the choice of prior distribution should not be affected by the nature of the inference problem, either philosophically or operationally. The practice of using a drastically different prior in testing of point-null hypothesis compared to more natural looking priors used in estimation is inconvenient and has led to a lot of disagreement (the Jeffreys-Lindley paradox) among statisticians in the past. This sharply contrasts with the case of estimation, at least when the sample size is reasonably large, where one can often match frequentist and Bayesian procedures up to the second order. Bernardo has made a valuable contribution by bringing prior selection mechanism for widely different inference problems (point estimation, testing and region estimation) under the same umbrella.

Bernardo's concept of intrinsic loss is fundamentally important here. Although it has become less fashionable nowadays, Wald's decision theoretic formulation of inference problems is the most elegant and useful way of describing various issues related to inference. It is therefore not surprising that Bernardo's elegant solution for unification of objective prior is based on decision theoretic concepts. By letting the loss function be dependent only on the distance between probability measures (rather than on the corresponding parameters), Bernardo has made invariance under parameterization a non-issue.

Nevertheless, we still need to make some choices. The first one is the choice of the divergence measure $\delta(p, q)$ itself. Bernardo's choice of $\delta(p, q)$ is the minimum of the two possible Kullback-Leibler divergence measures $\kappa(p|q)$ and $\kappa(q|p)$. This is a clever choice since the Kullback-Leibler divergence is very "likelihood friendly", and further the minimum is symmetric, and is zero only when the two densities are equal. The triangle inequality is still elusive, but that does not appear to be an issue here. But there is a potential problem with this symmetrization of $\kappa(p|q)$. Unlike $\kappa(p|q)$, the measure $\delta(p, q)$ does not appear to be additive for product densities for general non-identically distributed case, although it is additive for all i.i.d. cases. This is because the minimum and summation operation may not be interchanged unless the ordering between $\kappa(p_i|q_i)$ and $\kappa(q_i|p_i)$ does not depend on i . This has potential consequences when dealing with non-i.i.d. data.

Another possible candidate for the divergence measure to be used to construct the intrinsic loss is given by the *negative log-affinity*, defined by $\rho(p, q) = -\log \int \sqrt{pq}$. It is easy to check that

- (i) $\rho(p, q) \geq 0$;
- (ii) $\rho(p, q) = 0$ only if $p = q$;
- (iii) $\rho(p, q) < \infty$ unless p and q have completely disjoint support;
- (iv) $\rho(\prod_{i=1}^n p_i, \prod_{i=1}^n q_i) = \sum_{i=1}^n \rho(p_i, q_i)$ always.

Note that property (iv) makes the measure completely “likelihood friendly” unlike $\delta(p, q)$, which is so only in the i.i.d. cases. It may be noted that property (iii) makes $\rho(p, q)$ more useful even in some i.i.d. cases like the $\text{Un}(|\theta-1, \theta+1)$ family, for which $\delta(p_\theta, p_{\theta'}) = \infty$ for all $\theta \neq \theta'$. It will be curious to see how the resulting analysis goes through when $\rho(p, q)$ replaces $\delta(p, q)$. At least in two cases, normal location and uniform scale families, $\rho(p, q)$ is equivalent to $\delta(p, q)$, but the former does not need two integral evaluations and taking their minimum. It is also useful to note that $\rho(p, q)$ has local quadratic nature similar to that of $\delta(p, q)$ (cf., Theorem 3).

Although it is a diversion of the topic, one may wonder about the notion of relative concentration of a density p compared to another density q . Bernardo called p more concentrated than q if $\kappa(p|q) < \kappa(q|p)$. This certainly appears to be intuitively acceptable for the uniform scale family. However, in general the concept does not appear to be transitive. This is, however, a common problem for notions defined through pairwise distances. The better known of this type is perhaps Pitman closeness, which also fails to be transitive.

However, the main issue in the proposed procedure appears to be calibration in hypothesis testing. Bernardo has recommended using a scale like $\log 10^k$, for $k = 1, 2, \dots$; it would be nice to make the calibration issue more formal since ultimate decisions will be based on the choice of the cut-off. This point seems to be also related to the sample size, since larger sample sizes are likely to make the likelihood ratios more extreme, and so the intrinsic loss as well. As the intrinsic loss is multiplied n -fold in the i.i.d. case, I would expect the presence of a factor of n in the cut-off point.

Finally, it will be interesting to formulate Bernardo’s decision making procedure for multiple hypothesis testing, which has received considerable attention recently because of genomic and fMRI applications.

MALAY GHOSH (*University of Florida, USA*)

It is a distinct pleasure and honour to contribute a discussion to the more recent article of Professor Bernardo on objective Bayesian estimation and hypothesis testing. Bernardo (1979) pathbreaking discussion paper has inspired many researchers, old and young, to further the development of objective priors. It is safe to say that the present article will also stimulate future research on objective priors and their role in estimation and testing.

As I see it, there are two main issues in this article. The first, a general definition of reference priors, has been addressed very adequately in Berger, Bernardo and Sun (2009). The second, the introduction of intrinsic loss, to unify estimation and testing is clearly a novel idea which bears promise for future development.

Like Robert and Bernardo, I am very much in favour of using intrinsic losses which measure the discrepancy between two densities rather than measuring the discrepancy between a parameter and its estimate. Indeed, inspired by an article of

George, Liang and Xu (2006), I wrote two articles with colleagues (Ghosh, Mergel and Datta, 2008; Ghosh and Mergel, 2009) to exhibit the Stein phenomenon under a very general intrinsic loss. In the remainder of my discussion, I will confine my comments to this particular aspect of the paper.

It appears that Bernardo's procedure may overcome one of the problems which Jeffreys encountered. Consider the $N(\mu, \sigma^2)$ distribution with both parameters unknown. The problem is to test $H_0 \equiv \{\mu = 0\}$ against the alternatives $H_1 \equiv \{\mu \neq 0\}$. With the prior $\pi(\mu, \sigma) \propto \sigma^{-1}$, which is ideal for point and set estimation, Jeffreys discovered a problem in the testing context described. Specifically, the Bayes factor of H_0 relative to H_1 tends to a positive constant rather than zero when the t -statistic goes to infinity. The problem disappears with the Cauchy prior. However, Bernardo's approximation for d in his Example 6 suggests that even with the prior $\pi(\mu, \sigma) \propto \sigma^{-1}$, the right inference can be done for the testing problem. My question is: how good is this approximation? Can the author elaborate more on this?

It should be noted that the Kullback-Leibler (KL) divergence is a special case of a more general power divergence class, considered for example in Cressie and Read (1984). Admittedly, KL is the most well-used measure. It appears though that many of these results will hold for the general power divergence class. Other than the KL, the Hellinger divergence is an important member within this class.

To see this, I considered the simple normal example, where $p(x|\theta) = N(x|\theta, 1)$. In this example, with KL loss, $l(\theta_0, \theta) = (1/2)(\theta - \theta_0)^2$. With the general power divergence loss

$$\left[1 - \int p^{1-\beta}(x|\theta) p^\beta(x|\theta_0) dx \right] / [\beta(1-\beta)],$$

the expression reduces to $[1 - \exp\{-\beta(1-\beta)(\theta - \theta_0)^2/2\}] / [\beta(1-\beta)]$ for this problem. This is monotonically increasing in $(1/2)(\theta - \theta_0)^2$. While this monotonicity may not prevail, something qualitatively similar should happen for the general exponential family or possibly even for nonregular families. I will appreciate the author's comments on this.

Professor Bernardo has promoted Bayesian statistics by holding the Valencia meetings for more than three decades. In those days when Bayesianism was not at all popular, it took a lot of courage and effort to find the resources to organize them. He earns a well-deserved rest after so many highly successful ventures. Even with his retirement from holding Valencia meetings, I do not expect him to retire from Bayesian statistics. I take this opportunity to toast for his long and productive career. Cheers!

MIGUEL GÓMEZ-VILLEGAS (*Universidad Complutense de Madrid, Spain*)

The author proposes a method to build, estimate and test hypothesis from a Bayesian point of view by using an objective prior and a measure of discrepancy. The author is to be congratulated on the way that he has overcome the difficulties associated with objective priors. Objective Bayesian methods are those which use a prior distribution which only depends on the assumed model and the quantity of interest. Thus, the combined use of a modified Kullback-Leibler discrepancy and an appropriately defined prior, provides an integrated Bayesian solution for both estimation and hypothesis testing problems.

As it is frequent when the paradigm of the decision theory is used every thing is clarified. This happens in Section 2 with the point and region estimation problems.

With respect to the precise hypothesis testing, I think, with Jeffreys, that if θ is a continuous parameter this forces the use of a non-regular “sharp” prior, concentrating a positive probability mass at θ_0 . I do not share the author’s opinion about this formulation leading to the difficulties associated to Lindley-Jeffreys paradox; I think that the problem arises with the use of a too high value of the positive probability mass, as pointed in Gómez-Villegas *et al.* (2009).

One question relative to intrinsic discrepancy may be asked. Is it not possible to simply use

$$\delta\{p_i, p_j\} = \kappa\{p_j|p_i\}$$

where $\kappa\{p_j|p_i\}$ is the Kullback-Leibler directed logarithmic divergence of p_j from p_i ? I think this is adequate when robustness is being considered. We have made use of this idea in the context of Bayesian networks in Gómez-Villegas *et al.* (2008).

It should be pointed out that the reference priors advocated by the author violate the likelihood principle, but it must immediately be admitted that this is the price to be paid to obtain an objective prior.

EDUARDO GUTIÉRREZ-PEÑA and RAÚL RUEDA
(IIMAS-UNAM, Mexico)

We would first like to congratulate Professor Bernardo for an interesting and clearly-written paper. We could not agree more with him concerning the need for a natural, integrated approach to Bayesian estimation and hypothesis testing. It is somewhat surprising that such an approach has not yet made its way into the mainstream Bayesian textbooks.

The paper contains a wealth of ideas and examples, but here we will only comment here on two aspects:

Invariance. Bernardo places quite a lot of emphasis on the invariance of the loss function. While we agree it is a nice property for a loss function to have, we do not think this is essential to achieve invariant Bayesian procedures. One can always obtain invariant estimators, even if the loss function used is not intrinsically invariant, provided that it is suitably defined.

In the case of the quadratic loss, for example, if θ is a one-dimensional parameter one can use the ‘normalizing’ transformation $\phi(\cdot)$ defined in Theorem 3 of the paper in order to find a suitable parameterization $\phi = \phi(\theta)$ upon which the quadratic loss is a more natural choice, so that

$$\ell\{\phi_0, \phi\} = (\phi_0 - \phi)^2.$$

If one now wishes to work in terms of θ or, for that matter, any other parameterization $\vartheta = \vartheta(\phi)$, it suffices to define the corresponding loss function as

$$\ell_{\vartheta}\{\vartheta_0, \vartheta\} = [\phi_{\vartheta}(\vartheta_0) - \phi_{\vartheta}(\vartheta)]^2,$$

where $\phi_{\vartheta}(\cdot)$ is the inverse of the transformation $\vartheta(\cdot)$.

Integration. The author does indeed provide an integrated, decision theoretical approach to point estimation, region estimation and *precise* hypothesis testing.

However, his *ad hoc* solution for the *compound* case $H_0 \equiv \{\theta \in \Theta_0\}$ does not seem to fit into his general framework.

We quote: “Thus, depending on the context, a compound hypothesis of the form $H_0 \equiv \{\theta_0 \in \Theta_0\}$ could be accepted when *at least one* of its elements would be accepted...” “...or when *all* its elements would be accepted...”

What is the loss function behind each of these criteria for testing compound hypotheses? Is either of these loss functions consistent with the loss function used for the other procedures? We would like to see the explicit form of the loss difference, $\Delta \ell_h = \ell_h\{a_0, (\theta, \lambda)\} - \ell_h\{a_1, (\theta, \lambda)\}$, whose expected value leads to either of the criteria suggested by Bernardo for the compound case. In our view, unless this loss can be exhibited and shown to be consistent with the intrinsic loss used elsewhere in the paper, this approach cannot be regarded as fully ‘integrated’.

In closing, we would like to express our gratitude to José-Miguel for his constant efforts over all these years in organizing the Valencia meetings. The impact of these on our discipline has been enormous.

ANGELIKA VAN DER LINDE (*University of Bremen, Germany*)

In this very last talk of the Valencia meetings, Professor Bernardo presented a diamond: more brilliant than ever, sparkling with new ideas, exhibiting many facets in terms of examples and being definitely invariant. He has spent much of his lifetime cutting and polishing it. We are stunned, we admire it, and we thank José for his passion, his inexorability and his continuous challenge for stimulating discussions with a long applause.

We acknowledge his efforts and his success in securing the foundations of Bayesian statistics while most of us are busy computing... Are we going to wear this diamond in everyday life? Sorry, this is an inadequate female question. Are we going to use this diamond as a statistical tool in everyday life as Bernardo suggests? That depends on how well we understand why the tool has been constructed as it is, which problems, fallacies and even failures in Bayesian analysis had an impact on its development. Bernardo hardly discusses alternative suggestions to overcome acknowledged difficulties but mainly summarizes the driving problems as lack of invariance. Important to me is a message inherent in his solution(s): base statistical inference on information theoretically founded decision theory. Bayesian statistics as applied probability theory has to incorporate entropy and information as basic concepts of probability theory.

More closely related to the talk I have two questions:

- (i) The reference prior is motivated as the prior maximizing the missing information about θ , and in classical examples yield estimators which are close to but more stable than the maximum likelihood estimator. Is there any idea (experience or expectation) about the performance in a (relatively) “small n , large p ” scenario?
- (ii) Bernardo emphasized that the same prior could be used for different summaries (estimation/testing) of the posterior distribution. What about the posterior predictive distribution?

In the end, all ends turn out to be beginnings. But diamonds are forever.

DENNIS V. LINDLEY (*Minehead, UK*)

In the 1970's when I was head of the Department of Statistics at University College London, a student from Spain was accepted to read for a Ph.D. On arriving in the department one day, my secretary made a dramatic entrance saying "Your Spanish student has arrived". He had already made his mark in the department and in the College, where parking rules had been infringed. When he and I met to discuss research topics, I suggested a problem that appeared to be difficult; difficult because over the years I had had several stabs at it without success. Not long afterwards he handed in a paper that purported to give a solution. It was taken home to read, together with a blue pencil, but to my amazement it contained a reasonable and ingenious solution. As far as I was concerned, he could have his Ph.D.

That student was José (Bernardo, 1979b) whose initial success was no flash in the pan but has been followed by a series of important papers, culminating in this one for the ninth in the influential series of Bayesian meetings for which he has been the guiding light. In it he presents a definitive statement of the objective Bayesian approach, developing priors and loss functions from the supposedly-objective probability model by sensible and carefully-argued mathematics. It is a triumph and statistics owes him a great debt for clarifying so many issues and producing usable results. A colleague of mine, interested in applications, dismissed the ideas as mathematical posturing. This is unfair because if the basic principles Bernardo proposes are accepted, the mathematical development can be ignored by the practitioner and the operational procedures adopted without much difficulty in this computer age. This is a paper that is valuable both for its theory and its practicality. José has shown me that he deserves, not just a Ph.D, but the highest award that statistics has to offer.

It is a pity that, despite my enthusiasm, I disagree with the development, just as I think that Fisher did brilliant work of the highest order yet, at the foundational level was wrong, for example over tail-area probabilities. My view is that the scientific method, and statistics as part of that method, is fundamentally subjective, objectivity only appearing when scientists reach agreement. I therefore argue in favour of statistical procedures that are based on subjective probabilities; probabilities that reflect your beliefs. My doubts begin with Bernardo's model; is it really objective, the same for all? There are several examples in the literature of data that have been analysed using different models, but my objections go deeper than that.

We recognize two aspects to statistics, inference and decision-making. Some statisticians, following the founders of the Royal Statistical Society and, more recently, Fisher, have held that our remit is the collection and analysis of data, not its use in determining action. The subjective attitude includes the decision aspect, if only because the ideas easily lead to a recipe for action, namely the maximization of expected utility. Indeed, many justifications for the Bayesian position start from the decision aspect. On the practical side, what use is inference if it cannot be used as a guide to action? Bernardo treats inference as a decision process, producing his loss from the model. I presume that if the data in the model were to be used as a basis for action, that action would determine the loss, or utility, structure and his suggestion abandoned. I interpret his loss function as that needed for inference only; yet inference does not need a decision element but can be restricted to a statement of your probability distribution for Θ given the data. Admittedly that probability structure may be hard to comprehend and some approximation used, but does approximation call for the paraphernalia of loss? There are other objections and an

important one for me is the violation of the likelihood principle; a principle that is basic to the Bayesian method. (And to frequentist statistics, though usually unrecognized there.) This violation occurs as soon as an integration takes place over sample space \mathcal{Z} , since the principle says the elements of that space, apart from the data observed, are irrelevant. Definition 4 provides the first use of this banned operation. It would be interesting to see the application of objective Bayes to sequential analysis where frequentists usually violate the principle; for example where a sample of predetermined size n is distinguished from one in which n is random, so that the two sample spaces, and hence the models, differ. Within the objective view, does this make a difference?

Another difficulty for me lies in the use of the expectation operation when loss is introduced, and even in the concept of loss itself. The most satisfactory development of Bayesian concepts uses the notion of utility for outcomes, rather than losses, the latter presumably being the difference of two utilities, as suggested in §2.3. In this development the expectation operation can be justified if utility is itself measured on the scale of probability: if 0 and 1 are your utilities for a bad and a good outcome respectively, then an intermediate outcome E has utility u for you if you are indifferent between E for sure and a good outcome with probability u (and bad otherwise). Perhaps the paper lacks a clear explanation of loss and a justification for the sole use of expectation. These points are developed in chapter 10 of my book Lindley (2006). Similar doubts may also be expressed about the use of supremum in Definition 3, for it is often dangerous to replace a function by a number, with possible damage to one's understanding.

Example 7 was naturally of special interest to me. The analysis by the objective method is impressive and the outcome seems right. My analysis would have replaced the concentration of prior probability $\theta = 1/2$, which I regard as an approximation to reality, by a prior centred there and with small variance. The difficulty then is, how large is small? To answer that one has to go back to the design of the experiment, including the construction of the random mechanism. We know little about the construction of personal probabilities and this topic should be an important area for research. I am perhaps overfond of quoting de Finetti's question to me: "Why do statisticians always talk about Greek letters?". Yet its relevance is apparent here as in Example 5 where the classic, practical case involves the number θ of tramcars in a town and you see tramcar number x . The reference prior θ^{-1} , ignoring the discrete element, is unsatisfactory, for would you really think the town most likely had just one tramcar? The objective development arises because of the addiction to the Greek alphabet. In practice θ is merely the representative of something real, here tramcars, and reality should not be forgotten.

The objective approach to inference is like an exploration to find a pass through the mountains, failed to find it, but made wonderful discoveries on the way that were very useful to those who ultimately reached the pass.

MANUEL MENDOZA (*Instituto Tecnológico Autónomo de México, Mexico*)

As it often happens when listening a talk by Professor Bernardo, this presentation is so full of clever ideas, concepts and results that it is rather difficult to capture the global scenario behind all this work. If, in addition, the subject is one as controversial as the idea of an *objective* Bayesian analysis, the manuscript must be read at least twice to state clearly some of the questions that emerge from the very beginning. Let me start by recalling some specific assertions in the paper.

In Section 2.2, Bernardo says: Bayesian estimators are usually not invariant under one-to-one transformations', and remind us that, under quadratic loss, the Bayesian estimator of the variance is not the square of the Bayes estimator of the standard deviation. This assertion is not exactly true. Let us suppose that in a decision problem, and in accordance to the axiomatic foundations, you have elicited the loss function $L(d, \theta)$ where $d \in \mathcal{D}$ and $\theta \in \Theta$. Thus, L is defined as $L : \mathcal{D} \times \Theta \rightarrow \mathfrak{R}$. Now, if you relabel the action space so that $d' = g(d)$ where $g : \mathcal{D} \rightarrow \mathcal{D}'$ is a one-to-one function and a similar treatment is given to θ so that $\phi = h(\theta)$ where $h : \Theta \rightarrow \Phi$ is also a one-to-one function, then the loss function L , uniquely defined up to linear transformations, can be expressed in terms of d' and ϕ , as $L(g^{-1}(d'), (h^{-1}(\phi))) = L'(d', \phi)$ where $L' : \mathcal{D}' \times \Phi \rightarrow \mathfrak{R}$. Now if we get d'_* , the Bayesian solution to this problem in terms of the new labelling (and $L'(d', \phi)$), it follows that $d'_* = g(d_*)$ where d_* is the original Bayesian solution with $L(d, \theta)$. Thus, the Bayesian solution is invariant under one-to-one transformations and so is Bayesian pointwise estimation. The point here is that $L(\hat{\sigma}^2, \sigma^2) = (\hat{\sigma}^2 - \sigma^2)^2$ and $L(\hat{\sigma}, \sigma) = (\hat{\sigma} - \sigma)^2$ are different loss functions and thus, lead to different solutions. The coherent use of a quadratic loss requires the selection of the specific labelling for which the quadratic function describes our preferences. If this labelling is the standard deviation, then if we change to the variance, the appropriate expression for the same loss function is $L(\hat{\sigma}^2, \sigma^2) = (\sqrt{\hat{\sigma}^2} - \sqrt{\sigma^2})^2$. The same idea can be used to prove that coherent Bayesian credible intervals are also invariant (Section 2.3). All you have to do is to choose the labelling for which minimum length is desired. In any case, I think that a note should be introduced to clearly distinguish invariant Bayesian decisions from invariant loss functions, as discussed by Bernardo. Obviously, these concepts are related. If we use an invariant loss function, then the expressions for the loss function corresponding to different labellings have the same functional form.

In Section 2.3, when discussing the hypothesis testing problem in the compound case, we can read: Thus, depending on the context, a compound hypothesis of the form $H_0 \equiv \{\theta_0 \in \Theta_0\}$ could be accepted when *at least one* of its elements would be accepted, so that $\inf_{\theta_0 \in \Theta_0} \bar{\ell}(\theta_0 | z) < \ell_0$, or when *all* its elements would be accepted, so that $\sup_{\theta_0 \in \Theta_0} \bar{\ell}(\theta_0 | z) < \ell_0$ '. This looks like a minimax-type criterion and I would like to see how this can be derived from a loss structure $(L\{a_0, (\theta, \lambda)\}, L\{a_1, (\theta, \lambda)\})$.

In Section 3.1 (Example 2), it is stated: ...both the entropy loss and the standardized quadratic loss penalize far more severely overestimation than underestimation, and will both yield too small estimates for the variance'. In the quest for a loss function which would be one of the components of the *objective* Bayesian analysis, invariance might be considered a useful property because of the technical simplifications it entails. On the other hand, the requirement of symmetry is a little more difficult to understand. Is the author implying that a case where underestimation is preferred to overestimation cannot be handled by means of an *objective* Bayesian analysis?

In more general level, my personal feeling is that Bernardo has presented us with his *subjective* version of what an *objective* Bayesian analysis should be. In constructing his proposal he has made a number of decisions. Some of them will be shared by many of us, but I guess that in some cases, alternative formulations could be proposed. In any case, I think that this is a nice paper with many stimulating ideas and enlightening examples which may lead us to a fruitful debate on the future of Bayesian analysis.

ELÍAS MORENO (*Universidad de Granada, Spain*)

While we should acknowledge the efforts of Professor Bernardo to put together both Bayesian estimation and hypothesis testing, my position is that they are different problems that require different statistical tools. In particular, different prior distributions will be generally needed when more than one model is involved in the problem. At the very beginning of the paper the author proposes the following definition “*Objective Bayesian methods are defined as those which use a prior distribution which only depends on the assumed model and the quantity of interest*”. Using this definition it can be argued that since in estimation and testing the quantities of interest and the models are different the goal of the paper of unifying objective Bayesian parameter estimation and hypothesis testing seems to be, in general, unattainable.

In testing problems the quantity of interest is a discrete set of competing models, that for simplicity we assume it contains only two models. This has the nature of a decision problem on the model space $\{\mathcal{M}_i, i = 0, 1\}$, where model \mathcal{M}_i consists of a set of sampling models $\{f_i(x_i | \theta_i, \mathcal{M}_i), \theta_i \in \Theta_i\}$, the prior structure has the form $\pi_i(\theta_i, \mathcal{M}_i) = \pi_i(\theta_i | \mathcal{M}_i) \Pr(\mathcal{M}_i)$, and the decision space is $\{d_i, i = 0, 1\}$, where d_i is the decision of choosing model \mathcal{M}_i . To complete the formulation of the decision problem we need the function $\ell(d_i, \mathcal{M}_j)$, the loss which corresponds to making the decision d_i when the true model is \mathcal{M}_j . We note that the loss function in the paper is not defined in the product space $\{d_0, d_1\} \times \{\mathcal{M}_0, \mathcal{M}_1\}$, even when the decision problem is that of choosing between the models \mathcal{M}_0 and \mathcal{M}_1 .

When we want to minimize the proportion of times we make a wrong decision the 0–1 loss function is an appropriate one. It is useful, for instance, in cost-effectiveness analysis where transfers of health between patients is judged to be not reasonable (Moreno *et al.* 2010). For this loss function, and the data $\mathbf{z} = (x_1, \dots, x_n)$, assuming they come from either a sampling model in the \mathcal{M}_0 or in \mathcal{M}_1 , the optimal decision is that of choosing model \mathcal{M}_0 if its posterior probability is such that $\Pr(\mathcal{M}_0 | z) \geq \Pr(\mathcal{M}_1 | z)$, or equivalently

$$BF_{10}(z) \leq \frac{\Pr(\mathcal{M}_0)}{\Pr(\mathcal{M}_1)},$$

where $BF_{10}(z) = \int f_1(z | \theta_1, \mathcal{M}_1) \pi_1(\theta_1 | \mathcal{M}_1) d\theta_1 / \int f_0(z | \theta_0, \mathcal{M}_0) \pi_0(\theta_0 | \mathcal{M}_0) d\theta_0$ is the Bayes factor to compare \mathcal{M}_0 and \mathcal{M}_1 . An important particular example is that of testing a sharp null hypothesis of the form $H_0 \equiv \{\theta = \theta_0\}$, the case where Bernardo is more critical with the Bayes factors. In this case the sampling model f_0 is nested in f_1 . In the paper it is argued that for this problem, in which we necessarily have $\pi_0(\theta | \mathcal{M}_0) = \delta_{\theta_0}(\theta)$, the prior π_0 is polemic. I do not see why it is polemic under the above scheme. Also it is asserted that “*Moreover, this formulation is also known to lead to the difficulties associated to the Lindley’s paradox*”. It is not so easy, and the question is whether Lindley’s paradox conveys such a message.

Let us revise the paradox. As far as I know it was originally described when $f_0(x | \mathcal{M}_0) = N(x | 0, 1)$, $f_1(x | \theta, \mathcal{M}_1) = N(x | \theta, 1)$ and $\pi(\theta | \mathcal{M}_1) = N(\theta | 0, \tau)$, and it is desired to choose one of the models based on a sample $\mathbf{z} = \{x_1, \dots, x_n\}$. It is then easy to see that

$$BF_{10}(\bar{x}, n, \tau) = \frac{1}{(n\tau^2 + 1)^{1/2}} \exp \left\{ \frac{n \bar{x}^2}{2} \frac{n \tau^2}{n\tau^2 + 1} \right\},$$

where \bar{x} is the sample mean. When $\tau \rightarrow \infty$ we have that $B_{10}(\bar{x}, n, \tau) \rightarrow 0$; that is, whatever the sample mean \bar{x} the model \mathcal{M}_0 is always chosen. This is seen to be paradoxical, and is called the Lindley's paradox. But we note that the prior $N(\theta|0, \tau)$ degenerates to zero when $\tau \rightarrow \infty$ (Robert 1993). Therefore, there is nothing paradoxical but simply that the prior for the alternative model is such that in the limit there is only one model to be chosen, the model \mathcal{M}_0 . By the way, we recall that for any fixed value of τ the Bayes factor $B_{10}(\bar{x}, n, \tau)$ is consistent, that is when sampling from \mathcal{M}_0 the Bayes factor tends to zero as the sample size grows, and when sampling from \mathcal{M}_1 its tend to ∞ .

Certainly, similar "paradoxical" situations can be reproduced for other sampling distributions; for instance, Example 7 in Section 5 of the paper, where a binomial sampling model $\text{Bi}(x|\theta, n)$ is considered and a sample observation such that $x/n = 0.500177$ for n as large as 104,900,000 is available, and we want to test the null $\theta = 0.5$ vs. $\theta \in (0, 1)$. We first note that in a small neighbourhood of the point 0.500177 the likelihood is huge compared with the likelihood outside of this neighbourhood, for instance $L(0.500177) = 685L(0.5)$. If the prior on the alternative sampling models spreads out the mass in the interval $(0, 1)$, most of the probability mass is put on the region where the likelihood is extremely small, and the resulting likelihood of the alternative model will be very small compared with the likelihood of the null. Consequently, the null will be chosen.

This situation is exactly what is provoked by the use of either the uniform or Jeffreys prior. These priors, which do not depend on the null, are not appropriate for testing problems since they do not concentrate mass around the null; that is, they do not satisfy the Savage continuity condition (Jeffreys, 1961, Ch. 5; Gnel and Dickey, 1974; Berger and Sellke, 1987; Casella and Berger, 1987; Morris, 1987a,b; Berger, 1994). When the prior concentrates its mass around the null hypothesis, as the intrinsic priors do with a degree of concentration controlled by the training sample size, the resulting likelihood of the alternative model will be a much more serious competitor of the null likelihood, and in this case the null can be rejected. For the notion of intrinsic priors see Berger and Pericchi (1996) and Moreno *et al.* (1998), and for an analysis of the concentration of the intrinsic prior of the alternative model on the null and its implication on robustness in testing problems see Consonni and La Rocca (2008), and Casella and Moreno (2009).

Therefore, in my opinion the arguments against the Bayes factors and model posterior probabilities in the paper are not convincing. After all, when the alternative is either empty or constructed as a mixture of models having a extremely small likelihood, to accept the null hypothesis is the right thing to do, a behaviour that is not paradoxical. I am not sure that the integrated objective Bayesian method presented in this paper by the author is a general improvement over the current objective Bayesian methods for hypothesis testing based on Bayes factors and model posterior probabilities.

Finally, I would like to thank Professor Bernardo for the organization of the Valencia meetings that have served as a source of stimulus for so many statisticians. Congratulations for that, Jos-Miguel.

CHRISTIAN P. ROBERT and JUDITH ROUSSEAU
(*Universit Paris-Dauphine, France*)

In this discussion, we congratulate Professor Bernardo for his all-encompassing perspective on intrinsic inference and focus on the case of nuisance parameters.

Unified inference. The paper manages the *tour de force* of aggregating intrinsic loss functions with intrinsic (*aka* reference) priors. Thus, Bernardo presents us with a unified picture of Bayesian analysis as he sees it and it is obviously fitting to see this cohesive perspective appearing in the Valencia 9 proceedings as a kind of third unification! We appreciated very much the paper and our comments will thus concentrate on minor issues rather than on the big picture, since we mostly agree with it. Although the tendency in Bayesian analysis, along the years, and in particular in the Valencia proceedings (see, e.g., Polson and Scott in this volume who discuss shrinkage without a loss function), has been to shy away from the decision-theoretic perspective (see, e.g., Gelman (2008), it is worth reenacting this approach to the field, both because it sustains to a large extent the validation of a Bayesian analysis, and because it avoids the deterioration of its scope into a mechanical data analysis tool.

Down with point masses! The requirement that one uses a point mass as a prior when testing for point null hypotheses is always an embarrassment and often a cause of misunderstanding in our classes. Rephrasing the decision to pick the simpler model as the result of a larger advantage is thus much more likely to convince our students. What matters in pointwise hypothesis testing is not whether or not $\theta = \theta_0$ holds but what the consequences of a wrong decision are. Of course, there is a caveat in the reformulation of Professor Bernardo, which is that, in the event the null hypothesis $\theta = \theta_0$ is accepted, one has to act with the model \mathcal{M}_0 . One can of course assume that, given the model \mathcal{M}_0 , the intrinsic Bayesian statistician would start from the reference prior for \mathcal{M}_0 , but this involves a dual definition of the prior for the *same* problem that remains a bit of an itch...

The case of compound hypotheses is only half-convincing in that the “natural” solution would seem to us to compare the posterior expected losses under both models, rather than singling out H_0 in a most unbalanced and unBayesian way. We actually take issue with the repeated use of infima in the definition of loss functions.

Intrinsic losses. Most obviously, we welcome the recentering of objective Bayes analyses around the intrinsic losses we developed in Robert (1996). Note that the severe lack of invariance of HPD regions was further studied in Druilhet and Marin (2007), while integrating point estimation losses in the evaluation of credible regions was proposed in Robert and Casella (1994).

The handling of nuisance parameters always is a ... nuisance, so Definition 5 is a possible solution to this nuisance. While it shies away from using the unsatisfactory argument of λ being “common” to both models, one of us (CPR) somehow dislikes the introduction of the infimum over all values of λ_0 : a more agreeable alternative would be to integrate over the λ_0 's, using for instance an intrinsic prior $\pi(\lambda|\theta_0)$. We however acknowledge the relevance of projections in model comparisons, as illustrated by Robert and Rousseau (2002).

Another issue deals with cases when the nuisance parameter is ill-defined under the null hypothesis, as for instance in our favourite example of mixtures of distributions Titterton, Smith and Makov (1985), Maclachlan and Peel (2000): When the null has several possible representations, the nuisance parameter varies from one representation to the next. A connected issue is the case when the parameter of interest is a function (functional) of the whole parameter vector that is such that there is no explicit way of breaking the whole parameter into a parameter of interest and a nuisance parameter, a setting that typically occurs in semi-parametric problems. Although a natural extension to Bernardo's approach is to define the intrinsic loss

between the parameter $\theta = \theta(f)$ and θ_0 as

$$\delta(\theta_0, f) = \inf\{\min(k(f | f_0), k(f_0 | f)); f_0 \in \mathcal{F} \text{ satisfies } \theta(f_0) = \theta_0\}$$

such an approach seems impossible to implement in practice, even in simple semi-parametric problems.

When replacing regular testing with checking whether or not the new type of regret $\ell\{\theta_0, (\theta, \lambda)\} - \ell_0$ is positive, the so-called *context dependent positive constant* ℓ_0 is equal to

$$\int_{\Theta} \int_{\Lambda} \ell_h\{a_1, (\theta, \lambda)\} p(\theta, \lambda | \mathbf{z}) d\theta d\lambda$$

in the original formulation. We therefore wonder why the special values $\ell_0 = \log 10^k$ for $k = 1, 2, 3, \dots$, are of particular interest compared, say, with $\ell_0 = \log \sqrt{\pi}^k$ or $\ell_0 = \log e^k \dots$. The calibration of ℓ_0 suffers from the same difficulty as the calibration of Bayes factors in that the choice of the decision boundary between acceptance and rejection is not based on a loss function. In particular, it is surprising that, in a *objective* context, ℓ_0 does not depend on the number of observations. Typically, the Kullback–Leibler divergence between the densities f_θ and $f_{\theta'}$ associated with n (not necessarily i.i.d) observations increases with n . Should ℓ_0 be rescaled as $n\ell_0$ and is such a scaling appropriate in general? We argue that rescaling by n as such is as *arbitrary* as considering Jeffreys prior as default prior.

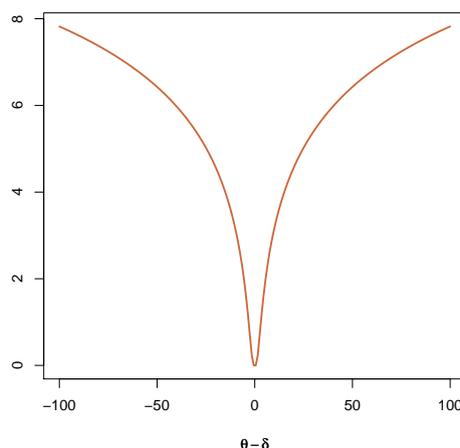


Figure 8: Kullback–Leibler loss function $\ell(\theta, \delta)$ associated with a Cauchy distribution with location parameter θ .

A last point of interest to us is whether or not an integrated reference analysis is always possible. Bypassing the issue of finding a reference prior, we wonder if there exist settings where the posterior Kullback–Leibler loss is *uniformly* infinite, thus preventing the choice of a Bayes estimator. For instance, when observing a Cauchy variate x , the intrinsic loss is of the form represented in Figure 8. Since the posterior under the flat prior is a Cauchy distribution with location parameter x , the loss may be increasing too fast for the Bayes estimator to exist

A family of models where the Kullback–Leibler loss cannot be applied corresponds to cases where the densities have supports that depend on the parameters in a non-trivial way, i.e.

$$f_\theta(x) = \mathbb{I}_{L(\theta)} g_\theta(x), \quad \text{where} \quad L(\theta) \cap L(\theta')^c \neq \emptyset \quad \text{and} \quad L(\theta') \cap L(\theta)^c \neq \emptyset$$

and $g_\theta(x) > 0$ everywhere.

In conclusion, our point here is to emphasize that, although the Kullback–Leibler loss has compelling features such as additivity, it also suffers from drawbacks, related to the requirement of comparing absolutely continuous distributions (one way or the other) and to its unboundedness. Some other *natural* intrinsic losses could be considered, in particular the Hellinger distance (Robert, 1996). How would both losses compare and what would their relative merits be? It seems to us that the *natural calibrations* found in Bernardo’s proposal could not be used with an Hellinger loss. Now, could that be such a bad thing...?!

Reference priors. Although we essentially agree with most of the construction of reference priors, we are doubtful about the systematic use of repeated (identically and independently) data sets. Indeed, in cases where the observations are modelled as a dependent process, say a time series, a part of the parameter vector addresses the dependence structure. Then, first, repeated i.i.d. sampling from the model will not provide useful knowledge about these parameters, since they can only be inferred correctly by letting the sample size increase to infinity. Second, for a fixed sample size, the Fisher information matrix depends in a non-trivial way on n and it usually has a non-explicit representation. Therefore, the reference prior under repeated sampling does not have an interesting formulation. For instance, when sampling from a stationary Gaussian process with spectral density f_θ , the Fisher information matrix associated with the covariance matrix includes terms of the form

$$\text{tr} \left[(T_n(f_\theta))^{-1} T_n(\nabla f_\theta)^2 \right],$$

where $T_n(f)$ is the n dimensional Toeplitz matrix associated with the function f and ∇f_θ is the first derivative of the spectral density, see Philippe and Rousseau (2003). This expression is not *user-friendly*, to say the least!, whereas the reference prior—obtained by letting the sample size go to infinity—actually corresponds to the limit of the above terms:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} (\nabla \log f_\theta)^2(x) dx$$

which are much more satisfactory for the construction of a prior distribution. The latter can also be obtained by considering the limit of the reference priors as n goes to infinity, however it is not clear whether it should be interpreted as the reference directly obtained from increasing n in the sampling or as the limit of Professor Bernardo’s reference prior when n goes to infinity. These two approaches might indeed lead to quite different results, as illustrated by non-stationary models.

NOZER SINGPURWALLA (*The George Washington University, USA*)

Professor Bernardo is one among a handful of skilled researchers who work on the methodological foundations of Bayesian statistics. Regrettably, this handful

seems to be dwindling, and thus papers like this that provide a summarization of recent work in the arena of inferential methodology are of archival value.

I found this paper demanding to read, and having read it, challenging to digest. All the same, I was amply rewarded by exposure to a wealth of material, and for his contribution to my learning, I thank José.

Now that the obligatory niceties which discussants are expected to bestow on an author have been dispensed, I will put forth my thoughts on the edifice that Bernardo and his coauthors have created.

General comments: Tempest in a teapot. My claim that this paper is demanding to read is based on the feeling that the paper exposit its material in a pedantic style that detracts from essentials. For example, the introduction of the nuisance parameter λ , tends to be a nuisance when it comes to focus. Similarly, the introduction of the parameter $\omega \in \Omega$, and then a function of ω , namely $\omega(\theta) \in \Theta$, are unnecessary.

The material in the paper is challenging to digest because it subscribes to the hierarchical and nested writing style of de Finetti. But de Finetti's essential thesis, namely, a categorical rejection of the focus on parameters, has been forsaken. Are parameters not just "Greek symbols" whose sole purpose, at least from a Bayesian perspective, is to *mechanize* the process of predicting observables by imparting on them the property of (conditional) independence? If such a point of view were to be adopted, then the entire enterprise of reference priors, parameter estimation and hypothesis testing, boils down to a mathematical exercise - and I do like mathematics!

Engineers and scientists could not care less about properties of unobservable parameters. They are interested in predicting and in controlling observables, a point of view that has been made before by several others. Of course, this viewpoint does not obviate the need for loss functions and prior distributions [cf. Singpurwalla (2006)]. It simply says that priors and losses should be functions of observables, not parameters. Indeed Bayes assigned a prior distribution on outcomes [cf. Stigler (1982)], the prior on parameters being the handiwork of Laplace, whose genius it was to interpret the propensity parameter of Bernoulli trials as the *cause* of the observed observables [cf. Singpurwalla (2002b)]. Therefore, with some trepidation, I call upon the author to develop a mechanism for prediction and control, that is fully Bayesian in character, but with priors on observables that can be claimed to be objective, in some definable sense. Some preliminary ideas along the above lines, albeit without any claims of objectivity, are in Singpurwalla (2002a). My conjecture is that in pursuing such a path, many of the hurdles faced by Bernardo will vanish.

Thesis of the paper: Where is the intuition? I am in full agreement with the driving thesis of this paper that estimation and hypothesis testing should be decision theoretic, and that normative decision making is based on two pillars and one principle. The pillars are (prior) probability and utility, and the principle is the maximization of expected utility. However, the devil is in the details, and this is where the essence of the paper comes into play. As pointed out by the author, the existing *modus operandi* in Bayesian inference is to use two radically different kinds of priors on the same parameter, say ω ; one for estimation and one for hypothesis testing. This may somehow seem unattractive, even though estimation is for a different set of actions than testing hypothesis. If a parameter has a physical meaning

(such as the limit of a relative frequency in Bernoulli trials) then the claim is that in the eyes of a single assessor, there should be one and only one prior for ω .

Having advocated the use of reference priors for estimation, Bernardo is left with but one choice to achieve his goal of using the same prior for hypothesis testing as well. The choice is to tinker with the utility (loss) function. This is done admirably well by introducing an *intrinsic loss function*, and then producing an impressive array of inspiring examples.

The reference prior and the intrinsic loss function share a common mathematical foundation, namely, the Kullback-Leibler measure of divergence, and the manner in which this measure is invoked is ingenious. In the former case it is the discrepancy between the joint $p(x, \omega)$ and the product of the marginals $p(x)p(\omega)$. In the latter case it is via the introduction of the notion of an *intrinsic discrepancy*, which for two distributions p_1 and p_2 is the minimum of the Kullback-Leibler divergence of p_1 from p_2 , and of p_2 from p_1 ; see Definition 4. The intrinsic loss function is based on a minimization of the intrinsic discrepancy; see Definition 5.

With the above as a methodological architecture, we see a menu of examples, each exhibiting attractive features, leading sceptics to conclude that the ends justify the means.

Personally, I find the Bayesian argument attractive because of: its completeness due to a firm grounding in the mathematics of probability; its coherence via an avoidance of a sure loss, and most important its scientific merit vis à vis allowing the incorporation of information generated by honest beliefs, and/or the physics of the problem. Thus when endowing priors to observables or to parameters, the Bayesian should act like a scientist by gaining a sound appreciation of the problem, and then proceed without leaning on the use of automated priors. This philosophical viewpoint has been voiced before; to paraphrase the late Dev Basu “you have no business working on a problem you don’t understand.” Similarly, with the utilities; they should be grounded in the economics of the decision making environment.

But suppose one were to accept (on pragmatics) the philosophy of using automated priors and utilities. Then one should still have a believable basis for proposing these. The reference prior seems to have the intuitive import, in that it is guided by the principle of *let the prior do the talking and the data do the walking*. What is unclear to me is the motivation behind the notion of the intrinsic discrepancy, the notion which gives birth to the intrinsic loss. A clearer justification of this would allay my concerns, and despite reservations about the enterprise, will go a long way towards a “buy in”.

Minor quibbles. (i) Figure 1 seems to me like a 2-edged sword. It really demonstrates the disadvantages of choosing stylized loss functions. For example, in the context of engineering design, underestimating the variance could lead to designs that are risk prone, whereas overestimating the variance can result in designs with a large margin of safety. In the former case l_δ is attractive; and in the latter case l_{quad} is attractive. A loss function which encapsulates a trade-off between cost and safety appears to be a hybrid between l_δ and l_{quad} . All this goes to suggest that the appropriateness of a loss function should be context dependent.

(ii) The material of Example 4 is fascinating. Why should non-dependence of the posterior on m be viewed as a virtue? When $a = 1/2$ and $m = 2$, the two posteriors are identical. The Dirichlet based posterior offers more generality.

(iii) The material of Example 5, with $\theta^* = 2^{1/n}t$ and $(t, t(1-q)^{1/n})$ as the q -credible region, suggests the possibility of exploring an inference mechanism for the threshold parameter of failure models. Threshold parameters are proxies for minimum life and are useful for setting warranties and insurance premiums. On the matter of the example itself, the rationale behind choosing l_0 is unclear, and the expected loss linearly increasing in θ_0 bothersome. A diminishing marginal loss function (of θ_0) seems meaningful.

(iv) The discussion following Example 7 pertaining to ESP is paradoxical. Here we have one of the world's foremost Bayesians leaning on p -values as a yardstick for calibrating his work, and rejecting Jefferys' (not Sir Harold Jeffreys - the e and r are transposed) standard Bayesian approach as an example of Lindley's paradox. It seems we have come around a full circle. *Paradox I win, paradox you lose*. My physics colleagues will say that with 52,263,471 successes in 104,490,000 trials, the raw propensity of success is .50017677, and you do not need a statistician (Bayesian or frequentist) to accept the hypothesis that $\theta_0 = .5$. Besides, what is the point of testing such a hypothesis?

To conclude, I thank José for his years of friendship, his kindness, and his invitation to comment on this paper.

JAN SPRENGER (*Tilburg University, The Netherlands*)

In his contribution, Professor Bernardo presents a variety of results on objective Bayesian inference in the classical problems of parameter estimation and hypothesis testing. A main motivation for these developments, and in particular for the use of intrinsic loss functions, is to avoid results that vary with the chosen parametrization. Such results are, he says, "difficult to explain when, as it is the case in theoretical inference, one merely wishes to report an estimate of some quantity of interest".

This statement has a striking resemblance to Ronald A. Fisher's reservations with respect to a decision-theoretic approach in statistical inference:

"In the field of pure research no assessment of the cost of wrong conclusions [...] can conceivably be more than a pretence, and in any case such an assessment would be inadmissible and irrelevant in judging the state of the scientific evidence". (Fisher 1935, 25-26)

Although Bernardo has no *principal* objections to a decision-theoretic perspective in pure statistical inference (Bernardo 1999), he agrees with Fisher that pure scientific inference demands different methods than applied inference. This becomes clear in the case of hypothesis testing where, according to Bernardo, scientists frequently ask the question of whether a particular parameter value $\theta = \theta_0$ is "compatible with the data". This question makes perfect sense for a frequentist objectivist like Fisher, but from a Bayesian perspective, it sounds somewhat odd and incomplete. What counts as "compatible"? Doesn't our standard for judging compatibility depend on how plausible we consider the alternatives, those close to θ_0 and those far from θ_0 (see also Berger and Delampady 1987)? In other words, I believe the idea of an objective, context-independent discrepancy measure between model and data to be a frequentist chimera that we should be careful to invoke.

Moreover, the intrinsic loss function that Bernardo suggests is in principle unbounded, making it conceptually inappropriate for a variety of estimation problems, including the ones from theoretical science that Bernardo has in mind. I believe that

the justification for using such functions should be *practical*, not foundational, coming from our ignorance of the structure of the true loss function, and the convenient mathematical properties that they possess. Indeed, some of Bernardo's writings (e.g., the reply to Lindley in his 1999 paper) indicate that intrinsic losses and reference priors should not be understood as a challenge to Bayesian subjectivism, but as conventional or default choices, and as a form of sensitivity analysis.

Fisher's program of developing a logic of objective scientific inference has long been pursued by frequentists, but, as we (Bayesians) know, without success. The approach by Bernardo is arguably our best attempt to revive this program from a Bayesian perspective, but it needs to be aware of the intrinsic tension in the program itself.

FRANK TUYL (*University of Newcastle, Australia*)

While Professor Bernardo has proposed an attractive integrated framework, I suggest that certain aspects will not impress our frequentist colleagues, nor all Bayesians. I would like to discuss two limitations, one of the proposed region estimation and one of reference priors in general.

Region estimation. Previously, Bernardo has emphasised the excellent sampling properties of credible intervals based on reference posteriors. However, central intervals tend to lead to zero coverage near parameter extremes, and HPD intervals do not always fix this problem—which it certainly is from a frequentist point of view. While intrinsic intervals avoid the lack of invariance of HPD intervals, they potentially share the central interval's zero minimum coverage *even when the HPD interval avoids it*.

First consider Figure 3. As pointed out by Bernardo, here *any* credible interval is also a confidence interval, so that excluding, for larger values of α that lead to a lower limit $l > 1.71$, a sliver to the right of $\theta = 1.71$ has no effect on frequentist coverage. However, to many frequentists and Bayesians, such an interval appears inferior to the HPD interval, which is also the short confidence interval: the *wider* intrinsic interval sacrifices values with high likelihood for values with lower likelihood.

Second, in the very similar Figure 7 of Bernardo (2005b) based on binomial data $x = 0$ and $n = 10$, and with left limit 0 instead of 1.71, is more serious. Due to the binomial model's lack of a pivotal quantity, coverage now varies with θ ; here, HPD intervals are clearly preferable to central intervals as they avoid zero minimum coverage. However, Bernardo's (2005b) Figure 7 suggests that when $\alpha > 0.2$, approximately, the intrinsic interval shares this undesirable property with the central interval. Also, as $\alpha \rightarrow 1$, the HPD interval converges to $\hat{\theta} = 0$, an estimate called "utterly useless" by Bernardo (2005b, p. 342), even though as a *data-based* point estimate it seems perfectly adequate.

When deriving an interval from a reference posterior, referring to a suitable likelihood function rather than the intrinsic loss function, appears to add better sampling properties to the invariance property (work in progress). Of course there is only one likelihood function for one-parameter models, when this approach (see e.g., Box and Tiao, 1973, p. 124) leads to HPD in the metric for which the reference prior is uniform. But in the case of σ in the Normal model, for example, it is the use of the marginal likelihood that leads to HPD in $\log(\sigma)$, and to an "unbiased" confidence interval (Lindley *et al.* 1960), less complicated and more attractive (to frequentists and, I suggest, many Bayesians) than the interval from Example 6.

Reference priors. I believe Bernardo's (1979) article to be one of the most important contributions in the history of statistics. It appears, however, that a reference prior may be too informative when it "shoots off to infinity" (Zellner's words in a comment to Geisser, 1984) at the extreme of a parameter range, if in fact the pdf is defined at such an extreme. The most common example of this is, of course, the binomial; setting $0 < \theta < 1$ does not take away the fact that this model is also valid for $\theta = 0$ and $\theta = 1$. This could be the reason behind why the uniform or Bayes-Laplace prior appears preferable, as a representation of prior ignorance and thus for the purpose of scientific communication and sensitivity analysis, to the reference/Jeffreys prior $\text{Be}(\theta | 1/2, 1/2)$. This can be most easily shown by considering $x = 0$ ($x = n$) (Tuyl *et al.* 2008). As a related illustration, consider the Bayesian Rule of Three which states that, based on $x = 0$ and an informative prior $\text{Be}(\theta | 1, b)$ ($b > 1$), the 95% one-sided upper credible limit may be approximated by $3/(n+b)$ when n is large (Jovanovic and Levy, 1997). It is easy to check that the reference prior leads to an approximate limit of $1.92/n$ here. Equating the two rules gives $b = 0.56n$, so that under this scenario the reference prior adds, in effect, $0.56n - 1$ observations (*i.e.*, failures) to the uniform prior. For another argument in favour of the uniform prior, and an example of the informativeness of $\text{Be}(\theta | 1/2, 1/2)$ priors in the context of a 2×2 contingency table, see Aitkin *et al.* (2005, p. 229).

In the current article, the approximate marginal reference posterior from Example 4, $\text{Be}(\theta_i | n_i + 1/m, n - n_i + (m-1)/m)$ seems of particular concern when m is large and $n_i = 0$, resulting in a credible interval (for θ_i) too close to 0. For $m = n = 100$, for example, the 95% reference upper limit is 0.000033, even though $n_i = 0$ would be a common occurrence for θ_i many times greater than this limit. Instead, the 95% upper limit resulting from the uniform Dirichlet prior is 0.015, which seems more reasonable. [As pointed out by Bernardo, Dirichlet posteriors depend on m , which, if the only prior knowledge is that m categories exist, seems more reasonable than dependence on n_i only. The reference posterior $\text{Be}(\theta_i | n_i + 1/2, n - n_i + 1/2)$ also given by Bernardo does have the latter property ("as one would hope") and happens to perform better here, but remains too informative in general, as shown above in the context of $m = 2$.] Philosophical differences aside, it is important to acknowledge Jaynes's (1976, p. 178) words, "The merits of any statistical method are determined by the results it gives when applied to specific problems." Jaynes gave examples for which credible intervals are clearly superior to certain confidence intervals, with the same frequentist properties. However, the interval (0, 0.000033) above leads to inadequate coverage, and would be embarrassing to give to a client.

Examples by Lindley in his discussion of Bernardo (1979) are relevant here. Lindley started off referring to the different reference/Jeffreys prior $\text{Be}(\theta | 0, \frac{1}{2})$ for the negative binomial parameter: many Bayesians agree with Lindley that violation of the likelihood principle is undesirable. While Bernardo has continued to defend this prior, it seems that Bayesians who have adopted "the" Jeffreys prior for binary data analysis, do *not* usually check how the sample was collected, and simply adopt the $\text{Be}(\theta | 1/2, 1/2)$ —just like most frequentists tend to use *their* standard binomial calculations without checking the sampling rule. Ultimately, Geisser's (1984) counter arguments, in his reply to discussants (including Bernardo), seem irrefutable; for example, which reference/Jeffreys prior results when the experiment is stopped when either x successes or n trials are achieved, whichever occurs first? A possibility would be to adopt $\text{Be}(\theta | 0, 1/2)$ when the former and $\text{Be}(\theta | 1/2, 1/2)$ when the latter occurs, but such rules seem to simply illustrate the need for the

likelihood principle: what if the x^{th} success occurred on the n^{th} trial?

Interestingly, this type of prior, although nothing to do with a stopping rule as such, plays a role in the trinomial example given next by Lindley. In the context of life-tables, the three probabilities are $\lambda\{1 - (1 - \delta)\mu\}$, $(1 - \delta\lambda)\mu$ and $(1 - \lambda)(1 - \mu)$, with frequencies D of deaths, W of withdrawals and S of survivors. Lindley challenged Bernardo's reference prior methodology for this example, stating that for $\delta = 0$ or $\delta = 1$, when we "isolate λ say", the reference prior is not the usual $\text{Be}(\lambda | \frac{1}{2}, \frac{1}{2})$. However, in his reply Bernardo showed that his method *does* give this result for $\delta = 0$, but $\text{Be}(\lambda | \frac{1}{2}, 0)$ when $\delta = 1$. About this different prior Bernardo remarked, "I suspect that $\delta = 1$ is a limiting condition which precisely implies this type of sampling rule.", which is hardly convincing. In contrast with genuine negative binomial sampling, there is a definite problem, as the reference posterior $\text{Be}(\lambda | D + \frac{1}{2}, W + S)$ is improper when $D = n$.

Bernardo's analysis was based on setting $\phi_1 = (1 - \delta\lambda)\mu$ and $\phi_2 = (1 - \lambda)(1 - \mu)$. Based on Bayes's (1763) original argument for a uniform prior predictive distribution, we should consider a uniform prior on (ϕ_1, ϕ_2) . For $\delta = 1$, the posterior for λ is now $\text{Be}(\lambda | D + 1, W + S + 2)$ which makes sense: before any observations the prior mean is $\frac{1}{3}$, the result of a prior 'balanced' view of the frequencies D , W and S . Of course for large frequencies, adopting a uniform prior on (λ, μ) , which in case of $\delta = 1$ results in a slightly different posterior $\text{Be}(\lambda | D + 1, W + S + 1)$ instead, is adequate, unlike Bernardo's $\text{Be}(\lambda | D + \frac{1}{2}, W + S)$ *not* breaking down for any zero frequencies.

It seems surprising that Bernardo himself found the potential impropriety of the Haldane posterior of the binomial parameter, based on the prior $\text{Be}(\theta | 0, 0)$, "less than adequate" (Bernardo 1979, p. 119), but was not concerned about the same consequence for the $D = n$ possibility in this example. It appears that Lindley's intuition was correct, and that the example provides strong evidence that reference priors can be suboptimal for certain models. In contrast, Bernardo's treatment of the Fieller-Creasy problem and Stein's paradox, for example, must have been truly remarkable in 1979.

Lindley finished off by stating, "...but he has been successful in overcoming other difficulties, and the rewards of success would be so great, that I am sure he will be able to overcome these teasers." Until Bernardo does so, it appears that non-regular models such as the ones discussed above should be excluded from the current reference prior methodology.

REPLY TO THE DISCUSSION

I am extremely grateful to all discussants for their relevant, interesting, and thought-provoking comments. Naturally, I have also much appreciated their very nice personal remarks (which sometimes have made me blush). In particular, it is indeed very nice to be told that one is able to produce diamonds: I only hope that people would not be shy to wear them! And I was really moved by the warm opening words of Professor Lindley, my *maestro*, the person who introduced me to Bayesian statistics, and who has always been a key figure in my professional life.

In this rejoinder, I will first concentrate in trying to give specific answers to the discussant's queries, grouping these by subject, and I will finally attempt to summarize what I perceive to be the main conclusions.

Objectivity. Professor Lindley has foundational objections to the use objective priors. As any Bayesian surely knows, Lindley moved from an objectivist Bayesian position closely related to Jeffreys (one of his mentors), nicely developed in his excellent pioneering book (Lindley, 1969), to an extreme subjectivist approach. I guess that (as it is often the case in genetics), in this issue I am closer to my grandfather. Of course, many choices in any statistical analysis are subjective, and it may be claimed the objectivity only arises when scientists reach agreement. However, by limiting and making explicit these subjective choices, and using a prior distribution and a loss function chosen by consensus, this agreement is more likely to be obtained. Reference priors and intrinsic loss functions are precisely proposed for this type of consensus. These also produce benchmarks against solutions derived from more subjective choices may be usefully compared.

Both Professor Lindley and Professor Singpurwalla object to the treatment of parameters as “Greek symbols” without a context specific meaning, reducing statistical analysis to a mathematical exercise. I believe that this *objective* mathematical exercise is precisely what scientist often demand in order to reach some basic consensus on what, for a given assumed model, the data imply on the likely values of parameters which label the model. If those parameter have a physical interpretation (which is not always the case) *and* the scientists are able to specify a context-based prior, they should indeed use this information in the form of a context-based prior distribution but, even then, computing the reference posterior will be helpful to verify to what extent the results they obtain depend of the particular context-dependent prior which they have decided to use.

I certainly agree with Singpurwalla in that prediction is often the final goal of an statistical investigation but, for any given model, computation of a posterior predictive technically requires the derivation of a joint posterior for the parameters. Naturally, a reference posterior predictive requires a reference prior. As Dr. van der Linde mentions, in prediction problems with multiparameter models the quantity of interest (required to derive the appropriate reference prior) is not obvious. Using as the quantity of interest the median $\theta(\omega) = \text{Median}[y | \omega]$ of the sampling distribution of the quantity y to predict seems to produce very attractive reference predictive densities. For many successful examples of this approach, see Román (2011).

As Professors Pericchi, Lindley and Moreno all suggest, I am sure that there are situations where the scientist is willing to use a prior distribution highly concentrated at a particular region and explore the consequences of this assumption. Lindley describes this as a subjective choice, while both Pericchi and Moreno argue that some of these choices are somewhat objective. What I claim is that, even in precise hypothesis testing situations, the scientist is often interested in an analysis which does *not* assume this type of sharp prior knowledge, and that standard reference priors may be used to give an objective Bayesian answer to the question of whether or not a particular parameter value is compatible with the data, without making such an important assumption.

In line with his subjectivist approach to Bayesian statistics, Singpurwalla advocates the use on context-based loss functions as opposed to automatic loss functions, like those provided by divergence measures. My reaction to this is very much the same as that provoked by the suggestion to use context-based priors. In recommending the optimal dose of a drug given available experimental data, the pharmacologist will probably have to consider that too large doses may be severely toxic, while too small doses may be just complemented, and consequently use a non symmetric loss

leading to Bayes estimators for the optimal dose far smaller than the intrinsic estimator. However, an astronomer estimating the speed of a galaxy would probably just want to know the speed values which are most compatible with available data, and those will be provided by the intrinsic estimator and by intrinsic credible regions. Even the pharmacologist will probably want to know the intrinsic estimator of the optimal dose, if only to compare this with the solution suggested by his context-based loss function. Very much like information-theoretical ideas provide an intuitive basis for the formal definition of reference priors, divergences between models derived from functional analysis provide an intuitive bases for objective loss functions. The intrinsic discrepancy has an additional important intuitive justification for statisticians, as it is directly related to the average log-likelihood ratios between models, a concept well understood and appreciated by most statisticians.

Mathematical formulation. Singpurwalla finds pedantic the explicit introduction of nuisance parameters in our formulation. While it is certainly true that some of the central ideas are easier to describe in simple one-parameter models, most real problems use models with many parameters, and the extension to multiparameter problems of the original ideas is not particularly trivial. If I had been writing a book rather than a review paper, I would surely had used a two-stages approach but, with the space limitations of a paper, I was obliged to directly describe the general solution.

As Professor Strenger points out, I agree with both Fisher and Lindley in that pure scientific inference is a somewhat different problem than context-based decision making. However as described in the *Annals* paper which Lindley quotes (Bernardo, 1979b), statistical inference may formally be described as a decision problem (where the action space is the class of posteriors, and the utility function an information measure) and, as a consequence, decision theoretical techniques may be used to obtain sound procedures for statistical inference. It is decision theory which makes precise the conditions under which, say a particular point estimator, may be appropriate. I perceive decision theory as an appropriate guide to guarantee good statistical practice.

Bayesians have used decision theory for decades to obtain good general estimators, but too often the loss function used (in many cases just the quadratic loss) has been rather naïve. I have proposed the use of the intrinsic loss (a context-independent, information-based, invariant loss) as a possible improvement in those pure inference problems. Of course, as mentioned above, context-dependent decision problems should make use of context dependent loss function. Both Lindley and Sprenger mention the fact that the intrinsic loss (as the ubiquitous quadratic loss) is not bounded. I do not think this is a problem in scientific inference: conceptually, because an unbounded loss is the mathematical code for the possibility of being totally wrong; pragmatically, because the tails of the posterior will typically make irrelevant the form of the loss when one is far from the more likely region. That said, as Sprenger quotes, I certainly believe that intrinsic losses and reference priors should not be understood as a challenge to a Bayesian context-dependent analysis, but as conventional consensus choices, and as a form of sensitivity analysis.

All the ideas presented in this paper could have been developed using utility functions rather than loss functions; this would have probably been closer to foundations, but I believe that the use of losses makes interpretation of the results far simpler in the context of pure statistical inference.

Invariance. Pericchi recognizes that invariance with respect to one-to-one transformations is a “convenient” property, but then quotes Box and Tiao in an attempt to defend the use of non-invariant procedures, such as HPD regions. But this quote does *not* support non invariant procedures for the choice of credible regions. Box and Tiao are obviously right when they say that “the relative credibility of any two parameter points ought to change with non linear transformations”. This is indeed a trivial consequence of probability theory. Given a posterior density, say $\pi(\theta | \mathbf{z})$, one always has

$$\frac{\pi(\phi_i | \mathbf{z})}{\pi(\phi_j | \mathbf{z})} = \frac{\pi(\theta_i | \mathbf{z})}{\pi(\theta_j | \mathbf{z})} \frac{|\partial\theta/\partial\phi|_{\theta=\theta(\phi_i)}}{|\partial\theta/\partial\phi|_{\theta=\theta(\phi_j)}},$$

and this will generally be different from $\pi(\theta_i | \mathbf{z})/\pi(\theta_j | \mathbf{z})$ unless $|\partial\theta/\partial\phi|$ is constant. This does *not* imply however, that credible intervals should not be invariant. Indeed, the statement $\theta \in B$ is *precisely the same* as $\phi \in \phi(B)$ and, we argue, any sensible approach to the choice of credible regions should take this rather obvious fact into account. And certainly, insistence in this invariance does *not* lead to inappropriate solutions, as demonstrated in the many examples contained in Bernardo (2005b).

Professors Gutiérrez-Peña, Rueda and Mendoza are certainly right when they point out that one obtains invariant results if one chooses a loss function in a particular parameterization and then uses the corresponding transformation in any other parameterization. This is indeed the appropriate procedure if one uses a context dependent loss function (such as a monetary loss). If, however, one is interested in the simplifications and generality associated to the use of a conventional loss function (such as the quadratic) this requires to specify in which particular parameterization the conventional loss is appropriate (a non-trivial decision) and, moreover, this would lose the automatic calibration aspect of intrinsic loss. More importantly, measuring discrepancies between models makes far more sense to me that measuring distances between parameters, and invariance is then an immediate by-product.

Consistency. Pericchi does not like the fact that, given any fixed cut-off point in the procedure we suggest for testing a precise null H_0 , the probability of rejecting H_0 when it is false goes to one as the sample size increases, but the probability of rejecting the null when it is true does not go to zero as $n \rightarrow \infty$. We however consider this mathematical fact as a very welcome feature of the proposed testing procedure. Indeed, most philosophers of science agree that one can *falsify* a hypothesis (*i.e.*, to reject it when it is false) but one can *never prove* it (*i.e.*, to accept it with probability one, even if it is true); one may only claim that data are probably *compatible* with H_0 , but many other explanations may also be compatible with the data. Indeed the intrinsic discrepancy test does *not* place the null and the alternative on an equal footing and, we argue, rightly so: the alternative (the full model) is true by assumption, and one is testing whether or nor the data are compatible with the restricted model described by H_0 , hardly a symmetric situation. That said, one should not limit to analysis to selecting a particular cut-off and simply report whether or not the intrinsic statistic $d(\boldsymbol{\theta}_0 | \mathbf{z})$ exceeds, or does not exceed, that value. Indeed, the whole behaviour of $d(\boldsymbol{\theta}_0 | \mathbf{z})$ as a function of $\boldsymbol{\theta}_0$ is of interest and, for each particular $\boldsymbol{\theta}_0$, the value $d(\boldsymbol{\theta}_0 | \mathbf{z})$ is a direct, operational measure of the possible incompatibility of $\boldsymbol{\theta}_0$ with the data, in terms of the expected average log-likelihood ratio against this particular parameter value.

Professors Pericchi, Ghosal, Robert and Rousseau do not like the fact that the threshold required in the proposed procedure for hypothesis testing is independent of the sample size. Yet, we feel this is a very sensible feature of the procedure, for this describes the upper limit of the sampling average log-likelihood ratio against the null which one is prepared to tolerate without rejection, and this *utility* constant should surely be independent of the data size. Indeed, we find very unappealing the frequent *ad hoc* sample size adjustments that people are forced to do with conventional testing procedures. Notice also that our procedure is described in terms of available data \mathbf{z} , and this may have a totally general structure that will often *not* be that of a random sample, so that the concept of “sample size” cannot possibly be generally relevant. In particular, Ghosal suggests that, as the intrinsic loss is multiplied n -fold in the i.i.d. case, a factor of n in the cut-off point should be expected. I do *not* think this should be the case. For any data set $\mathbf{z} = \{x_1, \dots, x_n\}$, a fixed (sample size independent) cut-off point typically forces the size of the acceptance region to be proportional to $1/\sqrt{n}$, which (under regularity conditions) is precisely what one would expect. To take the simplest example, testing the value for a normal mean μ_0 given a random sample $\mathbf{z} = \{x_1, \dots, x_n\}$ from $N(x|\mu, 1)$, the expected intrinsic loss is $n \delta\{N(\cdot|\mu, 1), N(\cdot|\mu_0, 1)\} = n(\mu - \mu_0)^2/2$ and, using the cut-off constant $k = \log_e(e^k)$, the null value μ_0 will be rejected whenever $|\bar{x} - \mu_0| > \sqrt{2k-1}/\sqrt{n}$ so that, as one would expect, the size of the acceptance region decreases with $1/\sqrt{n}$. In particular, for $k = 3$ (where the null is rejected whenever the sampling average log-likelihood ratio against μ_0 may be expected to be larger than 3, and hence the likelihood ratio about $e^3 \approx 20$), this happens whenever $|\bar{x} - \mu_0| > 2.23/\sqrt{n}$. For $k = 2.42$ one gets the conventional rejection rule $|\bar{x} - \mu_0| > 1.96/\sqrt{n}$, which corresponds to an expected likelihood ratio against the null of about only $e^{2.42} \approx 11$. This is hardly conclusive evidence against μ_0 ; no wonder therefore that, as often reported in the literature, many frequentist $\alpha = 0.05$ based rejections turn out to be wrong rejections. Notice that if the cut-off constant had been chosen of the form nk , the rejection rule would have been $|\bar{x} - \mu_0| > \sqrt{n(2kn-1)}/n$, converging to $|\bar{x} - \mu_0| > \sqrt{2k}$ for large n , a hardly acceptable proposal.

Compound hypothesis. Gutiérrez-Peña, Rueda, Mendoza, Robert and Rousseau all question the suitability of the simple approach proposed to deal with compound hypothesis. The proposed testing procedure is consciously defined in terms of testing whether or not a *particular* value θ_0 of the parameter is compatible with the data. Depending on the context, a compound hypothesis of the form $H_0 \equiv \{\theta_0 \in \Theta_0\}$ is to be rejected when (i) *at least one* of its elements would be rejected, or when (ii) *all* its elements would be rejected. This second case is likely to be the more frequent situation, but the solution proposed may be used to deal with both cases.

For instance, in a scientific context, where the parameter values Θ_0 are *all* those compatible with an established theory, rejecting the hypothesis H_0 is rejecting that theory, and this will be the case if *all* of the values in Θ_0 are considered to be incompatible with the data. Thus, in high energy physics, the accepted theory may imply that the energy of a type of particle *must* lie in a given interval; if all the values in that interval are incompatible with the data, then that theory must be revised, and new physics proposed. However, in a pharmacological context, where Θ_0 is the set of all the conditions under which the drug must work, the hypothesis that the drug is worth further study is to be rejected if *at least one* of those parameter values is considered to be incompatible with the data, for this means that the drug does not work under, at least, some of the required conditions.

Approximations. Professor Ghosh refers to the approximate solution for testing the value of a normal mean when both parameters are unknown (Example 6) and wonders about the precision of that approximation. Actually, the approximation is quite good. Here, the intrinsic divergence between the model $N(\cdot | \mu, \sigma)$ and the hypothesis H_0 , which is the set $\{N(\cdot | \mu_0, \sigma_0), \sigma_0 > 0\}$, is given by

$$\begin{aligned} \delta\{H_0 | \mu, \sigma\} &= \inf_{\sigma_0 > 0} n \delta\{N(\cdot | \mu, \sigma), N(\cdot | \mu_0, \sigma_0)\} \\ &= \inf_{\sigma_0 > 0} \frac{n}{2} \left[\log \frac{\sigma_0^2}{\sigma^2} - 1 + \frac{\sigma^2}{\sigma_0^2} + \frac{(\mu - \mu_0)^2}{\sigma_0^2} \right] \\ &= \frac{n}{2} \left[\log \left(1 + \left(\frac{\mu - \mu_0}{\sigma} \right)^2 \right) \right] = \frac{n}{2} \log(1 + \theta^2), \end{aligned}$$

where $\theta = (\mu - \mu_0)/\sigma$. Moreover $\pi(\theta | \mathbf{z})$, the marginal posterior distribution of θ given $\mathbf{z} = \{x_1, \dots, x_n\}$ which corresponds to the reference prior $\pi(\mu, \sigma) = \sigma^{-1}$, is a non-central chi-squared which is proper for any $n \geq 2$, and which only depends on the data through the sample size n and the absolute value of the conventional statistic $t = (\bar{x} - \mu_0)/(s/\sqrt{n-1})$, where $s^2 = \Sigma(x_j - \bar{x})^2/n$. The reference expected intrinsic loss may thus be numerically computed as

$$d(\mu_0 | \mathbf{z}) = d(t, n) = \frac{n}{2} \int_{-\infty}^{\infty} \log(1 + \theta^2) \pi(\theta | t) d\theta \approx \frac{n}{2} \log \left[1 + \frac{1}{n+1} (1 + t^2) \right].$$

Table 1: Exact and approximate values of the intrinsic statistic $d(t, n)$ to test the value of a normal mean μ_0 , for $n = 25$, where t is the conventional t statistic.

$ t $	Exact Value of $d(t, n)$	Approximation
0	0.473	0.472
1	0.915	0.926
2	2.157	2.199
3	3.995	4.068
4	6.192	6.289
5	8.555	8.664
6	10.950	11.063

The quality of the approximation may be appreciated from Table 1, which gives the exact and the approximate values of $d(t, n)$ for several $|t|$ values and $n = 25$. The limiting value of $d(t, n)$ as $n \rightarrow \infty$ is $(1 + t^2)/2$. For fixed n , the value of $d(t, n)$ goes to infinity as $|t|$ goes to infinity so that, as one would surely expect, null values with very large $|t|$ values will always be rejected.

Since the expected loss is only a function of t and n , any cut-off value d_0 will be numerically equivalent to a particular choice of the significance level in the conventional frequentist t test. However, the practical differences with the proposed procedure are rather radical. For instance, one finds that the choice $d_0 = 2.42 = \log 11$ corresponds to p -values of 0.039, 0.048 and 0.050 for sample sizes 10, 100 and 1000 respectively. Thus, the null would typically be rejected by conventional practice when the only evidence is that the likelihood ratio against the null is expected to be about 11. Thus, once again, a sizeable proportion of frequentist rejections may be expected to be wrong rejections.

Bayes factors. As one would expect from scholars who have produced many interesting results using Bayes factors, Pericchi, Gómez-Villegas and Moreno are all unhappy with my treatment of precise hypothesis testing. I now analyze further the ESP example, which we all seem to agree is a good test case. The question is whether or not $r = 52,263,471$ successes in $n = 104,490,000$ Bernoulli trials is, or is not, compatible with the *precise* value $\theta = 1/2$. With such a huge sample, the posterior density which corresponds to *any* non-pathological positive prior on $(0, 1)$ is actually $N(\theta | 0.50018, 0.000049)$, so that the specific prior choice is pretty much irrelevant. This is shown in the upper panel of Figure 9.

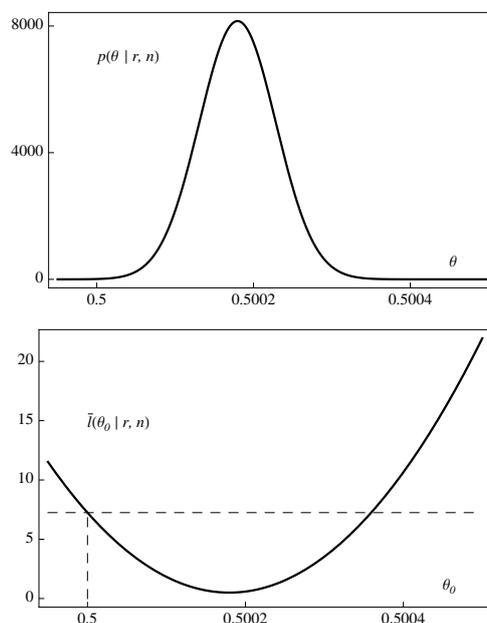


Figure 9: Posterior density and expected intrinsic loss for the parameter θ of a binomial model, given $r = 52,263,471$ successes in $n = 104,490,000$ trials.

It should be obvious from that figure that any Bayesian with a non-dogmatic prior is *forced* to conclude that the precise value $\theta = 0.5$ is *not* well supported by the data. More precisely, using the intrinsic discrepancy loss function (represented in the lower panel of Figure 9), the expected loss from using $\theta = 0.5$ in place of the true, unknown θ value is about $7.24 = \log[1400]$ so that, given the data, the sampling average log-likelihood ratio against $\theta = 1/2$ may be expected to be 7.24 (again with virtually any non-dogmatic prior) and hence, in any future use of the model the data may be expected to be about 1400 times more likely under the true value of θ (which should roughly be around 0.5002 ± 0.0001) than under $\theta = 1/2$. Thus, one should *not* work under the hypothesis that θ is *precisely* equal to $1/2$.

The fact that this conclusion agrees with the conclusion which one would obtain using p -values in this case does *not* mean (as Singpurwalla seems to suggest) that I lean on p -values as a yardstick. The argument rest is solidly on Bayesian arguments.

We conclude that H_0 should be rejected, but Pericchi computes the Bayes factor which corresponds to the use of the density $\text{Be}(\theta | 1/2, 1/2)$ as the *conditional* prior given that $H_0 \equiv \{\theta = 1/2\}$ is false, and obtains $B_{01} = 18.7$ suggesting a “ratio of evidences” of about 19 *in favour* of H_0 . And again, the conclusion will be qualitatively the same with any other non-pathological conditional prior under the full model. But he adds that this conclusion is reached with the same assumptions (the reference prior) which I propose to use. However, this is certainly *not* the case. As pointed out before by many authors, Bayes factors have no direct foundational meaning to a Bayesian: only posterior probabilities have a proper Bayesian interpretation. The fact that, under appropriate conditions, the Bayes factors contain all relevant data information to obtain the required posterior probabilities associated to a finite set of possibilities does *not* mean that one may reason in terms of Bayes factors rather than probabilities, very much as the fact that likelihood functions contain all data information to obtain posterior densities does not mean that one may reason in terms of likelihoods instead of in terms of posterior densities. In this particular case the Bayes factor B_{01} is only relevant if one wishes to obtain the posterior probabilities $\Pr[H_0 | \mathbf{z}]$ and $\Pr[H_1 | \mathbf{z}]$ and these only make sense in one assumes a prior of the form $\Pr[H_0] = p_0$, $\Pr[H_1] = 1 - p_0$, $p(\theta | H_1) = p(\theta)$, for some $p_0 > 0$ and some proper prior $p(\theta)$; clearly this is a non-regular “sharp” prior which will always be *very* different from the continuous reference prior $\pi(\theta)$ which I assume. Contrary to Pericchi’s assertion, the reference prior value $\pi(\theta_0) = 0$ is not in violation of Cromwell’s rule, but a simple consequence of the fact that H_0 is a measure zero set in this setting. We argue that only if one restricted the parameter space to a *finite* set of values $\Theta = \{\theta_0, \theta_1, \dots, \theta_k\}$ (and then one will be in an entirely different problem) would the assumption $\Pr(\theta_0) > 0$ be required.

Sprengrer suggests that our standard for judging compatibility should depend on how plausible we consider the alternatives. I do not believe this should be case. The fact that one is interested in a particular θ_0 value does not *require* that this value is judged to be more likely. Interest is measured in terms of utility, not in terms of probability. One is *not* interested in how likely θ_0 is, a less than appropriate question in a continuous setting, but on whether or not the simplified model $p(\mathbf{z} | \theta_0)$ may safely be used in place of the assumed model $\{p(\mathbf{z} | \theta), \theta \in (0, 1)\}$, a very different question. We give an operational answer to this problem by suggesting that the null should be rejected whenever the expected average log-likelihood ratio of the assumed model against the null is too large. In the ESP example the message is clear: one should *not* conclude that θ is *precisely* equal to 0.5, for there is sufficient evidence to state that the true value of θ is likely to be within 0.5002 ± 0.0001 . No matter what the physicist friends of Singpurwalla apparently believe, you have to know some statistics to see that a raw propensity of success of 0.500018 with this huge sample size of $n = 104,490,000$ does *not* lead to accept the *precise* value $\theta_0 = 0.5$, but to conclude that a small bias was very likely present. Whether or not this is an important practical conclusion is another matter, which (if required) could surely be treated as a formal decision problem, and analyzed with a context specific utility function.

In an effort to justify a large posterior probability for the null, Moreno mentions likelihood ratios, but a careful analysis of the likelihood ratios suggests precisely the opposite. Thus, if one follows the common practice of plotting the likelihood ratio *against* the null $\text{BF}_{10}(\theta) = [\theta^r(1 - \theta)^{n-r}]/(1/2)^n$ as a function of the parameter θ (see Figure 10) one naturally finds that this ratio is large for all values in a region close to the m.l.e., $\hat{\theta} \approx 0.50018$, essentially reproducing (at another scale) the upper

panel of Figure 9. Thus, $\text{BF}_{01}(0.50018) = 686$, and $\text{BF}_{10}(\theta)$ is only smaller than 1 if $\theta < 0.5$ or $\theta > 0.50035$. It should be obvious that only a *very* dogmatic prior *extremely* concentrated on $\theta = 1/2$ could possibly dominate the data and give a large probability to a tiny interval around the null (and this would hardly qualify as an objective procedure which lets the data speak for themselves!)

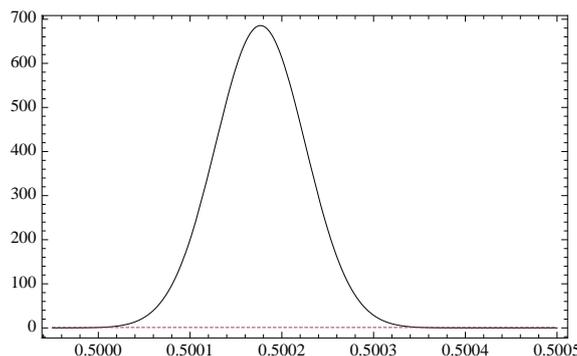


Figure 10: Likelihood ratio against $\theta = 1/2$ as a function the parameter θ of a binomial model, given $r = 52,263,471$ successes in $n = 104,490,000$ trials.

We all agree that scientists often need a summary of the implications of the data to the problem at hand and that “it is better to have large posterior probabilities in the right direction than to decide robustly in the wrong” but I am afraid that I do not agree with Pericchi on what the appropriate summary should be, or on precisely what is right and what is wrong here. To the best of my knowledge, the posterior density in the upper panel of Figure 9, a direct consequence of probability theory with no special assumptions about the prior, and possibly complemented by the expected loss in the lower panel of that figure, is the summary the scientist needs, while the conventional Bayes factor analysis is plainly misleading in this problem.

As Berger and Delampady (1987) correctly pointed out (precise) “nulls will always be rejected if enough data are taken because there will be uncontrolled sources of bias”, and this is possibly what data are showing here: the machine used possibly had a small positive bias, and this has been duly detected by the reference analysis. We do not have to believe in ESP, but the fact remains that the proportion of successes which the machine produces is found to be very likely *different* from *exactly* 50%, in direct contradiction with the Bayes factor results. As mentioned before, the analysis of practical consequences of this undeniable fact is another issue.

Lindley’s paradox. There are *two* rather different facts in the mathematical behaviour of the Bayes factor for testing whether a normal mean is zero, the example quoted by Moreno to discuss Lindley’s paradox. The fact mentioned by Moreno is that $B_{10} \rightarrow 0$ as the prior variance increases, proving that the usual objective prior in this problem, the uniform $\pi(\theta) = 1$ cannot be used with this formulation. The other fact, rather more upsetting, is that for any fixed value of $n\bar{x}^2$ (the square of the number of standard deviations \bar{x}/\sqrt{n} that the m.l.e. \bar{x} is from the null value $\theta = 0$), the Bayes factor $B_{10} \rightarrow 0$ as $n \rightarrow \infty$, hence leading to accept H_0 (for sufficiently large samples) no matter what the relevant data say. This is Lindley’s paradox,

and illustrates the fact that, when true parameter values are order $O(n^{-1/2})$ of the null and the sample size is large, one may obtain totally misleading answers from Bayes factors. This is a direct mathematical consequence of the fact that, in those conditions, a continuous prior concentrated on θ_0 cannot be properly approximated by a sharp prior with a point mass on θ_0 (Berger and Delampady, 1987).

Thus, I certainly do not agree with Pericchi's that Lindley's paradox has been "misunderstood" as an illness of Bayes factors for precise hypothesis testing. On the contrary, this clearly poses a very serious problem to Bayes factors, in that, under certain conditions, they may lead to misleading answers. Whether you call this a paradox or a disagreement, the fact that the Bayes factor for the null may be arbitrarily large for sufficiently large n , *however relatively unlikely the data may be under H_0* is, to say the least, deeply disturbing.

To further illustrate this point, consider again the EPS example discussed above. For large n and r/n close to $1/2$, the Bayes factor for testing whether or not the binomial parameter θ is precisely equal to $1/2$ becomes

$$B_{01}(r, n) \approx \frac{1}{2} \log \left[\frac{n\pi}{2} \right] - 2n \left(\hat{\theta} - \frac{1}{2} \right)^2, \quad \hat{\theta} = \frac{r}{n}.$$

With the EPS data of Example 7, this yields indeed $B_{01} = \exp[2.93] = 18.7$, as Pericchi mentions. Now suppose that the m.l.e. is at a distance of order $1/\sqrt{n}$ from the null, so that, say, $\hat{\theta} = 1/2 \pm \delta/(2\sqrt{n})$; then the Bayes factor becomes

$$B_{01}(r, n) \approx \frac{1}{2} \log \left[\frac{n\pi}{2} \right] - \frac{\delta^2}{2},$$

which (for any fixed δ) tends to infinity as n increases, therefore always strongly supporting the null for large enough samples. However, the likelihood ratio of the m.l.e. *against* the null will in this case be

$$\frac{\text{Bi}(r | \hat{\theta}, n)}{\text{Bi}(r | 1/2, n)} \approx \exp \left[\frac{\delta^2}{2} \right] = \exp \left[2n (\hat{\theta} - 1/2)^2 \right],$$

which, for all $\hat{\theta} \neq 1/2$, will be large for large n values. Thus, for large sample sizes, whenever the true value of the parameter is $O(n^{-1/2})$ from the null, the Bayes factor analysis may be completely misleading, in that it would suggest *accepting* the null, even if the likelihood ratio for the m.l.e. *against* the null is very large.

This is precisely what happens in the EPS example. Here the likelihood of the m.l.e. $\hat{\theta} = r/n$ *against* the null is $\text{Bi}(r | \hat{\theta}, n) / \text{Bi}(r | 1/2, n) \approx 686$, but the m.l.e. is $O(n^{-1/2})$ from $1/2$, with $\hat{\theta} = 1/2 + \delta/(2\sqrt{n})$ and $\delta = 3.614$. Thus, for any non-dogmatic continuous prior, the null is 3.614 posterior standard deviations from the m.l.e. (something most probabilists would regard as strong evidence against the null), the likelihood ratio of the m.l.e. against the null is about 686 and, yet, the Bayes factor suggests "evidence" for the null of about 19 to 1! We strongly believe that this behaviour (which is shared by *all* Bayes factor based procedures) is less than satisfactory.

Gómez-Villegas follows Jeffreys' conventional approach and claims that testing in a continuous parameter problem *forces* the use of a non-regular sharp prior. As

mentioned before, this is of course true if one insists in presenting the answer as a posterior probability for the null but, as demonstrated above, this is certainly *not* required if, for a given model, one wishes to test the compatibility of the available data with the null, which is precisely what I believe one should be doing. Placing (as he suggests) a larger value for p_0 , the prior probability of the null, than the conventional $p_0 = 1/2$ will often get closer compatibility with p -values practice, but I am less than convinced that this will generally provide a good answer. Moreover, for *any* p_0 choice, Lindley's paradox will always appear for specific values of the sample size.

As pointed out by Moreno, the formulation of model choice as a decision problem on the finite set of alternative models is of course fine. It is on the choices of the loss function and the prior distribution where where disagreement may occur. In particular, the 0–1 loss is possibly too naïve, for this cannot take into account the actual differences between using the alternative models for the problem under scrutiny. In nested models with continuous parameters it is precisely the use of this particular loss what forces the use of sharp priors, and this in turn leads to Lindley's paradox, with the possibly devastating effects illustrated by the ESP example discussed above.

Reference priors. Lindley, Gómez-Villegas and Tuyl all mention that reference priors are apparently incompatible with the likelihood principle. Once the data have been obtained, the likelihood “principle” is an immediate consequence of Bayes theorem, stating that inferences should only depend on observed data. However, integrating on the sample space is *mandatory* in many statistical tasks to be performed *before* the data are obtained. These include experimental design and reference analysis: one cannot optimize an experiment without some assumptions of its possible outcomes, and one cannot determine the (reference) prior which maximizes the missing information from a particular experiment without making precise the experiment one is willing to consider.

Reference posteriors are conditional answers of the “what if” type. They provide a meaningful answer to a very precise question: given a set of data which are assumed to have been generated by a particular model, what could be said about some function of its parameters *if* initial knowledge were described by that prior which maximizes the missing information which this experiment could possibly provide? Obviously, the answer is bound to depend of the experiment considered, and there is certainly no fundamental difficulty in simultaneously considering different plausible experimental settings as part of an (always welcome) sensitivity analysis.

Robert and Rousseau refer to the general definition of reference priors, where the required prior function is obtained from the behaviour of the posterior distribution of the quantity of interest under repeated replication of the original data structure, as opposed to simply letting the sample size increase. As they point out, the result may be very different, unless the original data structure already consists of n i.i.d. observations. I strongly believe that the original formulation is always the appropriate one. Indeed, by definition, the reference prior is designed to obtain that prior which maximizes the missing information which the experiment analyzed could possibly provide, and this would only be obtained by repeated observations from the precise experiment analyzed. Notice that this formulation naturally permits the reference prior to take the experiment design into account; for instance, in two sample problems (like that comparing two normal means), the reference prior typically depends on the ratio n/m of the two sample sizes, and this is obtained

by analyzing the posterior densities which correspond to k replications of pairs of samples of size n and m from the two populations, and letting k increase (for a detailed example, see *e.g.*, Bernardo and Pérez, 2007). If the data structure consists of n not i.i.d. observations, letting n go to infinity often produces an *approximate* reference prior, which could be used with actually *large* sample sizes; the results if one uses that prior with small data sets (where the relevance of the prior is largest) may well be inappropriate.

Dr. Tuyl argues for the use of uniform objective priors in models with bounded parameter range, where the reference prior often tends to infinity in the extremes, with special reference to the binomial model. However, the parameter range depends on the (once again) arbitrary parameterization; thus in the binomial model, the perfectly adequate logit parameterization $\phi(\theta) = \log[\theta/(1-\theta)]$ has the real line as the parameter space. The reference prior is actually uniform of the alternative parameterization $\psi(\theta) = \arcsin[\sqrt{\theta}]$, which also has a bounded parameter space. I strongly believe that *any* proposal for an objective prior which is not consistent under one-to-one reparameterization is simply not worth to be considered.

Alternative formulations. The whole objective integrated approach described in this paper may in principle be done with *any* invariant continuous loss function, and I would be very surprised if the results turned out to be dramatically different. Ghosal suggests the use negative logarithm of Matusita's *affinity* (Matusita, 1967), defined as

$$\rho\{p_1, p_2\} = -\log \left[\int_{\mathcal{X}_1 \cap \mathcal{X}_2} \sqrt{p_1(x)p_2(x)} dx \right],$$

whenever the integral exists and the two distributions have non-disjoint supports.

The negative log-affinity between two normals with different location and the same scale is $\rho\{N(\cdot | \mu_1, \sigma), N(\cdot | \mu_2, \sigma)\} = (\mu_1 - \mu_2)^2 / (8\sigma^2)$, just proportional to the intrinsic discrepancy, $\delta\{N(\cdot | \mu_1, \sigma), N(\cdot | \mu_2, \sigma)\} = (\mu_1 - \mu_2)^2 / (2\sigma^2)$. Using this result and standard asymptotic arguments, it should be relatively simple to establish the asymptotic equivalence of both discrepancies for *regular* problems, where asymptotic normality may be established.

On the other hand, the negative log-affinity between two uniform densities within the family $\text{Un}(x | \theta - 1, \theta + 1)$, the interesting *non-regular* case which Ghosal mentions (where the supports are not nested and the intrinsic discrepancy cannot be used) is given by

$$\rho\{\text{Un}(\cdot | \theta_1 - 1, \theta_1 + 1), \text{Un}(\cdot | \theta_2 - 1, \theta_2 + 1)\} = -\log [1 - |\theta_1 - \theta_2|/2],$$

whenever $|\theta_1 - \theta_2| < 2$, and $+\infty$ otherwise. The reference prior function for the uniform model $\text{Un}(x | \theta - 1, \theta + 1)$ is easily seen to be the uniform $\pi(\theta) = 1$, and the corresponding reference posterior given a random sample $\mathbf{z} = \{x_1, \dots, x_n\}$ is the uniform $\pi(\theta | \mathbf{z}) = \text{Un}(\theta | x_{\max} - 1, x_{\min} + 1)$. The corresponding expected negative log-affinity loss of using θ_0 rather than the true θ value will then be

$$\bar{\rho}(\theta_0 | \mathbf{z}) = \frac{1}{2 - (x_{\max} - x_{\min})} \int_{x_{\max} - 1}^{x_{\min} + 1} -\log \left[1 - \frac{1}{2} |\theta_0 - \theta| \right] d\theta,$$

a concave function of θ_0 with a unique minimum at $\theta^*(\mathbf{z}) = (x_{\max} + x_{\min})/2$, which is the (very sensible) reference Bayes point estimator for this particular loss function.

As these examples illustrate, the general method proposed in this paper may indeed be used with any intrinsic loss function and, as Ghosal indicates, there may be some advantages in using the negative log-affinity. Also, the power divergence class mentioned by Ghosh is certainly an interesting, general divergence measure. However, the pragmatically important interpretation of the expected loss as an expected average log-likelihood ratio, and hence the automatic calibration of the procedure in those terms, would be lost if one uses anything different than the proposed logarithmic divergence. And, to answer Robert, yes, I believe this would be a bad thing to lose.

As Robert indicates, the Kullback-Leibler loss cannot be used when the densities have supports that depend on the parameters; this is precisely an important reason for using the minimum of the two possible KL divergences. Indeed, using the KL divergence $\kappa\{p_j | p_i\}$ rather than the intrinsic discrepancy (as Gómez-Villegas suggests), would limit the applicability to those regular problems where $\kappa\{p_j | p_i\}$ is always finite. For instance, one could not use this to test the parameter value in a uniform $\text{Un}(\cdot | 0, \theta)$ model (Example 5).

To answer a point raised by Mendoza, I am certainly not suggesting that loss functions should necessarily be symmetric. In a context specific situation, this would naturally depend on the problem. However, in a pure inferential situation, where one is only interested in the true value of the parameter, one may well like to see some form of symmetry. This is not a requirement, but it may be a welcome feature when it happens, as in normal variance problem of Example 2.

Professor Robert suggests the use of an intrinsic prior $\pi(\boldsymbol{\lambda} | \boldsymbol{\theta}_0)$ to get rid of the nuisance parameter in the formulation of the loss (Definition 5). I do not like the idea of being forced to introduce further concepts then required (as the intrinsic prior formalism) but, more importantly, I believe the defining the discrepancy between a point and a set as the minimum discrepancy between the point and all the elements in the family has a long tradition in mathematics, and may be expected to produce appropriate results. The examples analyzed suggest that this is surely the case.

The formulation of Robert to deal with cases where the nuisance parameter is ill-defined under the null is certainly appropriate, and this has already been successfully used in practice. A relatively simple example is the Hardy-Weinberg equilibrium of Example 9, further detailed in Bernardo and Tomazella (2010).

Robert suggests that the use of the intrinsic discrepancy may not always work, and quotes a Cauchy model $\text{Ca}(x | \theta, 1)$ as a possible example. I am convinced that pathological examples may be found where the expected intrinsic discrepancy cannot be evaluated, but this is certainly *not* the case with Cauchy data. In this problem, the two KL divergences are identical, so that the intrinsic discrepancy is just one of them (represented in Figure 8). This is a location model; hence the reference prior is uniform and the reference posterior $\pi(\theta | \mathbf{z})$ is just the normalized likelihood. This may be used to obtain the posterior expected intrinsic discrepancy $d(\theta_0 | \mathbf{z})$ by one-dimensional numerical integration. To illustrate this, I simulated a random sample $\mathbf{z} = \{x_1, \dots, x_n\}$ of size $n = 20$ from a Cauchy $\text{Ca}(x | 3, 1)$. The upper panel of Figure 11 shows the corresponding reference posterior of θ , and the lower panel the intrinsic statistic function $d(\theta_0 | \mathbf{z}) = n \int_{\mathbb{R}} \kappa\{\theta_0 | \theta\} \pi(\theta | \mathbf{z}) d\theta$, where $\kappa\{\theta_0 | \theta\} = \kappa\{\theta | \theta_0\}$ is the KL divergence between $\text{Ca}(x | \theta_0, 1)$ and $\text{Ca}(x | \theta, 1)$ and $\pi(\theta | x_1, \dots, x_n) \propto \prod_{i=1}^n \text{Ca}(x_i | \theta, 1)$. This function has a unique minimum, the intrinsic estimator $\theta^* \approx 3.27$, and has the value $\log[1000] \approx 6.9$ at $\theta_0 = 1.96$ and $\theta_1 = 4.58$. Hence, values smaller than θ_0 or larger than θ_1 would be rejected at

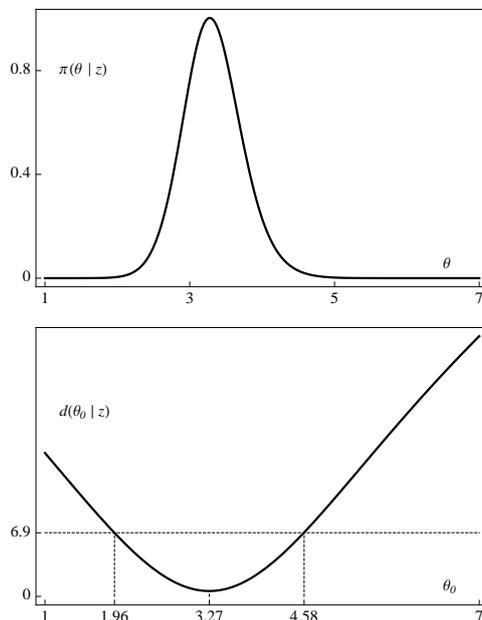


Figure 11: Reference posterior and intrinsic statistic function for the location parameter of a Cauchy $\text{Ca}(x | \theta, 1)$, given a random sample of size $n = 20$ simulated for a Cauchy, $\text{Ca}(x | \theta, 1)$.

that particular level, and the interval $[\theta_0, \theta_1]$ is an intrinsic credible interval, whose posterior probability may be found to be 0.9956. It may be verified that the procedure works even with a sample of size one where, if x is observed, the reference posterior is the Cauchy $\text{Ca}(\theta | x, 1)$ and the intrinsic estimator is just $\theta^* = x$.

In his opening remarks Professor Clarke provides a lucid, concise summary of what integrated reference analysis is all about. His suggestion of using a prior to find a reference likelihood, then using this to obtain a reference prior, and cycle until convergence is obtained is intriguing. A large collection of important solved case studies would be necessary however before such a programme could be appropriately evaluated.

As Clarke mentions, I have indeed assumed that the available data have been generated from some member of a well specified parametric family of probability models. This is certainly an idealized situation, but one which is systematically made in the literature. It is clear however that, even under this simplifying assumption, there has been an enormous amount of different (often incompatible) suggested procedures for both estimation and hypothesis testing. We believe that some clarification is in order before proceeding further, and we argue that foundational arguments provide the best tools for this task. That said, model mis-specification and robustness analysis are certainly two very important topics to consider. The idea of using leave-one-out techniques to analyze robustness, as in the particular proposal which Clarke sketches, is certainly worth exploring.

Both non-parametric density estimation, and parametric model elaboration are promising options to deal with the possibility of model mis-specification. The former is however bound to be crucially dependent on the particular density estimation procedure chosen, and we all know that there is not yet a consensus on how this may be better done. I find far more attractive the idea of model elaboration. Indeed, as mentioned in Section 2.3, the hypothesis testing scenario may often be precisely described in those terms: one begins from a model, $\mathcal{M}_0 \equiv \{p(\mathbf{z} | \theta_0, \lambda), \lambda \in \Lambda\}$ in my original notation or $p(\mathbf{z} | \theta)$ in that used by Clarke, and this is embedded into a more general model, $\mathcal{M}_z \equiv \{p(\mathbf{z} | \theta, \lambda), \lambda \in \Lambda, \theta \in \Theta\}$, constructed to include promising departures from \mathcal{M}_0 .

Specific queries. Van der Linde asks about the performance of reference intrinsic estimators in small n large p scenarios, where the number of parameters is large relative to the sample size, resulting in unstable conventional estimators. I am not aware of detailed systematic reference analysis of this type of problems, but my attitude would be to introduce a suitable hierarchical structure modelling the plausible relations between the parameters, and then using a joint reference prior for the hyperparameters thus introduced, derived from the corresponding integrated model.

Ghosal is certainly right in stressing the importance of multiple hypothesis testing problems. Some relevant results in this direction (from the approach proposed here) may be found in Bernardo and Rueda (2002), where simultaneous testing of $H_{0i} \equiv \{\mu_i = 0\}$, for $i = 1, \dots, k$, is considered in a multivariate $N_k(\mathbf{x} | \boldsymbol{\mu}, \Sigma)$ context; this is seen to provide a resolution of Rao's paradox. Further work is however needed in this area.

Both Singurwalla and Tuyl question the virtue of the non-dependence of the posterior of the multinomial parameters (Example 4) on the number m of categories. Well, I cannot imagine how an objective estimate of the proportion of votes which a party will obtain in an election given a random sample of results, should depend on something else than the votes for that party and the total number of votes counted. In particular, this should not depend on whether the small parties are jointly labeled as "small" or individually considered.

I am afraid I do not understand Singpurwalla's comment on the uniform model (Example 5). In that case, the expected intrinsic loss under repeated sampling is $(\theta/\theta_0)^n - n \log(\theta/\theta_0)$. This increases with n (not linearly with θ_0) for any $\theta \neq \theta_0$, thus leading to always rejecting a false null for sufficiently large samples. For fixed sample size n , it increases (not linearly) as the ratio θ/θ_0 moves away from one, producing a function of θ/θ_0 with a unique minimum at $\theta/\theta_0 = 1$.

Tuyl seems to prefer HPD regions to intrinsic regions because the later may "sacrifice values with high likelihood". However, it is *not* likelihood what drives HPD, but posterior density, and this totally depends on the (arbitrary) parametrization used. A Bayesian should always prefer values with minimum expected loss, and only the rather naïve, not invariant, 0–1 loss will yield HPD regions (and then only in the particular parametrization chosen). And the idea that in a binomial situation the m.l.e. estimator r/n is "perfectly adequate" even in extreme situations is hardly acceptable: would you really quote to your Provost precisely 0 as your best estimate for the proportion of AIDS infected persons in the campus, just because you have not observed any in a random sample of n ? (and this, even for small n values!) Incidentally, in one parameter regular models (where asymptotic normality may be verified) Jeffreys prior has been found to be optimal from so many viewpoints (of

which the reference algorithm is only one example) that using something else in those simple conditions is, to say the least, rather bizarre.

In his comments to the trinomial example, Tuyl seems to forget his earlier uniform recommendation on the original parameterization for the multinomial, and suggests a uniform prior on a different parametrization, a less than consistent behaviour, I would say. He mentions the obvious fact that posteriors should be proper; *by definition* a reference posterior *must* be proper (see Berger, Bernardo and Sun, 2009, for a detailed discussion). For a recent detailed reference analysis of the trinomial example, (were the posteriors are of course all proper) see Bernardo and Tomazella (2010). Finally, the notion by Tuyl that the simple binomial model is not “regular” and needs special *ad hoc* techniques to produce sensible objective Bayesian answers is, in my view, rather far removed from reality.

Conclusions. As Mendoza cunningly puts it, I have tried to present here my subjective view of what objective Bayesian methods should be: model divergence based loss functions, information-based reference priors, and the machinery of decision theory, can work together to derive attractive Bayesian solutions to pure inference problems. These solutions could be labeled objective, both in the narrow sense of only using model and data, and in the larger sense of making possible a much needed form of consensus.

As both Clarke and Strenger nicely remind, the integrated reference analysis advocated here is intended to be a *benchmark* against which other analysis, with context dependent loss functions and/or subjectively assessed prior functions could be compared, to help in the evaluation of the impact of these, possibly debatable inputs, in the results finally presented.

On the apparently more polemic aspect of this paper, it should be obvious to the reader that I do *not* agree with Pericchi that the “probability of a hypothesis given the data is perhaps the most relevant question for a scientist”. To my perception, the relevant question is whether or not available data are compatible with a hypothesis, and this is a decision problem which requires a loss function. Posterior probabilities are the answer if, *and only if*, the scientist preferences are well described by the naïve zero-loss function, a less than likely situation. Besides, this forces a totally different objective prior structure (unnecessary otherwise) than that used for estimation, and this entails the difficulties discussed above. I firmly believe that continuous invariant loss functions and relevant reference priors are more appropriate for the job.

To conclude, I am certainly not claiming to have discovered the ultimate pass through the statistical mountains but, as my *maestro* suggests, I am certainly enjoying the ride. Thanks again to all of you.

ADDITIONAL REFERENCES IN THE DISCUSSION

- Aitkin, M., Boys, R. J. and Chadwick, T. (2005). Bayesian point null hypothesis testing via the posterior likelihood ratio. *Statist. Computing* **15**, 217–230.
- Bayes, T. R. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc. London* **53**, 370–418.
- Berger, J. O. and Delampady, M. (1987). Testing precise hypothesis. *Statist. Science* **2**, 317–352 (with discussion).

- Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p -values and evidence *J. Amer. Statist. Assoc.* **82**, 112–139 (with discussion).
- Berger, J.O. (1994). An overview of robust Bayesian analysis (with discussion), *Test* **3**, 5–124.
- Berger, J. O. and Pericchi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *J. Amer. Statist. Assoc.* **91**, 109–122.
- Bernardo, J. M. (1979b). Expected information as expected utility. *Ann. Statist.* **7**, 686–690.
- 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).
- Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. New York: Wiley.
- Casella, G. and Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc.* **82**, 106–111.
- Casella G. and Moreno E. (2009). Assessing robustness of intrinsic tests of independence in two-way contingency tables *J. Amer. Statist. Assoc.* **104**, 1261–1271.
- Consonni, G. and La Rocca, L. (2008). Tests based on intrinsic priors for the equality of two correlated proportions, *J. Amer. Statist. Assoc.* **103**, 1260–1269.
- Dawid, A. P. (1982). The well-calibrated Bayesian. *J. Amer. Statist. Assoc.* **77**, 605–610.
- Dawid, A. P. (1984). Statistical theory: The prequential approach. *J. Roy. Statist. Soc. A* **147**, 278–292.
- Druilhet, P. and Marin, J.-M. (2007). Invariant HPD credible sets and MAP estimators. *Bayesian Analysis* **2**, 681–692.
- Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.
- Geisser, S. (1984). On prior distributions for binary trials. *Amer. Statist.* **38**, 244–251.
- Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis* **3**, 445
- George, E. I., Liang, F. and Xu, X. (2006). Improved minimax predictive densities under Kullback-Leibler loss. *Ann. Statist.* **34**, 78–92.
- Ghosh, M., Mergel, V. and Datta, G. S. (2008). Estimation, prediction and the Stein phenomenon under divergence loss. *J. Multivariate Analysis* **99**, 1941–1961.
- Ghosh, M. and Mergel, V. (2009). On the Stein phenomenon under divergence loss and an unknown variance-covariance matrix. *J. Multivariate Analysis* **100**, 2331–2336.
- Gómez-Villegas, M. A., Maín, P. and Sanz, L. (2009). A Bayesian analysis for the multivariate point null testing problem. *Statistics* **43**, 379–391.
- Gómez-Villegas, M. A., Maín, P. and Susi, R. (2008). Extreme inaccuracies in Gaussian Bayesian networks. *J. Multivariate Analysis* **99**, 1929–1940.
- Günel, E. and Dickey, J. (1974). Bayes factors for independence in contingency tables, *Biometrika* **61**, 545–557.
- Jaynes, E. T. (1976). Confidence intervals vs Bayesian intervals. *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* (W. L Harper and C. A. Hooker, eds.) Dordrecht: Reidel 175–257 (with discussion).
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine* **2**, 696–701.

- Jovanovic, B. D. and Levy, P. S. (1997). A look at the Rule of Three. *Amer. Statist.* **51**, 137–139.
- Lindley, D. V. (1969). *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge: University Press
- Lindley, D. V. (2006). *Understanding Uncertainty*. Chichester: Wiley
- Lindley, D. V., East, D. A. and Hamilton, P. A. (1960). Tables for making inferences about the variance of a Normal distribution. *Biometrika* **47**, 433–437.
- MacLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley
- Matusita, K. (1967). On the notion of affinity of several distributions and some of its applications. *Ann. Inst. Statist. Math.* **19**, 181–192.
- Morris, C. N. (1987a). Discussion of Casella and Berger (1987). *J. Amer. Statist. Assoc.* **82**, 106–111.
- Morris, C. N. (1987b). Discussion of Berger and Sellke (1987). *J. Amer. Statist. Assoc.* **82**, 112–122.
- Moreno, E., Bertolino, F. and Racugno, W. (1998). An intrinsic limiting procedure for model selection and hypothesis testing. *J. Amer. Statist. Assoc.* **93**, 1451–1460.
- Moreno E., Girón, F.J., Vazquez-Polo, F. J. and Negrin, M. A. (2010). Optimal healthcare decisions: Comparing treatments on a cost-effectiveness basis. *Eur. J. Oper. Res.* **204**, 180–187.
- Pericchi, L. R. (2010). How large should be the training sample? *Frontiers of Statistical Decision Making and Bayesian Analysis. In Honor of James O. Berger* (M.-H. Chen, D. K. Dey, P. Müller, D. Sun and K. Ye, eds.) New York: Springer,
- Philippe, A. and Rousseau, J. (2003). Non-informative priors for Gaussian long-memory processes. *Bernoulli* **8**, 451–473.
- Polson, N. and Scott, J. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics 9* (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,
- Robert, C. (1996). *Méthodes de Monte Carlo par Chaînes de Markov*. Paris: Economica.
- Robert, C. and Casella, G. (1994). Distance penalized losses for testing and confidence set evaluation. *Test* **3**, 163–182.
- Robert, C. and Rousseau, J. (2002). A mixture approach to Bayesian goodness of fit. *Tech. Rep.*, Université Paris Dauphine, France..
- Román, L. (2010). *Funciones Iniciales de Referencia para Predicción Bayesiana*. Ph.D. Thesis, Universidad de Valencia, Spain.
- Singpurwalla, N. D. (2002a). Some cracks in the empire of chance: Flaws in the foundations of reliability. *Internat. Statist. Rev.* **70**, 53–78 (with discussion).
- Singpurwalla, N. D. (2002b). On causality and causal mechanisms. *Internat. Statist. Rev.* **70**, 198–206.
- Singpurwalla, N. D. (2006). *Reliability and Risk: A Bayesian Perspective*. Chichester: Wiley.
- Stigler, S. M. (1982). Thomas Bayes' Bayesian Inference. *J. Roy. Statist. Soc. A* **145**, 250–258.
- Titterton, D., Smith, A. and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: Wiley.

- Tuyl, F., Gerlach, R. and Mengersen, K. (2008). A comparison of Bayes–Laplace, Jeffreys, and other priors: The case of zero events. *Amer. Statist.* **62**, 40–44.
- Wang, Q., Stefanski, L., Genton, M. and Boos, D. (2009). Robust time series analysis via measurement error modeling. *Statistica Sinica* **19**, 1263–1280.
- Yuan, A. and Clarke, B. (1999). A minimally informative likelihood for decision analysis: illustration and robustness. *Can. J. Statist.* **27**, 649–665.