Bayes and Discovery: Objective Bayesian Hypothesis Testing

José M. Bernardo
Universitat de València, Spain

Abstract
Hypothesis testing is formulated from a decision theoretical viewpoint. The combined use of intrinsic discrepancy, an invariant information-based loss function, and conventional reference priors provides an objective Bayesian solution to precise hypothesis testing problems which easily integrates with the standard formulation of objective Bayesian point and region estimation.

1 Precise Hypothesis Testing

Let $\boldsymbol{z}$ be the available data which are assumed to have been generated as one random observation from model $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z} \,|\, \boldsymbol{\omega}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$. Often, but not always, data will consist of a random sample $\boldsymbol{z} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from some distribution $q(\boldsymbol{x} \,|\, \boldsymbol{\omega})$, with $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$; in this case $p(\boldsymbol{z} \,|\, \boldsymbol{\omega}) = \prod_{i=1}^{n} q(\boldsymbol{x}_i \,|\, \boldsymbol{\omega})$ and $\boldsymbol{\mathcal{Z}} = \boldsymbol{\mathcal{X}}^n$. Let $\boldsymbol{\theta}(\boldsymbol{\omega})$ be the vector of interest. Without loss of generality, the model may explicitly be expressed in terms of $\boldsymbol{\theta}$ so that $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, where $\boldsymbol{\lambda}$ is some appropriately chosen nuisance parameter vector. Let $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \pi(\boldsymbol{\lambda} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$ be the assumed prior, and let $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{z})$ be the corresponding marginal posterior distribution of $\boldsymbol{\theta}$. Appreciation of the inferential contents of $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{z})$ may be enhanced by providing both point and region estimates of the vector of interest $\boldsymbol{\theta}$, and by declaring whether or not some context-suggested specific value $\boldsymbol{\theta}_0$ is compatible with the observed data $\boldsymbol{z}$ (precise hypothesis testing). 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.

Let $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ describe, as a function of the (unknown) parameter values $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ which have generated the available data, *the loss* to be suffered if, working with model $\mathcal{M}_{\boldsymbol{z}}$, the value $\boldsymbol{\theta}_0$ were used as a proxy for the unknown value of $\boldsymbol{\theta}$. Point estimation, region estimation and hypothesis testing procedures may all be appropriately described as specific decision problems using a common prior distribution and a common loss function of this type. The results, which are obviously all conditional on the assumed model $\mathcal{M}_{\boldsymbol{z}}$, may dramatically depend on the particular choices made for both the prior and the loss functions but, given the available data $\boldsymbol{z}$, they all only depend on those through the corresponding posterior expected loss,

$$\overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z}) = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, \pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{z})\, d\boldsymbol{\theta} d\boldsymbol{\lambda}. \tag{1}$$

As a function of $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, the expected loss $\overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$ 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.

In this paper, we will concentrate on precise hypothesis testing, with objective reference priors. For a more general perspective and many examples, see Bernardo (2011) and references therein.

1.1 Decision Theoretic Formulation

Consider a value $\boldsymbol{\theta}_0$ of the vector of interest which deserves special consideration, either because assuming $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ would noticeably simplify the model, or because there are additional context-specific arguments suggesting that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Intuitively, the value $\boldsymbol{\theta}_0$ should be judged to be *incompatible* with

the observed data $\boldsymbol{z}$ if the posterior expected loss $\overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$ of using $\boldsymbol{\theta}_0$ as a proxy for $\boldsymbol{\theta}$ is too large. This notion is now made precise.

Formally, testing the hypothesis $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ 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 specification of a loss function $\ell_h\{a_i, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ measuring the consequences of accepting or rejecting H_0 as a function of the actual parameter values. By assumption, a_0 means to *act as if* H_0 were true, that is to work with the model $\mathcal{M}_0 = \{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}\}$, while a_1 means to reject this simplification and to keep working with model $\mathcal{M}_{\boldsymbol{z}}$=$\{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$. Alternatively, an already established model $\mathcal{M}_0$ may have been embedded into a more general model $\mathcal{M}_{\boldsymbol{z}}$, constructed to include promising departures from $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and it is required to verify whether presently available data $\boldsymbol{z}$ are still compatible with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, or whether the extension to $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is really necessary. Given the available data $\boldsymbol{z}$, 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 $\int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} [\ell_h\{a_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - \ell_h\{a_1, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}] \, \pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{z}) \, d\boldsymbol{\theta} d\boldsymbol{\lambda} > 0$. Hence, only the loss difference $\Delta\ell_h\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \ell_h\{a_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - \ell_h\{a_1, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, which measures the *advantage* of rejecting $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ 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

$$\Delta\ell_h\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - \ell_0$$

where, as mentioned above (and precisely as in estimation problems), $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ describes the non-negative loss to be suffered if $\boldsymbol{\theta}_0$ were used as a proxy for $\boldsymbol{\theta}$. Since $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}_0, \boldsymbol{\lambda})\} = 0$, so that $\Delta\ell_h\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}_0, \boldsymbol{\lambda})\} = -\ell_0$, the value $\ell_0 > 0$ describes (in the same loss units) the context-dependent non-negative advantage of accepting $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ 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 $\overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$, the posterior expectation of $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, is larger than ℓ_0. Thus, as intuition suggested, the solution to the precise hypothesis testing decision problem posed is found in terms of the value of the expected loss $\overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$ of using $\boldsymbol{\theta}_0$ as a proxy for the unknown value of $\boldsymbol{\theta}$.

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) $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0 \,|\, \boldsymbol{z}) < p_0$ 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 requires the use of a non-regular "sharp" prior, concentrating a positive probability mass at $\boldsymbol{\theta}_0$. With no mention of the (rather naïve) loss structure which is implicit in the formulation, this type of solution was early advocated by Jeffreys (1961). Notice however, that this formulation implies the use of radically different priors for hypothesis testing than those used for estimation, and a different prior for each value to be tested. Moreover, this formulation is known to lead to the difficulties associated with Lindley's paradox (Lindley, 1957; Bartlett, 1957; Robert, 1993).

There are many real world situations where there is really a concentration of prior probability around particular value, and a sound Bayesian analysis should then certainly use this information. Under some conditions, those situations may well be described with a probability mass at a (measure zero) point. However, even in these cases, robustness concerns suggest that it may well be worth exploring the consequences of using a regular reference prior with the same data, if only to verify the possible dependence of the conclusions reached on the particular prior assumptions made.

Using the quadratic loss function leads to rejecting a $\boldsymbol{\theta}_0$ value whenever its Euclidean distance to $\mathrm{E}[\boldsymbol{\theta} \,|\, \boldsymbol{z}]$, 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, and the same prior for all values to be tested. 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 shown below, a judicious choice of the loss function leads to calibrated expected losses, where the relevant threshold constant has an immediate, operational interpretation.

2 The Intrinsic Divergence Loss

Conditional on model $\mathcal{M}_{\boldsymbol{z}}$=$\{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, 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 $\mathcal{M}_{\boldsymbol{z}}$, 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.

A reasonable measure of the dissimilarity $\delta\{p_{\boldsymbol{z}}, q_{\boldsymbol{z}}\}$ between two probability densities $p(\boldsymbol{z})$ and $q(\boldsymbol{z})$ for a random vector $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ should surely be non-negative, zero if (and only if), $p(\boldsymbol{z}) = q(\boldsymbol{z})$ almost everywhere, and preferably symmetric. Moreover it should be invariant under one-to-one transformations of $\boldsymbol{z}$; indeed, if $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{z})$ is such a transformation and J is the appropriate Jacobian, $p_{\boldsymbol{y}} = p_{\boldsymbol{z}}/|J|$, and $q_{\boldsymbol{y}} = q_{\boldsymbol{z}}/|J|$ are expressions of precisely the same uncertainties and, therefore, one should certainly have $\delta\{p_{\boldsymbol{z}}, q_{\boldsymbol{z}}\} = \delta\{p_{\boldsymbol{y}}, q_{\boldsymbol{y}}\}$. 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. These desiderata are all satisfied by the *intrinsic discrepancy* (Bernardo and Rueda, 2002), a divergence measure which has both an information theoretical justification, and a simple operational interpretation in terms of average log-density ratios.

Definition 1 *The* intrinsic discrepancy $\delta\{p_1, p_2\}$ *between two probability distributions for the random vector* $\boldsymbol{z}$ *with densities* $p_1(\boldsymbol{z})$, $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}_1$, *and* $p_2(\boldsymbol{z})$, $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}_2$, *is*

$$\delta\{p_1, p_2\} = \min\left[\, \kappa\{p_1 \,|\, p_2\},\, \kappa\{p_2 \,|\, p_1\} \,\right] \tag{2}$$

where $\kappa\{p_j \,|\, p_i\} = \int_{\boldsymbol{\mathcal{Z}}_i} p_i(\boldsymbol{z}) \log[p_i(\boldsymbol{z})/p_j(\boldsymbol{z})]\, d\boldsymbol{z}$ *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,*

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

The intrinsic discrepancy $\delta\{p_1, p_2\}$ is the minimum average 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. This may be used to define an appropriate loss function for the decision problem considered in this paper as the intrinsic discrepancy between the model, labelled by $(\boldsymbol{\theta}, \boldsymbol{\lambda})$, and the family $\mathcal{M}_0$ of models which satisfy the hypothesis to be tested:

Definition 2 *Consider* $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$. *The* intrinsic discrepancy 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 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\mathcal{M}_0 = \{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}\}$,

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} = \delta\{p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \mathcal{M}_0\} = \inf_{\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}} \delta\{p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0), p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})\}. \tag{3}$$

Notice the complete generality of Definition 2; this may be used with either discrete or continuous data models (in the discrete case, the integrals in Definition 1 will obviously be sums), and with either discrete or continuous parameter spaces of any dimensionality.

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

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

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

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} = \ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{t}}\}.$$

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

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} = n\, \ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{x}}\}.$$

An interesting interpretation of the intrinsic discrepancy loss follows directly from Definitions 1 and 2. Indeed, $\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\}$ 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 discrepancy 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 $\overline{\ell}_\delta(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$, the posterior expectation of $\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\}$, 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 $\{\log 100, \log 1000, \log 10000\} \approx \{4.6, 6.9, 9.2\}$.

Under regularity conditions, the intrinsic discrepancy loss has an alternative expression which is generally much simpler to compute:

Theorem 1 (Juárez, 2004, Sec. 2.4) *If the support of* $p(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ *is convex for all* $(\boldsymbol{\theta}, \boldsymbol{\lambda})$*, then the intrinsic discrepancy loss may also be written as*

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} = \min\left[\inf_{\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}} \kappa\{\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0 \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}\},\ \inf_{\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}} \kappa\{\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0\}\right], \tag{4}$$

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

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

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

Example 1 (Multivariate normal model). Let $\boldsymbol{z} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a random sample from a k-variate normal distribution $\mathrm{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with known covariance matrix $\boldsymbol{\Sigma}$. The KL divergence of $\mathrm{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ from $\mathrm{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ 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 $\boldsymbol{\mu}$.

2.1 Approximations

Under regularity conditions, the result of Example 1 may be combined with conventional asymptotic results to obtain large sample approximations to intrinsic discrepancy losses:

Theorem 2 (Bernardo, 2011) *Let data $\boldsymbol{z} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ consist of a random sample from $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$, 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,*

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

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

The invariance of the intrinsic discrepancy 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 is a one-parameter example of this technique, which makes use of the variance stabilization transformation.

Theorem 3 (Bernardo, 2005b) *Let $\boldsymbol{z} = \{x_1, \ldots, x_n\}$ be a random sample of size n from model $p(x \,|\, \theta)$, and let $\tilde{\theta}_n = \tilde{\theta}_n(\boldsymbol{z})$ 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,*

$$\ell\{\theta_0, \theta \,|\, \mathcal{M}_{\boldsymbol{z}}\} = \frac{n}{2}\,[\phi(\theta_0) - \phi(\theta)]^2 + o(1).$$

3 Reference Analysis

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, frequentist 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 is this limited, but precise sense. For lively discussions of this, and related issues, see Bernardo (1997), Berger (2006), and ensuing discussions.

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 synthesizing the basic definition and briefly reviewing some recent developments.

Note first that the same mathematical concepts which lie behind the definition of the intrinsic discrepancy provide the intuitive basis for the definition of reference priors. Indeed, for the one parameter model $\mathcal{M} = \{p(\boldsymbol{z} \,|\, \theta), \boldsymbol{z} \in \mathcal{Z}, \theta \in \Theta \subset \Re\}$, the intrinsic discrepancy $I\{p_\theta \,|\, \mathcal{M}\} = \delta\{p(\boldsymbol{z}, \theta), p(\boldsymbol{z})\, p(\theta)\}$ between the joint prior $p(\boldsymbol{z}, \theta)$ and the product of their marginals $p(\boldsymbol{z})\, p(\theta)$ 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 $\boldsymbol{z}$ 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 $\{\pi_k(\theta)\}_{i=1}^k$ 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 (Berger, Bernardo and Sun, 2009). *Let $\boldsymbol{z}^{(k)} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\}$ denote k conditionally independent observations from $\mathcal{M}_{\boldsymbol{z}}$. Then, the reference prior is defined as an appropriate limit of*

$$\pi_k(\theta) \propto \exp\left\{ \mathrm{E}_{\boldsymbol{z}^{(k)} \,|\, \theta}\big[\log p_h(\theta \,|\, \boldsymbol{z}^{(k)}) \big] \right\} \tag{5}$$

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

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 $\{\boldsymbol{z}_1^{(k)}, \ldots, \boldsymbol{z}_s^{(k)}\}$ of s replications $\boldsymbol{z}_i^{(k)} = \{\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{ik}\}$ of size k from the original model $p(\boldsymbol{z} \,|\, \theta)$, computes the corresponding s posterior densities at θ, $\{p_h(\theta \,|\, \boldsymbol{z}_j^{(k)})\}_{j=1}^s$, and approximates the reference prior at this point by its logarithmic average,

$$\pi(\theta) \approx \exp\Big\{ \frac{1}{s} \sum\nolimits_{j=1}^{s} \log p_h(\theta \,|\, \boldsymbol{z}_j^{(k)}) \Big\}. \tag{6}$$

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 $p(\boldsymbol{x} \,|\, \theta)$ is asymptotically normal with standard deviation $s(\tilde{\theta}_n)/\sqrt{n}$, where $\tilde{\theta}_n$ is a consistent estimator of θ, then the reference prior is $\pi(\theta) = s(\theta)^{-1}$. This includes as a particular case the famous Jeffreys-Perks prior (Jeffreys, 1946, independently formulated by Perks, 1947)

$$\pi(\theta) \propto i(\theta)^{1/2}, \quad i(\theta) = \mathrm{E}_{\boldsymbol{x} \,|\, \theta}[-\partial^2 \log p(\boldsymbol{z} \,|\, \theta)/\partial\theta^2]. \tag{7}$$

Similarly, if $p(\boldsymbol{x} \,|\, \theta)$ is a non regular model with a support $S(\theta)$ which depends on the paramenter in the form $S(\theta) = \{x;\, a_1(\theta) < x < a_2(\theta)\}$, 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

$$\pi(\theta) \propto \mathrm{E}_{\boldsymbol{x} \,|\, \theta}[|\partial \log p(\boldsymbol{z} \,|\, \theta)/\partial\theta|]. \tag{8}$$

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).

If the model has more than one parameter, the required joint reference prior is derived sequentially. Thus, if the model is $p(\boldsymbol{z} \,|\, \theta, \lambda)$ 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* $p(\boldsymbol{z} \,|\, \theta) = \int_\Lambda p(\boldsymbol{z} \,|\, \theta, \lambda)\, \pi(\lambda \,|\, \theta)\, d\lambda$, 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 derives the corresponding limiting result.

In general, reference priors are sequentially derived with respect to an ordered parameterization. Thus, given a model $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z} \,|\, \boldsymbol{\omega}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ with m parameters, the reference prior with respect to a particular ordered parameterization $\boldsymbol{\phi}(\boldsymbol{\omega}) = \{\phi_1, \ldots, \phi_m\}$ (where the ϕ_i's are ordered by inferential importance) is sequentially obtained as $\pi(\boldsymbol{\phi}) = \pi(\phi_m \,|\, \phi_{m-1}, \ldots, \phi_1) \times \cdots \times \pi(\phi_2 \,|\, \phi_1)\, \pi(\phi_1)$. Unless all reference priors turn out to be proper, the model must be endowed with an appropriate compact approximation to the parameter space $\{\boldsymbol{\Omega}_j\}_{j=1}^{\infty} \subset \boldsymbol{\Omega}$, 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 $\pi(\theta \,|\, \boldsymbol{z})$ of the quantity of interest.

4 Objective Bayesian Hypothesis Testing

With the loss function chosen to be the intrinsic discrepancy loss, all that is required to define an objective Bayesian testing procedure is to specify an objective prior distribution. It will not come as a surprise that we recommend the use of a reference prior. Thus, one must obtain the posterior expectation of the intrinsic discrepancy loss with respect to the appropriate joint reference posterior

$$d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z}) = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\}\, \pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{z})\, d\boldsymbol{\theta} d\boldsymbol{\lambda}. \tag{9}$$

and decide whether or not this is big enough to reject that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. The function $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$ is the relevant *intrinsic* test statistic, a direct measure of the incompatibility of $\boldsymbol{\theta}_0$ with the data $\boldsymbol{z}$ in terms of the expected average log-likelihood ratio against the null.

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

Example 2 (Poisson data). Let $\boldsymbol{z} = \{x_1, \ldots, x_n\}$ be a random sample from a Poisson model, so that $p(x \,|\, \lambda) = \text{Po}(x \,|\, \lambda) = e^{-\lambda}\lambda^x / x!$. This is a regular model, and using (7), the reference prior is immediately found to be $\pi(\lambda) = \lambda^{-1/2}$. This leads to the gamma reference posterior $\pi(\lambda \,|\, \boldsymbol{z}) = \pi(\lambda \,|\, t, n) = \text{Ga}(\lambda \,|\, t + 1/2, n) \propto e^{-n\,\lambda} \lambda^{t-1/2}$, with $t = \sum_{j=1}^{n} x_j$.

Using Definition 2 and the additive property of the intrinsic discrepancy, the intrinsic discrepancy loss of using λ_0 as a proxy for λ with a random sample of size n from a Poisson distribution with parameter λ is

$$\delta\{\lambda_0, \lambda \,|\, \mathcal{M}_{\boldsymbol{z}}\} = n\, \delta\{\lambda_0, \lambda \,|\, \mathcal{M}_x\} = n \, \min\left\{ \text{E}_{x \,|\, \lambda}\left[\log \frac{\text{Po}(x \,|\, \lambda)}{\text{Po}(x \,|\, \lambda_0)} \right], \text{E}_{x \,|\, \lambda_0}\left[\log \frac{\text{Po}(x \,|\, \lambda_0)}{\text{Po}(x \,|\, \lambda)} \right] \right\}$$

which immediately yields

$$\delta\{\lambda_0, \lambda \,|\, \mathcal{M}_{\boldsymbol{z}}\} = \begin{cases} n\,(\lambda - \lambda_0 + \lambda_0 \log \frac{\lambda_0}{\lambda}) & \text{if} \quad \lambda_0 \le \lambda, \\ n\,(\lambda_0 - \lambda + \lambda \log \frac{\lambda}{\lambda_0}) & \text{if} \quad \lambda_0 \ge \lambda. \end{cases} \tag{10}$$

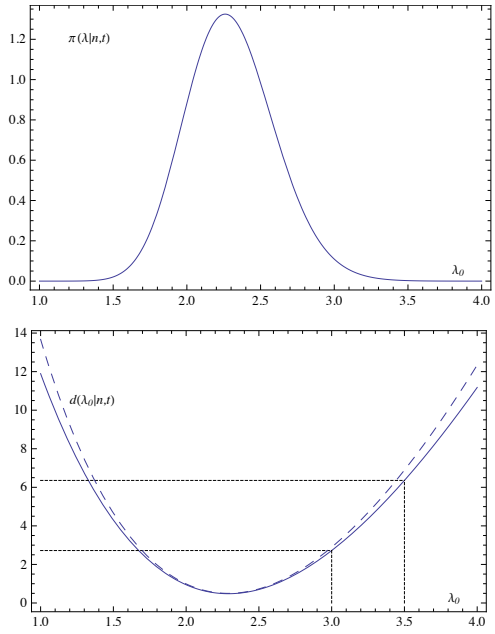

Figure 1: Posterior reference analysis of the parameter of a Poisson model.

a non-negative concave function of λ and λ_0, with minimum equal to zero when $\lambda = \lambda_0$.

The intrinsic statistic $d(\lambda_0 \,|\, \boldsymbol{z})$ is the corresponding reference posterior expectation,

$$d(\lambda_0 \,|\, \boldsymbol{z}) = \int_0^\infty \delta\{\lambda_0, \lambda \,|\, \mathcal{M}_{\boldsymbol{z}}\}\, \mathrm{Ga}(\lambda \,|\, t + 1/2, n)\, d\lambda. \tag{11}$$

This has no simple analytical expression, but is easily computed by numerical integration. Moreover, using Theorem 3 and the fact that the sampling distribution of the sufficient and consistent mle, $\hat{\lambda} = \bar{x}$ is asymptotically normal with standard deviation $\sqrt{\lambda}/\sqrt{n}$, one finds

$$\begin{aligned} d(\lambda_0 \,|\, \boldsymbol{z}) &\approx \int_0^\infty \frac{n}{2} \left(2\sqrt{\lambda_0} - 2\sqrt{\lambda}\right)^2 \mathrm{Ga}(\lambda \,|\, t + 1/2, n)\, d\lambda \\ &= 1 + 2\,t + 2n\,\lambda_0 - 4\sqrt{n\,\lambda_0}\, \frac{\Gamma(t+1)}{\Gamma(t+1/2)} \qquad (12) \\ &\approx 1 + 2\,t + 2n\,\lambda_0 - 4\sqrt{n\,\lambda_0} \left(\sqrt{t} + \frac{1}{8\sqrt{t}}\right). \qquad (13) \end{aligned}$$

To illustrate the type of results obtained, a sample of size $n = 25$ was simulated from a Poisson distribution with parameter $\lambda = 2$ resulting in $t = 57$. The corresponding reference posterior density is plotted in the top panel of Figure 1. The expected intrinsic discrepancy loss $d(\lambda_0 \,|\, t, n)$ (both computed from (11) (continuous line) and analytically approximated with (13) (dashed line)) are plotted in bottom panel of Figure 1. It may be appreciated that, even with this rather small sample size, the approximation is quite good.

Suppose that the value $\lambda_0 = 3$ is to be tested. The corresponding intrinsic statistic is $d(3 \,|\, \boldsymbol{z}) = 2.72 \approx \log(15)$; thus the average likelihood ratio against the $H_0 \equiv \{\lambda = 3\}$ may be expected to be about 15, not really strong evidence against this value, even if this may be seen to be in the right tail of the reference posterior of λ_0. On the other hand, if the value to test is $\lambda_0 = 3.5$, the corresponding intrinsic statistic is $d(3.5 \,|\, \boldsymbol{z}) = 6.36 \approx \log(576)$; hence the average likelihood ratio against this value may be expected to be about 576 and the hypothesis $H_0 \equiv \{\lambda = 3.5\}$ should be rejected in most scenarios. Notice that, in marked difference to conventional testing using Bayes factors, the same prior has been used to test these two possible parameter values (as it would obviously be for any other value).

The following example illustrates the use of the methods described to derive a new solution to a classical precise testing problem.

Example 3 (Equality of Normal means). Let $\boldsymbol{z} = \{\boldsymbol{x}, \boldsymbol{y}\}$ be two independent random samples, $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from $\mathrm{N}(x \,|\, \mu_x, \sigma\}$, and $\boldsymbol{y} = \{y_1, \ldots, y_m\}$ from $\mathrm{N}(x \,|\, \mu_y, \sigma\}$, and suppose that one in 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 $\boldsymbol{z}$. Using the additive property of the intrinsic discrepancy loss and the fact that the KL divergence between two normals distrubutions with the same variance is simply $\kappa\{\mu_j, \sigma \,|\, \mu_i, \sigma\} = (\mu_i - \mu_j)^2/(2/\sigma^2)$ to derive the logarithmic divergence of $p(\boldsymbol{z} \,|\, \mu_0, \mu_0, \sigma_0)$ from $p(\boldsymbol{z} \,|\, \mu_x, \mu_y, \sigma)$, and then minimizing over both μ_0 and σ_0 yields $\inf_{\mu_0 \in \Re,\, \sigma_0 > 0} \kappa\{\mu_0, \mu_0, \sigma_0 \,|\, \mu_x, \mu_y, \sigma\} = k_{nm}\ \theta^2$, 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, $\inf_{\mu_0 \in \Re,\, \sigma_0 > 0}\ \kappa\{\mu_x, \mu_y, \sigma \,|\, \mu_0, \mu_0, \sigma_0\}$ yields $[(m+n)/2]\, \log[1 + (k_{nm}/(2(m+n)))\, \theta^2]$, which is always smaller. Hence, the intrinsic discrepancy loss of accepting H_0 is

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

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 $t = t(\boldsymbol{z}) = (\bar{x} - \bar{y})/(s/\sqrt{2/k_{nm}})$, where s is the m.l.e. of σ. Therefore, the required marginal reference posterior of θ is $\pi(\theta \,|\, \boldsymbol{z}) = \pi(\theta \,|\, t, m, n) \propto p(t \,|\, \theta)\, \pi(\theta)$ 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 relevant expected 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

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

The sampling distribution of $d(H_0 \,|\, t, n, m)$ is asymptotically $(1/2)[1 + \chi^2_1(\lambda)]$, where $\chi^2_1(\lambda)$ is a non-central chi-squared distribution with one degree of freedom and non-centrality parameter $\lambda = k_{nm}\theta^2/2$. If 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.

For many more sophisticated examples of precise hypothesis testing, see Bernardo (2011) and references therein.

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 in 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. (2011). Integrated objective Bayesian estimation and hypothesis testing. *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, 1–68, (with discussion).

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

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

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.

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).