International Statistical Review (2002), **70,** 3, 351–372, Printed in The Netherlands

Bayesian Hypothesis Testing: a Reference Approach

José M. Bernardo[1] and Raúl Rueda[2]

[1]*Dep. d'Estadística e IO, Universitat de València, 46100-Burjassot, Valencia, Spain. E-mail: jose.m.bernardo@uv.es* [2]*IIMAS, UNAM, Apartado Postal 20-726, 01000 Mexico DF, Mexico. E-mail: pinky@sigma.iimas.unam.mx*

Summary

For any probability model $M \equiv \{p(x \mid \theta, \omega), \theta \in \Theta, \omega \in \Omega\}$ assumed to describe the probabilistic behaviour of data $x \in X$, it is argued that testing whether or not the available data are *compatible* with the hypothesis $H_0 \equiv \{\theta = \theta_0\}$ is best considered as a formal decision problem on whether to use (a_0), or not to use (a_1), the simpler probability model (or *null* model) $M_0 \equiv \{p(x \mid \theta_0, \omega), \omega \in \Omega\}$, where the loss difference $L(a_0, \theta, \omega) - L(a_1, \theta, \omega)$ is proportional to the amount of information $\delta(\theta_0, \theta, \omega)$ which would be lost if the simplified model M_0 were used as a proxy for the assumed model M. For any prior distribution $\pi(\theta, \omega)$, the appropriate normative solution is obtained by rejecting the null model M_0 whenever the corresponding posterior expectation $\int \int \delta(\theta_0, \theta, \omega)\, \pi(\theta, \omega \mid x) d\theta\, d\omega$ is sufficiently large.

Specification of a subjective prior is always difficult, and often polemical, in scientific communication. Information theory may be used to specify a prior, the *reference* prior, which only depends on the assumed model M, and mathematically describes a situation where no prior information is available about the quantity of interest. The reference posterior expectation, $d(\theta_0, x) = \int \delta\, \pi(\delta \mid x) d\delta$, of the amount of information $\delta(\theta_0, \theta, \omega)$ which could be lost if the null model were used, provides an attractive non-negative test function, the *intrinsic statistic*, which is invariant under reparametrization.

The intrinsic statistic $d(\theta_0, x)$ is measured in units of information, and it is easily calibrated (for any sample size and any dimensionality) in terms of some average log-likelihood ratios. The corresponding Bayes decision rule, the *Bayesian reference criterion (BRC)*, indicates that the null model M_0 should only be rejected if the posterior expected loss of information from using the simplified model M_0 is too large or, equivalently, if the associated expected average log-likelihood ratio is large enough.

The BRC criterion provides a general reference Bayesian solution to hypothesis testing which does not assume a probability mass concentrated on M_0 and, hence, it is immune to Lindley's paradox. The theory is illustrated within the context of multivariate normal data, where it is shown to avoid Rao's paradox on the inconsistency between univariate and multivariate frequentist hypothesis testing.

Key words: Amount of Information; Decision Theory; Lindley's Paradox; Loss function; Model Criticism; Model Choice; Precise Hypothesis Testing; Rao's Paradox; Reference Analysis; Reference Prior.

1 Introduction

1.1 Model Choice and Hypothesis Testing

Hypothesis testing has been subject to polemic since its early formulation by Neyman and Pearson in the 1930s. This is mainly due to the fact that its standard formulation often constitutes a serious oversimplification of the problem intended to solve. Indeed, many of the problems which traditionally

have been formulated in terms of hypothesis testing are really complex decision problems on *model choice*, whose appropriate solution naturally depends on the structure of the problem. Some of these important structural elements are the motivation to choose a particular model (*e.g.*, simplification or prediction), the class of models considered (say a finite set of alternatives or a class of nested models), and the available prior information (say a sharp prior concentrated on a particular model or a relatively diffuse prior).

In the vast literature of model choice, reference is often made to the "true" probability model. Assuming the existence of a "true" model would be appropriate whenever one knew for sure that the real world mechanism which has generated the available data was one of a specified class. This would indeed be the case if data had been generated by computer simulation, but beyond such controlled situations it is difficult to accept the existence of a "true" model in a literal sense. There are many situations however where one is prepared to proceed "as if" such a true model existed, and furthermore belonged to some specified class of models. Naturally, any further conclusions will then be conditional on this (often strong) assumption being reasonable in the situation considered.

The natural mathematical framework for a systematic treatment of model choice is decision theory. One has to specify the range of models which one is willing to consider, to decide whether or not it may be assumed that this range includes the true model, to specify probability distributions describing prior information on all unknown elements in the problem, and to specify a loss function measuring the eventual consequences of each model choice. The best alternative within the range of models considered is then that model which minimizes the corresponding expected posterior loss. Bernardo & Smith (1994, Ch. 6) provide a detailed description of many of these options. In this paper attention focuses on one of the simplest problems of model choice, namely *hypothesis testing*, where a (typically large) model M is tentatively accepted, and it is desired to test whether or not available data are *compatible* with a particular submodel M_0. Note that this formulation includes most of the problems traditionally considered under the heading of hypothesis testing in the frequentist statistical literature.

1.2 Notation

It is assumed that probability distributions may be described through their probability mass or probability density functions, and no distinction is generally made between a random quantity and the particular values that it may take. Roman fonts are used for *observable* random quantities (typically data) and for known constants, while Greek fonts are used for *unobservable* random quantities (typically parameters). Bold face is used to denote row vectors, and $\boldsymbol{x}'$ to denote the transpose of the vector $\boldsymbol{x}$. Lower case is used for variables and upper case for their domains. The standard mathematical convention of referring to *functions*, say f and g of $\boldsymbol{x} \in X$, respectively, by $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$, will often be used. In particular, $p(\boldsymbol{x} \mid C)$ and $p(\boldsymbol{y} \mid C)$ will respectively represent general *probability densities* of the *observable* random vectors $\boldsymbol{x} \in X$ and $\boldsymbol{y} \in Y$ under conditions C, without any suggestion that the random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ have the same distribution. Similarly, $\pi(\boldsymbol{\theta} \mid C)$ and $\pi(\boldsymbol{\omega} \mid C)$ will respectively represent general probability densities of the *unobservable* parameter vectors $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\omega} \in \Omega$ under conditions C. Thus, $p(\boldsymbol{x} \mid C) \geq 0$, $\int_X p(\boldsymbol{x} \mid C) d\boldsymbol{x} = 1$, and $\pi(\boldsymbol{\theta} \mid C) \geq 0$, $\int_\Theta \pi(\boldsymbol{\theta} \mid C) d\boldsymbol{\theta} = 1$. If the random vectors are discrete, these functions are probability mass functions, and integrals over their values become sums. $\mathrm{E}[\boldsymbol{x} \mid C]$ and $\mathrm{E}[\boldsymbol{\theta} \mid C]$ are respectively used to denote the expected values of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ under conditions C. Finally, $\Pr(\boldsymbol{\theta} \in A \mid \boldsymbol{x}, C) = \int_A p(\boldsymbol{\theta} \mid \boldsymbol{x}, C) d\boldsymbol{\theta}$ denotes the probability that the parameter $\boldsymbol{\theta}$ belongs to A, given data $\boldsymbol{x}$ and conditions C.

Specific density functions are denoted by appropriate names. Thus, if x is a univariate random quantity having a Normal distribution with mean μ and variance σ^2, its probability density function will be denoted $\mathrm{N}(x \mid \mu, \sigma^2)$; if θ has a Beta distribution with parameters a and b, its density function

will be denoted $\text{Be}(\theta \mid a, b)$.

A *probability model* for some data $\boldsymbol{x} \in X$ is defined as a *family* of probability distributions for $\boldsymbol{x}$ indexed by some *parameter*. Whenever a model has to be fully specified, the notation $\{p(\boldsymbol{x} \mid \boldsymbol{\phi}),\ \boldsymbol{\phi} \in \Phi,\ \boldsymbol{x} \in X\}$ is used, and it is assumed that $p(\boldsymbol{x} \mid \boldsymbol{\phi})$ is a probability density function (or a probability mass function) so that $p(\boldsymbol{x} \mid \boldsymbol{\phi}) \geq 0$, and $\int_X p(\boldsymbol{x} \mid \boldsymbol{\phi}) d\boldsymbol{x} = 1$ for all $\boldsymbol{\phi} \in \Phi$. The parameter $\boldsymbol{\phi}$ will generally be assumed to be a vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)$ of finite dimension $k \geq 1$, so that $\Phi \subset \Re^k$. Often, the parameter vector $\boldsymbol{\phi}$ will be written in the form $\boldsymbol{\phi} = \{\boldsymbol{\theta}, \boldsymbol{\omega}\}$, where $\boldsymbol{\theta}$ is considered to be the vector of interest and $\boldsymbol{\omega}$ a vector of nuisance parameters. The sets X and Φ will be referred to, respectively, as the *sample space* and the *parameter space*. Occasionally, if there is no danger of confusion, reference is made to 'model' $\{p(\boldsymbol{x} \mid \boldsymbol{\phi}), \boldsymbol{\phi} \in \Phi\}$, or even to 'model' $p(\boldsymbol{x} \mid \boldsymbol{\phi})$, without recalling the sample and the parameter spaces. In non-regular problems the sample space X depends on the parameter value $\boldsymbol{\phi}$; this will explicitly be indicated by writing $X = X(\boldsymbol{\phi})$. Considered as a function of the parameter $\boldsymbol{\phi}$, the probability density (or probability mass) $p(\boldsymbol{x} \mid \boldsymbol{\phi})$ will be referred to as the *likelihood function* of $\boldsymbol{\phi}$ given $\boldsymbol{x}$. Whenever this exists, a maximum of the likelihood function (*maximum likelihood estimate* or *mle*) will be denoted by $\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\phi}}(\boldsymbol{x})$.

The *complete* set of available data is represented by $\boldsymbol{x}$. In many examples this will be a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from a model of the form $\{p(x \mid \boldsymbol{\phi}), x \in \Re, \boldsymbol{\phi} \in \Phi\}$ so that the likelihood function will be of the form $p(\boldsymbol{x} \mid \boldsymbol{\phi}) = \prod_{j=1}^n p(x_j \mid \boldsymbol{\phi})$ and the sample space will be $X \subset \Re^n$, but it will *not* be assumed that this has to be the case. The notation $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$, $\boldsymbol{t} \in T$, is used to refer to a general function of the data; often, but not necessarily, this will be a sufficient statistic.

1.3 Simple Model Choice

The simplest example of a model choice problem (and one which centers most discussions on model choice and model comparison) is one where (i) the range of models considered is a finite class $\mathcal{M} = \{M_1, \ldots, M_m\}$, of m fully specified models

$$M_i \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\phi}_i),\ \boldsymbol{x} \in X\}, \quad i = 1, \ldots, m \tag{1}$$

(ii) it is assumed that that the 'true' model is a member $M_t \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\phi}_t), \boldsymbol{x} \in X\}$ from that class, and (iii) the loss function is the simple step function

$$\begin{cases} \ell(a_t, \boldsymbol{\phi}_t) & = 0, \\ \ell(a_i, \boldsymbol{\phi}_t) & = c > 0, \quad i \neq t, \end{cases} \tag{2}$$

where a_i denotes the *decision to act as if* the true model was M_i. In this simplistic situation, it is immediate to verify that the optimal model choice is that which maximizes the posterior probability, $\pi(\boldsymbol{\phi}_i \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\phi}_i)\pi(\boldsymbol{\phi}_i)$. Moreover, an intuitive measure of paired comparison of plausibility between any two of the models M_i and M_j is provided by the ratio of the posterior probabilities $\pi(\boldsymbol{\phi}_i \mid \boldsymbol{x})/\pi(\boldsymbol{\phi}_j \mid \boldsymbol{x})$. If, in particular, all m models are judged to be equally likely a priori, so that $\pi(\boldsymbol{\phi}_i) = 1/m$, for all i, then the optimal model is that which maximizes the likelihood, $p(\boldsymbol{x} \mid \boldsymbol{\phi}_i)$, and the ratio of posterior probabilities reduces to the corresponding *Bayes factor* $B_{ij} = p(\boldsymbol{x} \mid \boldsymbol{\phi}_i)/p(\boldsymbol{x} \mid \boldsymbol{\phi}_j)$ which, in this simple case (with no nuisance parameters), it is also the corresponding likelihood ratio.

The natural extension of this scenario to a continuous setting considers a non-countable class of models $\mathcal{M} = \{M_{\boldsymbol{\phi}}, \boldsymbol{\phi} \in \Phi \subset \Re^k\}$,

$$M_{\boldsymbol{\phi}} \equiv p(\boldsymbol{x} \mid \boldsymbol{\phi}), \quad \text{with} \quad p(\boldsymbol{x} \mid \boldsymbol{\phi}) > 0, \quad \int_X p(\boldsymbol{x} \mid \boldsymbol{\phi})\, d\boldsymbol{x} = 1, \tag{3}$$

an absolutely continuous and strictly positive prior, represented by its density $p(\phi) > 0$, and a simple step loss function $\ell(a_\phi, \phi)$ such that

$$\begin{cases} \ell(a_\phi, \phi_t) &= 0, & \phi \in B_\epsilon(\phi_t) \\ \ell(a_\phi, \phi_t) &= c > 0, & \phi \notin B_\epsilon(\phi_t), \end{cases} \tag{4}$$

where a_ϕ denotes the decision to act as if the true model was M_ϕ, and $B_\epsilon(\phi_t)$ is a radius ϵ neighbourhood of ϕ_t. In this case, it is easily shown that, as ϵ decreases, the optimal model choice converges to the model labelled by the mode of the corresponding posterior distribution $\pi(\phi \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \phi)\, \pi(\phi)$. Note that with this formulation, which strictly parallels the conventional formulation for model choice in the finite case, the problem of model choice is mathematically equivalent to the problem of point estimation with a zero-one loss function.

1.4 Hypothesis Testing

Within the context of an accepted, possibly very wide class of models, $\mathcal{M} = \{M_\phi, \phi \in \Phi\}$, a subset $\mathcal{M}_0 = \{M_\phi,\ \phi \in \Phi_0 \subset \Phi\}$ of the class $\mathcal{M}$, where Φ_0 may possibly consist of a single value ϕ_0, is sometimes suggested in the course of the investigation as deserving special attention. This may either be because restricting ϕ to Φ_0 would greatly simplify the model, or because there are additional (context specific) arguments suggesting that $\phi \in \Phi_0$. The conventional formulation of a *hypothesis testing* problem is stated within this framework. Thus, given data $\boldsymbol{x} \in X$ which are *assumed* to have been generated by $p(\boldsymbol{x} \mid \phi)$, for some $\phi \in \Phi$, a procedure is required to advise on whether or not it may safely be assumed that $\phi \in \Phi_0$. In conventional language, a procedure is desired to *test* the *null* hypothesis $H_0 \equiv \{\phi \in \Phi_0\}$. The particular case where Φ_0 contains a *single* value ϕ_0, so that $\Phi_0 = \{\phi_0\}$, is further referred to as a problem of *precise* hypothesis testing.

The standard frequentist approach to *precise* hypothesis testing requires to propose some one-dimensional test statistic $t = t(\boldsymbol{x}) \in T \subset \Re$, where large values of t cast doubt on H_0. The p-value (or observed significance level) associated to some observed data $\boldsymbol{x}_0 \in X$ is then the probability, conditional on the null hypothesis being true, of observing data as or more extreme than the data actually observed, that is,

$$p = \Pr[t \geq t(\boldsymbol{x}_0) \mid \phi = \phi_0] = \int_{\{\boldsymbol{x};\ t(\boldsymbol{x}) \geq t(\boldsymbol{x}_0)\}} p(\boldsymbol{x} \mid \phi_0)\, d\boldsymbol{x}. \tag{5}$$

Small values of the p-value are considered to be evidence against H_0, with the values 0.05 and 0.01 typically used as conventional cut-off points.

There are many well-known criticisms to this common procedure, some of which are briefly reviewed below. For further discussion see Jeffreys (1961), Edwards, Lindman & Savage (1963), Rao (1966), Lindley (1972), Good (1983), Berger & Delampady (1987), Berger & Sellke (1987), Matthews (2001), and references therein.

- *Arbitrary choice of the test statistic*. There is no generally accepted theory on the selection of the appropriate test statistic, and different choices may well lead to incompatible results.
- *Not a measure of evidence*. Observed significance levels are not direct measures of evidence. Although most users would like it to be true, in precise hypothesis testing there is no mathematical relation between the p-value and $\Pr[H_0 \mid \boldsymbol{x}_0]$, the probability that the null is true given the evidence.
- *Arbitrary cut-off points*. Conventional cut-off points for p-values (as the ubiquitous 0.05) are arbitrary, and ignore power. Moreover, despite frequent warnings in the literature, they are

typically chosen with no regard for either the dimensionality of the problem or the sample size (possibly due to the fact that there is no accepted methodology to perform that adjustment).

- *Exaggerate significance*. Different arguments have been used to suggest that the conventional use of p-values exaggerate significance. Indeed, with common sample sizes, a 0.05 p-value is typically better seen as an indication that more data are needed than as firm evidence against the null.
- *Improper conditioning*. Observed significance levels are not based on the *observed* evidence, namely $t(\boldsymbol{x}) = t(\boldsymbol{x}_0)$, but on the (less than obviously relevant) event $\{t(\boldsymbol{x}) \geq t(\boldsymbol{x}_0)\}$ so that, to quote Jeffreys (1980, p. 453), the null hypothesis may be rejected by not predicting something that has not happened.
- *Contradictions*. Using fixed cut-off points for p-values easily leads to contradiction. For instance, in a multivariate setting, one may simultaneously reject all components $\phi_i = \phi_{i0}$ and yet accept $\boldsymbol{\phi} = \boldsymbol{\phi}_0$ (Rao's paradox).
- *No general procedure*. The procedure is not directly applicable to general hypothesis testing problems. Indeed, the p-value is a function of the sampling distribution of the test statistic under the null, and this is only well defined in the case of *precise* hypothesis testing. Extensions to the general case, $\mathcal{M}_0 = \{M_\phi,\ \boldsymbol{\phi} \in \Phi_0\}$, where Φ_0 contains more than one point, are less than obvious.

Hypothesis testing has been formulated as a decision problem. No wonder therefore that Bayesian approaches to hypothesis testing are best described within the unifying framework of decision theory. Those are reviewed below.

2 Hypothesis Testing as a Decision Problem

2.1 General Structure

Consider the probability model $M \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega}), \boldsymbol{\theta} \in \Theta, \boldsymbol{\omega} \in \Omega\}$ which is currently assumed to provide an appropriate description of the probabilistic behaviour of observable data $\boldsymbol{x} \in X$ in terms of some *vector of interest* $\boldsymbol{\theta} \in \Theta$ and some *nuisance parameter vector* $\boldsymbol{\omega} \in \Omega$. From a Bayesian viewpoint, the complete final outcome of a problem of inference about any unknown quantity is the appropriate posterior distribution. Thus, given data $\boldsymbol{x}$ and a (joint) prior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\omega})$, all that can be said about $\boldsymbol{\theta}$ is encapsulated in the corresponding posterior distribution

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) = \int_\Omega \pi(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{x})\, d\boldsymbol{\omega}, \qquad \pi(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})\, \pi(\boldsymbol{\theta}, \boldsymbol{\omega}). \tag{6}$$

In particular, the (marginal) posterior distribution of $\boldsymbol{\theta}$ immediately conveys information on those values of the vector of interest which (given the assumed model) may be taken to be *compatible* with the observed data $\boldsymbol{x}$, namely, those with a relatively high probability density. In some occasions, a particular value $\boldsymbol{\theta} = \boldsymbol{\theta}_0 \in \Theta$ of the quantity of interest is suggested in the course of the investigation as deserving special consideration, either because assuming $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ would greatly simplify the model, or because there are additional (context specific) arguments suggesting that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Intuitively, the *(null) hypothesis* $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ should be judged to be *compatible* with the observed data $\boldsymbol{x}$ if $\boldsymbol{\theta}_0$ has a relatively high posterior density; however, a more precise conclusion is often required, and this may be derived from a decision-oriented approach.

Formally, testing the hypothesis $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ is defined as a *decision problem* where the action space has only two elements, namely to accept (a_0) or to reject (a_1) the use of the restricted model $M_0 \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}),\ \boldsymbol{\omega} \in \Omega\}$ as a convenient proxy for the *assumed* model $M \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega}),\ \boldsymbol{\theta} \in \Theta,\ \boldsymbol{\omega} \in \Omega\}$. To solve this decision problem, it is necessary to specify an appropriate loss function, $\{\ell[a_i, (\boldsymbol{\theta}, \boldsymbol{\omega})],\ i = 0, 1\}$, measuring the consequences of accepting or rejecting H_0 as a function of

the actual values $(\boldsymbol{\theta}, \boldsymbol{\omega})$ of the parameters. Notice that this requires the statement of an *alternative* action a_1 to accepting H_0; this is only to be expected, for an action is taken not because it is good, but because it is better than anything else that has been imagined.

Given data $\boldsymbol{x}$, the optimal action will be to reject H_0 if (and only if) the expected posterior loss of accepting, $\int_\Theta \int_\Omega \ell[a_0, (\boldsymbol{\theta}, \boldsymbol{\omega})]\, \pi(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{x})\, d\boldsymbol{\theta} d\boldsymbol{\omega}$, is larger than the expected posterior loss of rejecting, $\int_\Theta \int_\Omega \ell[a_1, (\boldsymbol{\theta}, \boldsymbol{\omega})]\, \pi(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{x})\, d\boldsymbol{\theta} d\boldsymbol{\omega}$, *i.e.,* iff

$$\int_\Theta \int_\Omega \{\ell[a_0, (\boldsymbol{\theta}, \boldsymbol{\omega})] - \ell[a_1, (\boldsymbol{\theta}, \boldsymbol{\omega})]\}\, \pi(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{x})\, d\boldsymbol{\theta} d\boldsymbol{\omega} > 0. \tag{7}$$

Therefore, *only* the *loss difference*

$$\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega}) = \ell[a_0, (\boldsymbol{\theta}, \boldsymbol{\omega})] - \ell[a_1, (\boldsymbol{\theta}, \boldsymbol{\omega})], \tag{8}$$

which measures the *advantage* of rejecting H_0 as a function of $\{\boldsymbol{\theta}, \boldsymbol{\omega}\}$, has to be specified. Notice that no constraint has been imposed in the preceding formulation. It follows that *any* (generalized) Bayes solution to the decision problem posed (and hence any *admissible* solution, see *e.g.,* Berger, 1985, Ch. 8) *must* be of the form

$$\text{Reject } H_0 \quad \text{iff} \quad \int_\Theta \int_\Omega \Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega})\, \pi(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{x})\, d\boldsymbol{\theta} d\boldsymbol{\omega} > 0, \tag{9}$$

for some loss difference function $\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega})$, and some (possibly improper) prior $\pi(\boldsymbol{\theta}, \boldsymbol{\omega})$. Thus, as common sense dictates, the hypothesis H_0 should be rejected whenever the expected advantage of rejecting H_0 is positive. In some examples, the loss difference function does not depend on the nuisance parameter vector $\boldsymbol{\omega}$; if this is the case, the decision criterion obviously simplifies to rejecting H_0 iff $\int_\Theta \Delta\ell(H_0, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\, d\boldsymbol{\theta} > 0$.

A crucial element in the specification of the loss function is a description of what is precisely meant by rejecting H_0. By assumption, a_0 means to act *as if* model M_0 were true, *i.e.,* as if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, but there are at least two options for the alternative action a_1. This might mean the *negation* of H_0, that is to act as if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, or it might rather mean to reject the simplification to M_0 implied by $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and to keep the unrestricted model M (with $\boldsymbol{\theta} \in \Theta$), which is acceptable by assumption. Both of these options have been analyzed in the literature, although it may be argued that the problems of scientific data analysis where precise hypothesis testing procedures are typically used are better described by the second alternative. Indeed, this is the situation in two frequent scenarios: (i) an established model, identified by M_0, is *embedded* into a more general model M (so that $M_0 \subset M$), constructed to include possibly promising departures from M_0, and it is required to verify whether or not the extended model M provides a significant improvement in the description of the behaviour of the available data; or, (ii) a large model M is accepted, and it is required to verify whether or not the simpler model M_0 may be used as a sufficiently accurate approximation.

2.2 Bayes Factors

The *Bayes factor* approach to hypothesis testing is a particular case of the decision structure outlined above; it is obtained when the alternative action a_1 is taken to be to act as if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, and the difference loss function is taken to be a simplistic zero-one function. Indeed, if the *advantage* $\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ of rejecting H_0 is of the form

$$\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega}) = \Delta\ell(H_0, \boldsymbol{\theta}) = \begin{cases} -1 & \text{if } \boldsymbol{\theta} = \boldsymbol{\theta}_0 \\ +1 & \text{if } \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \end{cases} \tag{10}$$

then the corresponding decision criterion is

$$\text{Reject } H_0 \quad \text{iff} \quad \Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0 \mid \boldsymbol{x}) < \Pr(\boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \mid \boldsymbol{x}). \tag{11}$$

If the prior distribution is such that $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0) = \Pr(\boldsymbol{\theta} \neq \boldsymbol{\theta}_0) = 1/2$, and $\{\pi(\boldsymbol{\omega} \mid \boldsymbol{\theta}_0), \pi(\boldsymbol{\omega} \mid \boldsymbol{\theta})\}$ respectively denote the conditional prior distributions of $\boldsymbol{\omega}$, when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and when $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, then the criterion becomes

$$\text{Reject } H_0 \quad \text{iff} \quad B_{01}\{\boldsymbol{x}, \pi(\boldsymbol{\omega} \mid \boldsymbol{\theta}_0), \pi(\boldsymbol{\omega} \mid \boldsymbol{\theta})\} = \frac{\int_\Omega p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega})\, \pi(\boldsymbol{\omega} \mid \boldsymbol{\theta}_0)\, d\boldsymbol{\omega}}{\int_\Theta \int_\Omega p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})\, \pi(\boldsymbol{\omega} \mid \boldsymbol{\theta})\, d\boldsymbol{\theta} d\boldsymbol{\omega}} < 1 \tag{12}$$

where $B_{01}\{\boldsymbol{x}, \pi(\boldsymbol{\omega} \mid \boldsymbol{\theta}_0), \pi(\boldsymbol{\omega} \mid \boldsymbol{\theta})\}$ is the *Bayes factor* (or integrated likelihood ratio) in favour of H_0. Notice that the Bayes factor B_{01} crucially depends on the conditional priors $\pi(\boldsymbol{\omega} \mid \boldsymbol{\theta}_0)$ and $\pi(\boldsymbol{\omega} \mid \boldsymbol{\theta})$, which must typically be proper for the Bayes factor to be well-defined.

It is important to realize that this formulation *requires* that $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0) > 0$, so that the hypothesis H_0 must have a strictly positive prior probability. If $\boldsymbol{\theta}$ is a continuous parameter, this *forces* the use of a *non-regular* (not absolutely continuous) 'sharp' prior concentrating a positive probability mass on $\boldsymbol{\theta}_0$. One unappealing consequence of this non-regular prior structure, noted by Lindley (1957) and generally known as *Lindley's paradox*, is that for any *fixed* value of the pertinent test statistic, the Bayes factor typically increases as $\sqrt{n}$ with the sample size; hence, with large samples, "evidence" in favor of H_0 *may* be overwhelming with data sets which are both extremely implausible under H_0 and quite likely under alternative $\boldsymbol{\theta}$ values, such as (say) the mle $\hat{\boldsymbol{\theta}}$. For further discussion of this polemical issue see Bernardo (1980), Shafer (1982), Berger & Delampady (1987), Casella & Berger (1987), Robert (1993), Bernardo (1999), and discussions therein.

The Bayes factor approach to hypothesis testing in a continuous parameter setting deals with situations of *concentrated* prior probability; it *assumes* important prior knowledge about the value of the vector of interest $\boldsymbol{\theta}$ (described by a prior sharply spiked on $\boldsymbol{\theta}_0$) and analyzes how such *very strong* prior beliefs about the value of $\boldsymbol{\theta}$ should be modified by the data. Hence, Bayes factors should *not* be used unless this strong prior formulation is an appropriate assumption. In particular, Bayes factors should *not* be used to test the *compatibility* of the data with H_0, for they inextricably combine what data have to say with (typically subjective) *strong* beliefs about the value of $\boldsymbol{\theta}$.

2.3 Continuous Loss Functions

It is often natural to assume that the loss difference $\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega})$, a conditional measure of the loss suffered if $p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega})$ were used as a proxy for $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$, has to be some *continuous* function of the 'discrepancy' between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. Moreover, one would expect $\Delta\ell(H_0, \boldsymbol{\theta}_0, \boldsymbol{\omega})$ to be negative, for there must be some positive advantage, say $\ell^* > 0$, in accepting the null when it is true. A simple example is the quadratic loss

$$\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega}) = \Delta\ell(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^2 - \ell^*, \quad \ell^* > 0. \tag{13}$$

Notice that continuous difference loss functions do not require the use of non-regular priors. As a consequence, their use does not *force* the assumption of strong prior beliefs and, in particular, they may be used with improper priors. However, (i) there are many possible choices for continuous difference loss functions; (ii) the resulting criteria are typically not invariant under one-to-one reparametrization of the quantity of interest; and (iii) their use requires some form of calibration, that is, an appropriate choice of the utility constant ℓ^*, which is often context dependent.

In the next section we justify the choice of a particular continuous invariant difference loss function, the *intrinsic discrepancy*. This is combined with reference analysis to propose an attractive Bayesian solution to the problem of hypothesis testing, defined as the problem of deciding whether

or not available data are statistically compatible with the hypothesis that the parameters of the model belong to some subset of the parameter space. The proposed solution sharpens a procedure suggested by Bernardo (1999) to make it applicable to non-regular models, and extends previous results to multivariate probability models. For earlier, related references, see Bernardo (1982, 1985), Bernardo & Bayarri (1985), Ferrándiz (1985), Gutiérrez-Peña (1992), and Rueda (1992). The argument lies entirely within a Bayesian decision-theoretical framework (in that the proposed solution is obtained by minimizing a posterior expected loss), and it is *objective* (in the precise sense that it only uses an "objective" prior, a prior uniquely defined in terms of the assumed model and the quantity of interest).

3 The Bayesian Reference Criterion

Let model $M \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega}), \boldsymbol{\theta} \in \Theta, \boldsymbol{\omega} \in \Omega\}$ be a currently accepted description of the probabilistic behaviour of data $\boldsymbol{x} \in X$, let a_0 be the decision to work under the restricted model $M_0 \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, and let a_1 be the decision to keep the general, unrestricted model M. In this situation, the loss advantage $\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ of rejecting H_0 as a function of $(\boldsymbol{\theta}, \boldsymbol{\omega})$ may safely be assumed to have the form

$$\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega}) = \delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega}) - d^*, \qquad d^* > 0, \tag{14}$$

where

(i) the function $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ is some non-negative measure of the *discrepancy* between the assumed model $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ and its closest approximation within $\{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, such that $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0, \boldsymbol{\omega}) = 0$, and

(ii) the constant $d^* > 0$ is a context dependent *utility value* which measures the (necessarily positive) advantage of being able to work with the simpler model when it is true.

Choices of both $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ and d^* which might be appropriate for general use will now be discussed.

3.1 The Intrinsic Discrepancy

Conventional loss functions typically focus on the "distance" between the true and the null values of the quantity of interest, rather than on the "distance" between the models they label and, typically, they are *not* invariant under reparametrization. Intrinsic losses however (see *e.g.*, Robert, 1996) directly focus on how different the true model is from the null model, and they typically produce invariant solutions. We now introduce a new, particularly attractive, intrinsic loss function, the *intrinsic discrepancy loss*.

The basic idea is to define the discrepancy between two probability densities $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ as $\min\{k(p_1 \mid p_2), k(p_2 \mid p_1)\}$, where

$$k(p_2 \mid p_1) = \int_X p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} d\boldsymbol{x} \tag{15}$$

is the *directed logarithmic divergence* (Kullback & Leibler, 1951; Kullback, 1959) of $p_2(\boldsymbol{x})$ from $p_1(\boldsymbol{x})$. The discrepancy from a point to a set is further defined as the discrepancy from the point to its closest element in the set. The introduction of the minimum makes it possible to define a symmetric discrepancy between probability densities which is *finite* with strictly nested supports, a crucial property if a general theory (applicable to non-regular models) is required.

Definition 1. Intrinsic Discrepancies. The intrinsic discrepancy $\delta(p_1, p_2)$ between two probability densities $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ for the random quantity $\boldsymbol{x} \in X$ is

$$\delta\{p_1(\boldsymbol{x}), p_2(\boldsymbol{x})\} = \min \left\{ \int_X p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} d\boldsymbol{x}, \int_X p_2(\boldsymbol{x}) \log \frac{p_2(\boldsymbol{x})}{p_1(\boldsymbol{x})} d\boldsymbol{x} \right\}.$$

The intrinsic discrepancy between two families of probability densities for the random quantity $\boldsymbol{x} \in X$, $M_1 \equiv \{p_1(\boldsymbol{x} \mid \boldsymbol{\phi}), \boldsymbol{\phi} \in \Phi\}$ and $M_2 \equiv \{p_2(\boldsymbol{x} \mid \boldsymbol{\psi}), \boldsymbol{\psi} \in \Psi\}$, is given by

$$\delta(M_1, M_2) = \min_{\boldsymbol{\phi} \in \Phi, \boldsymbol{\psi} \in \Psi} \delta\{p_1(\boldsymbol{x} \mid \boldsymbol{\phi}), p_2(\boldsymbol{x} \mid \boldsymbol{\psi})\}. \quad \triangleleft$$

It immediately follows for Definition 1 that $\delta\{p_1(\boldsymbol{x}), p_2(\boldsymbol{x})\}$ provides the minimum expected log-density ratio $\log[p_i(\boldsymbol{x})/p_j(\boldsymbol{x})]$ in favour of the true density that one would obtain if data $\boldsymbol{x} \in X$ were sampled from either $p_1(\boldsymbol{x})$ or $p_2(\boldsymbol{x})$. In particular, if $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ are fully specified alternative probability models for data $\boldsymbol{x} \in X$, and it is assumed that one of them is true, then $\delta\{p_1(\boldsymbol{x}), p_2(\boldsymbol{x})\}$ is the minimum expected log-likelihood ratio for the true model.

Intrinsic discrepancies have a number of attractive properties. Some are directly inherited from the directed logarithmic divergence. Indeed,

(i) The intrinsic discrepancy $\delta\{p_1(\boldsymbol{x}), p_2(\boldsymbol{x})\}$ between $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ is *non-negative* and vanishes iff $p_1(\boldsymbol{x}) = p_2(\boldsymbol{x})$ almost everywhere.
(ii) The intrinsic discrepancy $\delta\{p_1(\boldsymbol{x}), p_2(\boldsymbol{x})\}$ is invariant under one-to-one transformations $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$ of the random quantity $\boldsymbol{x}$.
(iii) The intrinsic discrepancy is *additive* in the sense that if the available data $\boldsymbol{x}$ consist of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from either $p_1(x)$ or $p_2(x)$, then $\delta\{p_1(\boldsymbol{x}), p_2(\boldsymbol{x})\} = n\,\delta\{p_1(x), p_2(x)\}$.
(iv) If the densities $p_1(\boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{\phi}_1)$ and $p_2(\boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{\phi}_2)$ are two members of a parametric family $p(\boldsymbol{x} \mid \boldsymbol{\phi})$, then $\delta\{p(\boldsymbol{x} \mid \boldsymbol{\phi}_1), p(\boldsymbol{x} \mid \boldsymbol{\phi}_2)\} = \delta\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2\}$ is *invariant* under one-to-one transformations for the parameter, so that for any such transformation $\boldsymbol{\psi}_i = \boldsymbol{\psi}(\boldsymbol{\phi}_i)$, one has $\delta\{p(\boldsymbol{x} \mid \boldsymbol{\psi}_1), p(\boldsymbol{x} \mid \boldsymbol{\psi}_2)\} = \delta\{\boldsymbol{\psi}(\boldsymbol{\phi}_1), \boldsymbol{\psi}(\boldsymbol{\phi}_2)\} = \delta\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2\}$.
(v) The intrinsic discrepancy between $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ measures the minimum amount of information (in natural information units, *nits*) that one observation $\boldsymbol{x} \in X$ may be expected to provide in order to discriminate between $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ (Kullback, 1959).

Moreover, the intrinsic discrepancy has two further important properties which the directed logarithmic divergence does *not* have:

(vi) The intrinsic discrepancy is *symmetric* so that $\delta\{p_1(\boldsymbol{x}), p_2(\boldsymbol{x})\} = \delta\{p_2(\boldsymbol{x}), p_1(\boldsymbol{x})\}$.
(vii) If the two densities have strictly nested supports, so that $p_1(\boldsymbol{x}) > 0$ iff $\boldsymbol{x} \in X_1$, $p_2(\boldsymbol{x}) > 0$ iff $\boldsymbol{x} \in X_2$, and either $X_1 \subset X_2$ or $X_2 \subset X_1$, then the intrinsic discrepancy is still typically *finite*. More specifically, the intrinsic discrepancy then reduces to one of the directed logarithmic divergences while the other diverges, so that $\delta\{p_1, p_2\} = k(p_1 \mid p_2)$ when $X_2 \subset X_1$, and $\delta\{p_1, p_2\} = k(p_2 \mid p_1)$ when $X_1 \subset X_2$.

Example 1. Discrepancy between a Binomial distribution and its Poisson approximation. Let $p_1(x)$ be a binomial distribution $\mathrm{Bi}(x \mid n, \theta)$, and let $p_2(x)$ be its Poisson approximation $\mathrm{Pn}(x \mid n\theta)$. Since $X_1 \subset X_2$, $\delta(p_1, p_2) = k(p_2 \mid p_1)$; thus,

$$\delta\{p_1(x), p_2(x)\} = \delta(n, \theta) = \sum_{x=0}^{n} \mathrm{Bi}(x \mid n, \theta) \log \frac{\mathrm{Bi}(x \mid n, \theta)}{\mathrm{Pn}(x \mid n\theta)}.$$

The resulting discrepancy, $\delta(n, \theta)$ is plotted in Figure 1 as a function of θ for several values of n. As one might expect, the discrepancy converges to zero as θ decreases and as n increases, but it is

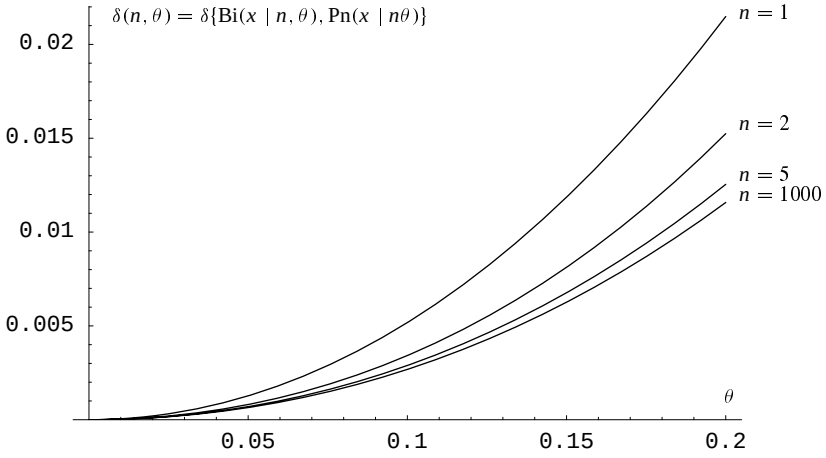

Figure 1. *Intrinsic discrepancy between a Binomial distribution* $\text{Bi}(x \mid n, \theta)$ *and a Poisson distribution* $\text{Pn}(x \mid n\theta)$ *as a function of* θ*, for* $n = 1, 2, 5$ *and* 1000*.*

apparent from the graph that the important condition for the approximation to work is that θ has to be small. ◁

The definition of the intrinsic divergence suggests an interesting new form of convergence for probability distributions:

Definition 2. Intrinsic Convergence. A sequence of probability distributions represented by their density functions $\{p_i(\boldsymbol{x})\}_{i=1}^{\infty}$ is said to converge *intrinsically* to a probability distribution with density $p(\boldsymbol{x})$ whenever $\lim_{i\to\infty} \delta(p_i, p) = 0$, that is, whenever the intrinsic discrepancy between $p_i(x)$ and $p(x)$ converges to zero. ◁

Example 2. Intrinsic convergence of Student densities to a Normal density. The intrinsic discrepancy between a standard Normal and a standard Student with α degrees of freedom is $\delta(\alpha) = \delta\{\text{St}(x \mid 0, 1, \alpha), \text{N}(x \mid 0, 1)\}$, *i.e.,*

$$\min\left\{\int_{-\infty}^{\infty} \text{St}(x \mid 0, 1, \alpha) \log \frac{\text{St}(x \mid 0, 1, \alpha)}{\text{N}(x \mid 0, 1)} dx,\ \int_{-\infty}^{\infty} \text{N}(x \mid 0, 1) \log \frac{\text{N}(x \mid 0, 1)}{\text{St}(x \mid 0, 1, \alpha)} dx\right\};$$

The second integral may be shown to be always smaller than the first, and to yield an analytical result (in terms of the Hypergeometric and Beta functions) which, for large α values, may be approximated by Stirling to obtain

$$\delta(\alpha) = \int_{-\infty}^{\infty} \text{N}(x \mid 0, 1) \log \frac{\text{N}(x \mid 0, 1)}{\text{St}(x \mid 0, 1, \alpha)} dx = \frac{1}{(1+\alpha)^2} + o(\alpha^{-2}),$$

a function which rapidly converges to zero. Thus, a sequence of standard Student densities with increasing degrees of freedom intrinsically converges to a standard normal density. ◁

In this paper, intrinsic discrepancies are basically used to measure the "distance" between alternative model assumptions about data $\boldsymbol{x} \in X$. Thus, $\delta\{p_1(\boldsymbol{x} \mid \phi), p_2(\boldsymbol{x} \mid \psi)\}$ is a symmetric measure (in natural information units, *nits*) of how different the probability densities $p_1(\boldsymbol{x} \mid \phi)$ and $p_2(\boldsymbol{x} \mid \psi)$ are from each other as a function of ϕ and ψ. Since, for any given data $\boldsymbol{x} \in X$, $p_1(\boldsymbol{x} \mid \phi)$ and $p_2(\boldsymbol{x} \mid \psi)$ are the respective likelihood functions, it follows from Definition 1 that

$\delta\{p_1(\boldsymbol{x} \mid \phi), p_2(\boldsymbol{x} \mid \psi)\} = \delta(\phi, \psi)$ may immediately be interpreted as the *minimum expected log-likelihood ratio in favour of the true model*, assuming that one of the two models is true. Indeed, if $p_1(\boldsymbol{x} \mid \phi_0) = p_2(\boldsymbol{x} \mid \psi_0)$ almost everywhere (and hence the models $p_1(\boldsymbol{x} \mid \phi_0)$ and $p_2(\boldsymbol{x} \mid \psi_0)$ are indistinguishable), then $\delta\{\phi_0, \psi_0)\} = 0$. In general, if either $p_1(\boldsymbol{x} \mid \phi_0)$ or $p_2(\boldsymbol{x} \mid \psi_0)$ is correct, then an intrinsic discrepancy $\delta(\phi_0, \psi_0) = d$ implies an average log-likelihood ratio for the true model of at least d, *i.e.*, minimum likelihood ratios for the true model of about e^d. If $\delta\{\phi_0, \psi_0)\} = 5$, $e^5 \approx 150$, so that data $\boldsymbol{x} \in X$ should then be expected to provide *strong evidence* to discriminate between $p_1(\boldsymbol{x} \mid \phi_0)$ and $p_2(\boldsymbol{x} \mid \psi_0)$. Similarly, if $\delta\{\phi_0, \psi_0)\} = 2.5$, $e^{2.5} \approx 12$, so that data $\boldsymbol{x} \in X$ should then only be expected to provide *mild evidence* to discriminate between $p_1(\boldsymbol{x} \mid \phi_0)$ and $p_2(\boldsymbol{x} \mid \psi_0)$.

Definition 3. Intrinsic Discrepancy Loss. The intrinsic discrepancy loss $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ from replacing the probability model $M = \{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega}),\ \boldsymbol{\theta} \in \Theta,\ \boldsymbol{\omega} \in \Omega,\ \boldsymbol{x} \in X\}$ by its restriction with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $M_0 = \{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}),\ \boldsymbol{\omega} \in \Omega,\ \boldsymbol{x} \in X\}$ is the intrinsic discrepancy between the probability density $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ and the family of probability densities $\{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}),\ \boldsymbol{\omega} \in \Omega\}$, that is

$$\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega}) = \min_{\boldsymbol{\omega}_0 \in \Omega} \delta\{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega}),\ p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}_0)\}. \qquad \triangleleft$$

The intrinsic discrepancy $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ between $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ and M_0 is the intrinsic discrepancy between the assumed probability density $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ and its closest approximation with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Notice that $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ is invariant under reparametrization of either $\boldsymbol{\theta}$ or $\boldsymbol{\omega}$. Moreover, if $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$ is a sufficient statistic for model M, then

$$\int_X p(\boldsymbol{x} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}) \log \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega})}{p(\boldsymbol{x} \mid \boldsymbol{\theta}_j, \boldsymbol{\omega}_j)}\, d\boldsymbol{x} = \int_T p(\boldsymbol{t} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}) \log \frac{p(\boldsymbol{t} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega})}{p(\boldsymbol{t} \mid \boldsymbol{\theta}_j, \boldsymbol{\omega}_j)}\, d\boldsymbol{t}; \tag{16}$$

thus, if convenient, $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ may be computed in terms of the sampling distribution of the sufficient statistic $p(\boldsymbol{t} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$, rather than in terms of the complete probability model $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$. Moreover, although not explicitly shown in the notation, the intrinsic discrepancy function typically depends on the sample size. Indeed, if data $\boldsymbol{x} \in X \subset \Re^n$, consist of a *random sample* $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of size n from $p(x \mid \boldsymbol{\theta}_i, \boldsymbol{\omega})$, then

$$\int_X p(\boldsymbol{x} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}) \log \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega})}{p(\boldsymbol{x} \mid \boldsymbol{\theta}_j, \boldsymbol{\omega}_j)}\, d\boldsymbol{x} = n \int_{\Re} p(x \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}) \log \frac{p(x \mid \boldsymbol{\theta}_i, \boldsymbol{\omega})}{p(x \mid \boldsymbol{\theta}_j, \boldsymbol{\omega}_j)}\, dx, \tag{17}$$

so that the intrinsic discrepancy associated with the full model $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ is simply n times the intrinsic discrepancy associated to the model $p(x \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ which corresponds to a single observation. Definition 3 may be used however in problems (say time series) where $\boldsymbol{x}$ does *not* consist of a random sample.

It immediately follows from (9) and (14) that, with an intrinsic discrepancy loss function, the hypothesis H_0 should be rejected if (and only if) the posterior expected advantage of rejecting $\boldsymbol{\theta}_0$, given model M and data $\boldsymbol{x}$, is sufficiently large, so that the decision criterion becomes

$$\text{Reject } H_0 \quad \text{iff} \quad d(\boldsymbol{\theta}_0, \boldsymbol{x}) = \int_\Theta \int_\Omega \delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})\, \pi(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{x})\, d\boldsymbol{\theta} d\boldsymbol{\omega} > d^*, \tag{18}$$

for some $d^* > 0$. Since $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ is non-negative, $d(\boldsymbol{\theta}_0, \boldsymbol{x})$ is nonnegative. Moreover, if $\phi = \phi(\boldsymbol{\theta})$ is a one-to-one transformation of $\boldsymbol{\theta}$, then $d(\phi(\boldsymbol{\theta}_0), \boldsymbol{x}) = d(\boldsymbol{\theta}_0, \boldsymbol{x})$, so that the expected intrinsic loss of rejecting H_0 is invariant under reparametrization.

The function $d(\boldsymbol{\theta}_0, \boldsymbol{x})$ is a continuous, non-negative measure of how inappropriate (in loss of information units) may be expected to be to simplify the model by accepting H_0. Indeed, $d(\boldsymbol{\theta}_0, \boldsymbol{x})$ is a precise measure of the (posterior) expected amount information (in *nits*) which would be necessary to recover the assumed probability density $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ from its closest approximation within $M_0 \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}),\ \boldsymbol{\omega} \in \Omega\}$; it is a measure of the 'strength of evidence' against M_0 given $M \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega}),\ \boldsymbol{\theta} \in \Theta,\ \boldsymbol{\omega} \in \Omega\}$ (cf. Good, 1950). In traditional language, $d(\boldsymbol{\theta}_0, \boldsymbol{x})$ is a

(monotone) *test statistic* for H_0, and the null hypothesis should be rejected if the value of $d(\boldsymbol{\theta}_0, \boldsymbol{x})$ exceeds some *critical value* d^*. Notice however that, in sharp contrast to conventional hypothesis testing, the critical value d^* is found to be a positive *utility constant*, which may precisely be described as the number of information units which the decision maker is prepared to lose in order to be able to work with the simpler model H_0, and which does *not* depend on the sampling properties of the test statistic. The procedure may be used with standard, continuous (possibly improper) regular priors when $\boldsymbol{\theta}$ is a continuous parameter (and hence $M_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ is a zero measure set).

Naturally, to implement the decision criterion, both the prior $\pi(\boldsymbol{\theta}, \boldsymbol{\omega})$ and the utility constant d^* must be chosen. These two important issues are now successively addressed, leading to a general decision criterion for hypothesis testing, the *Bayesian reference criterion*.

3.2 The Bayesian Reference Criterion (BRC)

Prior specification. An objective Bayesian procedure (objective in the sense that it depends exclusively on the assumed model and the observed data), requires an objective "non-informative" prior which mathematically describes lack on relevant information about the quantity of interest, and which only depends on the assumed statistical model and on the quantity of interest. Recent literature contains a number of requirements which may be regarded as necessary properties of any algorithm proposed to derive these 'baseline' priors; those requirements include general applicability, invariance under reparametrization, consistent marginalization, and appropriate coverage properties. The *reference analysis* algorithm, introduced by Bernardo (1979) and further developed by Berger & Bernardo (1989, 1992) is, to the best of our knowledge, the only available method to derive objective priors which satisfy all these desiderata. For an introduction to reference analysis, see Bernardo & Ramón (1998); for a textbook level description see Bernardo & Smith (1994, Ch. 5); for a critical overview of the topic, see Bernardo (1997), references therein and ensuing discussion.

Within a given probability model $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$, the joint prior $\pi_\phi(\boldsymbol{\theta}, \boldsymbol{\omega})$ required to obtain the (marginal) *reference* posterior $\pi(\phi \mid \boldsymbol{x})$ of some function of interest $\phi = \phi(\boldsymbol{\theta}, \boldsymbol{\omega})$ generally depends on the function of interest, and its derivation is not necessarily trivial. However, under regularity conditions (often met in practice) the required reference prior may easily be found. For instance, if the marginal posterior distribution of the function of interest $\pi(\phi \mid \boldsymbol{x})$ has an asymptotic approximation $\hat{\pi}(\phi \mid \boldsymbol{x}) = \hat{\pi}(\phi \mid \hat{\phi})$ which only depends on the data through a consistent estimator $\hat{\phi} = \hat{\phi}(\boldsymbol{x})$ of ϕ, then the ϕ-reference prior is simply obtained as

$$\pi(\phi) \propto \hat{\pi}(\phi \mid \hat{\phi})\Big|_{\hat{\phi}=\phi}. \tag{19}$$

In particular, if the posterior distribution of ϕ is asymptotically normal $\mathrm{N}(\phi \mid \hat{\phi}, s(\hat{\phi})/\sqrt{n})$, then $\pi(\phi) \propto s(\phi)^{-1}$, so that the reference prior reduces to Jeffreys' prior in one-dimensional, asymptotically normal conditions. If, moreover, the sampling distribution of $\hat{\phi}$ only depends on ϕ, so that $p(\hat{\phi} \mid \boldsymbol{\theta}, \boldsymbol{\omega}) = p(\hat{\phi} \mid \phi)$, then, by Bayes theorem, the corresponding reference posterior is

$$\pi(\phi \mid \boldsymbol{x}) \approx \pi(\phi \mid \hat{\phi}) \propto p(\hat{\phi} \mid \phi)\, \pi(\phi), \tag{20}$$

and the approximation is exact if, given the ϕ-reference prior $\pi_\phi(\boldsymbol{\theta}, \boldsymbol{\omega})$, $\hat{\phi}$ is marginally sufficient for ϕ (rather than just *asymptotically* marginally sufficient).

In our formulation of hypothesis testing, the function of interest (*i.e.*, the function of the parameters which drives the utility function) is the intrinsic discrepancy $\delta = \delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$. Thus, we propose to use the joint reference prior $\pi_\delta(\boldsymbol{\theta}, \boldsymbol{\omega})$ which corresponds to the function of interest $\delta = \delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$. This implies rejecting the null if (and only if) the reference posterior expectation of the intrinsic discrepancy, which will be referred to as the *intrinsic statistic* $d(\boldsymbol{\theta}_0, \boldsymbol{x})$, is sufficiently large. The

proposed test statistic is thus

$$d(\boldsymbol{\theta}_0, \boldsymbol{x}) = \int_{\Delta} \delta\, \pi_\delta(\delta \mid \boldsymbol{x})\, d\delta = \int_{\Theta} \int_{\Omega} \delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})\, \pi_\delta(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{x})\, d\boldsymbol{\theta} d\boldsymbol{\omega}, \tag{21}$$

where $\pi_\delta(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})\, \pi_\delta(\boldsymbol{\theta}, \boldsymbol{\omega})$ is the posterior distribution which corresponds to the δ-reference prior $\pi_\delta(\boldsymbol{\theta}, \boldsymbol{\omega})$.

Loss calibration. As described in Section 3.1, the intrinsic discrepancy between two fully specified probability models is simply the minimum expected log-likelihood ratio for the true model from data sampled from either of them. It follows that $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ measures, as a function of $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$, the minimum expected log-likelihood ratio for $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$, against a model of the form $p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}_0)$, for some $\boldsymbol{\omega}_0 \in \Omega$.

Consequently, given some data $\boldsymbol{x}$, the intrinsic statistic $d(\boldsymbol{\theta}_0, \boldsymbol{x})$, which is simply the reference posterior expectation of $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$, is an estimate (given the available data) of the expected log-likelihood ratio against the null model. This is a continuous measure of the evidence provided by the data against the (null) hypothesis that a model of the form $p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}_0)$, for some $\boldsymbol{\omega}_0 \in \Omega$, may safely be used as a proxy for the assumed model $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$. In particular, values of $d(\boldsymbol{\theta}_0, \boldsymbol{x})$ of about 2.5 or 5.0 should respectively be regarded as mild and strong evidence against the (null) hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Example 3. Testing the value of a Normal mean, σ known. Let data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from a normal distribution $\mathrm{N}(x \mid \mu, \sigma^2)$, where σ is assumed to be known, and consider the canonical problem of testing whether these data are (or are not) compatible with some precise hypothesis $H_0 \equiv \{\mu = \mu_0\}$ on the value of the mean. Given σ, the logarithmic divergence of $p(\boldsymbol{x} \mid \mu_0, \sigma)$ from $p(\boldsymbol{x} \mid \mu, \sigma)$ is the symmetric function

$$k(\mu_0 \mid \mu) = n \int_{\Re} N(x \mid \mu, \sigma^2) \log \frac{N(x \mid \mu, \sigma^2)}{N(x \mid \mu_0, \sigma^2)}\, dx = \frac{n}{2} \left(\frac{\mu - \mu_0}{\sigma} \right)^2. \tag{22}$$

Thus, the intrinsic discrepancy in this problem is simply

$$\delta(\mu_0, \mu) = \frac{n}{2} \left(\frac{\mu - \mu_0}{\sigma} \right)^2 = \frac{1}{2} \left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}} \right)^2, \tag{23}$$

half the square of the standardized distance between μ and μ_0. For known σ, the intrinsic discrepancy $\delta(\mu_0, \mu)$ is a piecewise invertible transformation of μ and, hence, the δ-reference prior is simply $\pi_\delta(\mu) = \pi_\mu(\mu) = 1$. The corresponding reference posterior distribution of μ is $\pi_\delta(\mu \mid \boldsymbol{x}) = \mathrm{N}(\mu \mid \overline{x}, \sigma^2/n)$ and, therefore, the intrinsic statistic (the reference posterior expectation of the intrinsic discrepancy) is

$$d(\mu_0, \boldsymbol{x}) = \frac{n}{2} \int_{\Re} \left(\frac{\mu - \mu_0}{\sigma} \right)^2 \mathrm{N}\left(\mu \,\middle|\, \overline{x}, \frac{\sigma^2}{n} \right) d\mu = \tfrac{1}{2}(1 + z^2), \tag{24}$$

where $z = (\overline{x} - \mu_0)/(\sigma/\sqrt{n})$. Thus, $d(\mu_0, \boldsymbol{x})$ is a simple transformation of z, the number of standard deviations which μ_0 lies away from the data mean $\overline{x}$. The sampling distribution of z^2 is noncentral Chi squared with one degree of freedom and noncentrality parameter 2δ, and its expected value is $1 + 2\delta$, where $\delta = \delta(\mu_0, \mu)$ is the intrinsic discrepancy given by (23). It follows that, in this canonical problem, the expected value under repeated sampling of the reference statistic $d(\mu_0, \boldsymbol{x})$ is equal to one if $\mu = \mu_0$, and increases linearly with n if $\mu \neq \mu_0$.

Scientists have often expressed the view (see *e.g.*, Jaynes, 1980, or Jeffreys, 1980) that, in this canonical situation, $|z| \approx 2$ should be considered as a mild indication of evidence against $\mu = \mu_0$, while $|z| > 3$ should be regarded as strong evidence against $\mu = \mu_0$. In terms of the intrinsic statistic $d(\mu_0, \boldsymbol{x}) = (1 + z^2)/2$ this precisely corresponds to issuing warning signals whenever $d(\mu_0, \boldsymbol{x})$ is

about 2.5 nits, and to reject the null whenever $d(\mu_0, \boldsymbol{x})$ is larger than 5 nits, in perfect agreement with the log-likelihood ratio calibration mentioned above. ◁

Notice, however, that the information scale suggested is an *absolute* scale which is independent of the problem considered, so that rejecting the null whenever its (reference posterior) expected intrinsic discrepancy from the true model is larger than (say) $d^* = 5$ natural units of information is a *general* rule (and one which corresponds to the conventional '3σ' rule in the canonical normal case). Notice too that the use of the ubiquitous 5% confidence level in this problem would correspond to $z = 1.96$, or $d^* = 2.42$ nits, which only indicates mild evidence against the null; this is consistent with other arguments (see *e.g.*, Berger & Delampady, 1987) suggesting that a p-value of about 0.05 does *not* generally provide sufficient evidence to definitely reject the null hypothesis.

The preceding discussion justifies the following formal definition of an (objective) Bayesian reference criterion for hypothesis testing:

Definition 3. Bayesian Reference Criterion (BRC). Let $\{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega}), \boldsymbol{\theta} \in \Theta, \boldsymbol{\omega} \in \Omega\}$, be a statistical model which is assumed to have been generated some data $\boldsymbol{x} \in X$, and consider a precise value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ among those which remain *possible* after $\boldsymbol{x}$ has ben observed. To decide whether or not the precise value $\boldsymbol{\theta}_0$ may be used as a proxy for the unknown value of $\boldsymbol{\theta}$,

(i) compute the intrinsic discrepancy $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$;
(ii) derive the corresponding reference posterior expectation $d(\boldsymbol{\theta}_0, \boldsymbol{x}) = \mathrm{E}[\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega}) \mid \boldsymbol{x}]$, and state this number as a measure of evidence against the (null) hypothesis $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$.
(iii) If a formal decision is required, reject the null if, and only if, $d(\boldsymbol{\theta}_0, \boldsymbol{x}) > d^*$, for some context dependent d^*. The values $d^* \approx 1.0$ (no evidence against the null), $d^* \approx 2.5$ (mild evidence against the null) and $d^* > 5$ (significant evidence against the null) may conveniently be used for scientific communication.

◁

The results derived in Example 3 may be used to analyze the large sample behaviour of the proposed criterion in one-parameter problems. Indeed, if $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is a large random sample from a one-parameter regular model $\{p(x \mid \theta), \theta \in \Theta\}$, the relevant reference prior will be Jeffreys' prior $\pi(\theta) \propto i(\theta)^{1/2}$, where $i(\theta)$ is Fisher's information function. Hence, the reference prior of $\phi(\theta) = \int^{\theta} i(\theta)^{1/2} d\theta$ will be uniform, and the reference posterior of ϕ approximately normal $\mathrm{N}(\phi \mid \hat{\phi}, 1/\sqrt{n})$. Thus, using Example 3 and the fact that the intrinsic statistic is invariant under one-to-one parameter transformations, one gets the approximation $d(\theta_0, \boldsymbol{x}) = d(\phi_0, \boldsymbol{x}) \approx \frac{1}{2}(1 + z^2)$, where $z = \sqrt{n}(\hat{\phi} - \phi_0)$. Moreover, the sampling distribution of z will approximately be a non-central χ^2 with one degree of freedom and non centrality parameter $n(\phi - \phi_0)^2$. Hence, the expected value of $d(\phi_0, \boldsymbol{x})$ under repeated sampling from $p(x \mid \theta)$ will approximately be one if $\theta = \theta_0$ and will linearly increase with $n(\theta - \theta_0)^2$ otherwise. More formally, we may state

Proposition 1. One-Dimensional Asymptotic Behaviour. If $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is a random sample from a regular model $\{p(x \mid \theta), \theta \in \Theta \subset \Re, x \in X \subset \Re\}$ with one continuous parameter, and $\phi(\theta) = \int^{\theta} i(\theta)^{1/2} d\theta$, where $i(\theta) = -\mathrm{E}_{x \mid \theta}[\partial^2 \log p(x \mid \theta)/\partial \theta^2]$, then the intrinsic statistic $d(\theta_0, \boldsymbol{x})$ to test $\{\theta = \theta_0\}$ is

$$d(\theta_0, \boldsymbol{x}) = \tfrac{1}{2}[1 + z^2(\theta_0, \hat{\theta})] + o(n^{-1}), \quad z(\theta_0, \hat{\theta}) = \sqrt{n}[\phi(\hat{\theta}) - \phi(\theta)],$$

where $\hat{\theta} = \hat{\theta}(\boldsymbol{x}) = \arg\max p(\boldsymbol{x} \mid \theta)$. Moreover, the expected value of $d(\theta_0, \boldsymbol{x})$ under repeated sampling is

$$\mathrm{E}_{x \mid \theta}[d(\theta_0, \boldsymbol{x})] = 1 + n[\phi(\theta) - \phi(\theta_0)]^2 + o(n^{-1}),$$

so that $d(\theta_0, \boldsymbol{x})$ will concentrate around the value one if $\theta = \theta_0$, and will linearly increase with n otherwise.

◁

The arguments leading to Proposition 1 may be extended to multivariate situations, with or without nuisance parameters.

In the final section of this paper we illustrate the behaviour of the Bayesian reference criterion with three examples: (i) hypothesis testing on the value of a binomial parameter, which is used to illustrate the shape of an intrinsic discrepancy, (ii) a problem of precise hypothesis testing within a non-regular probability model, which is used to illustrate the exact behaviour of the BRC criterion under repeated sampling, and (iii) a multivariate normal problem which illustrates how the proposed procedure avoids Rao's paradox on incoherent multivariate frequentist testing.

4 Examples

4.1 Testing the Value of the Parameter of a Binomial Distribution

Let data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ consist of n conditionally independent Bernoulli observations with parameter θ, so that $p(x \mid \theta) = \theta^x (1-\theta)^{1-x}$, $0 < \theta < 1$, $x \in \{0, 1\}$, and consider testing whether or not the observed data $\boldsymbol{x}$ are compatible with the null hypothesis $\{\theta = \theta_0\}$. The directed logarithmic divergence of $p(x \mid \theta_j)$ from $p(x \mid \theta_i)$ is

$$k(\theta_j \mid \theta_i) = \theta_i \log \frac{\theta_i}{\theta_j} + (1-\theta_i) \log \frac{(1-\theta_i)}{(1-\theta_j)}\,, \tag{25}$$

and it is easily verified that $k(\theta_j \mid \theta_i) < k(\theta_i \mid \theta_j)$ iff $\theta_i < \theta_j < 1 - \theta_i$; thus, the intrinsic discrepancy between $p(\boldsymbol{x} \mid \theta_0)$ and $p(\boldsymbol{x} \mid \theta)$, represented in Figure 2, is

$$\delta(\theta_0, \theta) = n \begin{cases} k(\theta \mid \theta_0) & \theta \in (\theta_0, 1-\theta_0), \\ k(\theta_0 \mid \theta) & \text{otherwise.} \end{cases} \tag{26}$$

Since $\delta(\theta_0, \theta)$ is a piecewise invertible function of θ, the δ-reference prior is just the θ-reference prior and, since Bernoulli is a regular model, this is Jeffreys' prior, $\pi(\theta) = \text{Be}(\theta \mid 1/2, 1/2)$. The reference posterior is the Beta distribution $\pi(\theta \mid \boldsymbol{x}) = \pi(\theta \mid r, n) = \text{Be}(\theta \mid r + 1/2, n - r + 1/2)$, with $r = \sum x_i$, and the intrinsic statistic $d(\theta_0, \boldsymbol{x})$ is the concave function

$$d(\theta_0, \boldsymbol{x}) = d(\theta_0, r, n) = \int_0^1 \delta(\theta_0, \theta)\, \pi(\theta \mid r, n)\, d\theta = \tfrac{1}{2}[1 + z(\theta_0, \hat{\theta})^2] + o(n^{-1}) \tag{27}$$

where $z(\theta_0, \hat{\theta}) = \sqrt{n}[\phi(\hat{\theta}) - \phi(\theta_0)]$, and $\phi(\theta) = 2\text{ArcSin}(\sqrt{\theta})$. The exact value of the intrinsic statistic may easily be found by one-dimensional numerical integration, or may be expressed in terms of Digamma and incomplete Beta functions, but the approximation given above, directly obtained from Proposition 1, is quite good, even for moderate samples.

The canonical particular case where $\theta_0 = 1/2$ deserves special attention. The exact value of the intrinsic statistic is then

$$d(1/2, r, n) = \psi(n+1) + \tilde{\theta}\, \psi(r + 1/2) + (1 - \tilde{\theta})\, \psi(n - r + 1/2) - \log 2 \tag{28}$$

where $\tilde{\theta} = (r + 1/2)/(n+1)$ is the reference posterior mean. As one would certainly expect, $d(1/2, 0, n) = d(1/2, n, n)$ increases with n; moreover, it is found that $d(1/2, 0, 6) = 2.92$ and that $d(1/2, 0, 10) = 5.41$. Thus, when $r = 0$ (all failures) or $r = n$ (all successes) the null value $\theta_0 = 1/2$ should be questioned ($d > 2.5$) for all $n > 5$ and definitely rejected ($d > 5$) for all $n > 9$.

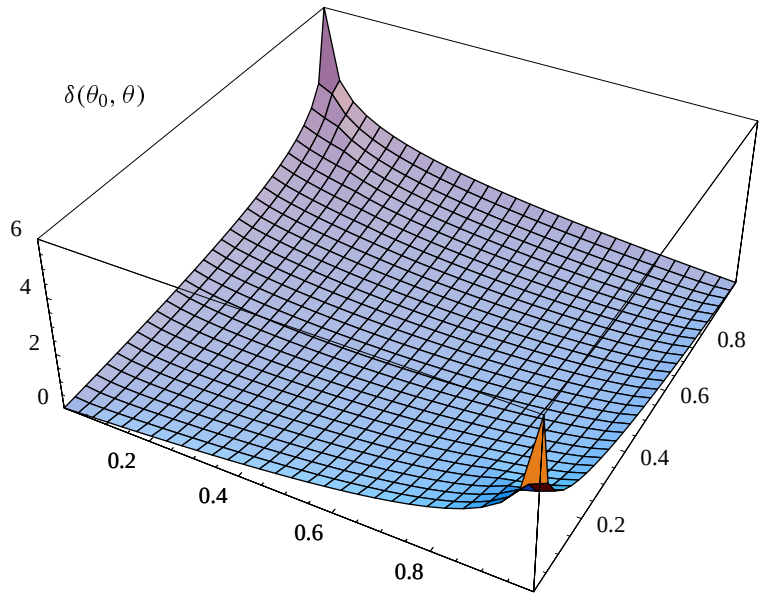

Figure 2. *Intrinsic discrepancy between two Bernoulli probability models.*

4.2 Testing the Value of the Upper Limit of a Uniform Distribution

Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}, x_i \in X(\theta) = [0, \theta]$ be a random sample of n uniform observations in $[0, \theta]$, so that $p(x_i \mid \theta) = \theta^{-1}$, and consider testing the compatibility of data $\boldsymbol{x}$ with the precise value $\theta = \theta_0$. The logarithmic divergence of $p(\boldsymbol{x} \mid \theta_j)$ from $p(\boldsymbol{x} \mid \theta_i)$ is

$$k(\theta_j \mid \theta_i) = n \int_0^{\theta_i} p(x \mid \theta_i) \log \frac{p(x \mid \theta_i)}{p(x \mid \theta_j)}\, dx = \begin{cases} n \log(\theta_j/\theta_i) & \text{if } \theta_i < \theta_j \\ \infty & \text{otherwise} \end{cases} \tag{29}$$

and, therefore, the intrinsic discrepancy between $p(x \mid \theta)$ and $p(x \mid \theta_0)$ is

$$\delta(\theta_0, \theta) = \min\{k(\theta_0 \mid \theta), k(\theta \mid \theta_0)\} = \begin{cases} n \log(\theta_0/\theta) & \text{if } \theta_0 > \theta \\ n \log(\theta/\theta_0) & \text{if } \theta_0 \le \theta. \end{cases} \tag{30}$$

Let $x_{(n)} = \max\{x_1, \ldots, x_n\}$ be the largest observation in the sample. The likelihood function is $p(\boldsymbol{x} \mid \theta) = \theta^{-n}$, if $\theta > x_{(n)}$, and zero otherwise; hence, $x_{(n)}$ is a sufficient statistic, and a simple asymptotic approximation $\hat{\pi}(\theta \mid \boldsymbol{x})$ to the posterior distribution of θ is given by

$$\hat{\pi}(\theta \mid \boldsymbol{x}) = \hat{\pi}(\theta \mid x_{(n)}) = \frac{\theta^{-n}}{\int_{x_{(n)}}^{\infty} \theta^{-n}\, d\theta} = (n-1)\, x_{(n)}^{n-1}\, \theta^{-n}, \quad \theta > x_{(n)}. \tag{31}$$

It immediately follows from (31) that $x_{(n)}$ is a consistent estimator of θ; hence, using (19), the θ-reference prior is given by

$$\pi_\theta(\theta) \propto \hat{\pi}(\theta \mid x_{(n)})\Big|_{x_{(n)}=\theta} \propto \theta^{-1}. \tag{32}$$

Moreover, for any θ_0, $\delta = \delta(\theta_0, \theta)$ is a piecewise invertible function of θ and, hence, the δ-reference prior is also $\pi_\delta(\theta) = \theta^{-1}$. Using Bayes theorem, the corresponding reference posterior is

$$\pi_\delta(\theta \mid \boldsymbol{x}) = \pi_\delta(\theta \mid x_{(n)}) = n\, x_{(n)}^n\, \theta^{-(n+1)}, \quad \theta > x_{(n)}; \tag{33}$$

thus, the intrinsic statistic to test the compatibility of the data with any possible value θ_0, *i.e.*, such that $\theta_0 > x_{(n)}$, is given by

$$d(\theta_0, \boldsymbol{x}) = d(t) = \int_{x_{(n)}}^{\infty} \delta(\theta_0, \theta)\, \pi_\delta(\theta \mid x_{(n)})\, d\theta = 2t - \log t - 1, \quad t = (x_{(n)}/\theta_0)^n, \tag{34}$$

which only depends on $t = t(\theta_0, x_{(n)}, n) = (x_{(n)}/\theta_0)^n \in [0, 1]$. The intrinsic statistic $d(t)$ is the concave function represented in Figure 3, which has a unique minimum at $t = 1/2$. Hence, the value of $d(\theta_0, \boldsymbol{x})$ is minimized iff $(x_{(n)}/\theta_0)^n = 1/2$, *i.e.*, iff $\theta_0 = 2^{1/n} x_{(n)}$, which is the Bayes estimator for this loss function (and the median of the reference posterior distribution).

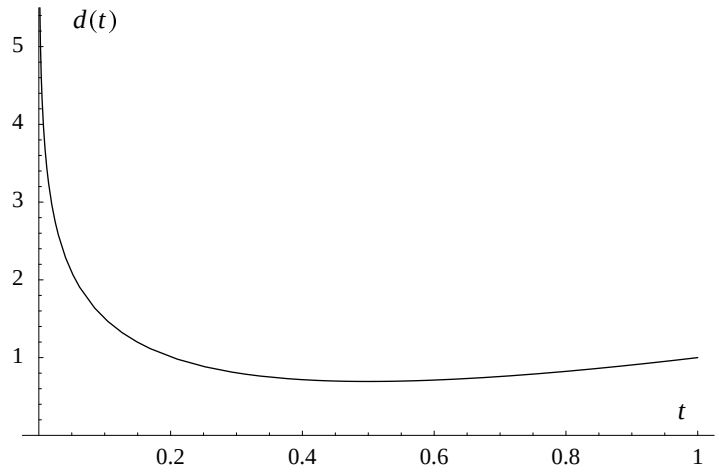

Figure 3. *The intrinsic statistic* $d(\theta_0, \boldsymbol{x}) = d(t) = 2t - \log t - 1$ *to test* $\theta = \theta_0$ *which corresponds to a random sample* $\{x_1 \ldots, x_n\}$ *from uniform distribution* $Un(x \mid 0, \theta)$, *as a function of* $t = (x_{(n)}/\theta_0)^n$.

It may easily be shown that the distribution of t under repeated sampling is uniform in $[0, (\theta/\theta_0)^n]$ and, hence, the expected value of $d(\theta_0, \boldsymbol{x}) = d(t)$ under repeated sampling is

$$\mathrm{E}[d(t) \mid \theta] = \int_0^{(\theta/\theta_0)^n} (2t - \log t - 1)\, dt = (\theta/\theta_0)^n - n \log(\theta/\theta_0), \tag{35}$$

which is precisely equal to one if $\theta = \theta_0$, and increases linearly with n otherwise. Thus, once again, one would expect $d(t)$ values to be about one under the null, and one would expect to always reject

a false null for a large enough sample. It could have been argued that $t = (x_{(n)}/\theta_0)^n$ is indeed a 'natural' intuitive measure of the evidence provided by the data against the precise value θ_0, but this is not needed; the procedure outlined *automatically* provides an appropriate test function for *any* hypothesis testing problem.

The relationship between BRC and both frequentist testing and Bayesian tail area testing procedures is easily established in this example. Indeed,

(i) The sampling distribution of t under the null is uniform in $[0, 1]$, so that t is precisely the p-value which corresponds to a frequentist test based on any one-to-one function of t.
(ii) The posterior tail area, that is, the reference posterior probability that θ is larger than θ_0, is $\int_{\theta_0}^{\infty} \pi(\theta \mid x_{(n)})\, d\theta = (x_{(n)}/\theta_0)^n = t$, so that t is *also* the reference posterior tail area.

It is immediately verified that $d(0.035) = 2.42$, and that $d(0.0025) = 5$. It follows that, in this problem, the bounds $d^* = 2.42$ and $d^* = 5$, respectively correspond to the p-values 0.035 and 0.0025. Notice that these numbers are *not* equal to the values 0.05 and 0.0027 obtained when testing a value $\mu = \mu_0$ for a univariate normal mean. This illustrates an important general point: for comparable strength of evidence in terms of information loss, the significance level should depend on the assumed statistical model (even in simple, one-dimensional problems).

4.3 Testing the Value of a Multivariate Normal Mean

Let $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a random sample from $\mathrm{N}_k(\boldsymbol{x} \mid \boldsymbol{\mu}, \sigma^2\Sigma)$, a multivariate normal distribution of dimension k, where Σ is a known symmetric positive-definite matrix. In this final example, tests on the value of $\boldsymbol{\mu}$ are presented for the case where σ is known. Tests for the case where σ is unknown, tests on the value of some of the components of $\boldsymbol{\mu}$, and tests on the values of regression coefficients $\boldsymbol{\beta}$ in normal regression models of the form $\mathrm{N}_k(\boldsymbol{y} \mid X\boldsymbol{\beta}, \sigma^2\Sigma)$, may be obtained from appropriate extensions of the results described below, and will be presented elsewhere.

Intrinsic discrepancy Without loss of generality, it may be assumed that $\sigma = 1$, for otherwise σ may be included in the matrix Σ; since Σ is known, the vector of means $\overline{\boldsymbol{x}}$ is a sufficient statistic. The sampling distribution of $\overline{\boldsymbol{x}}$ is $p(\overline{\boldsymbol{x}} \mid \boldsymbol{\mu}) = \mathrm{N}_k(\overline{\boldsymbol{x}} \mid \boldsymbol{\mu}, n^{-1}\Sigma)$; thus, using (16), the logarithmic divergence of $p(\boldsymbol{x} \mid \boldsymbol{\mu}_j)$ from $p(\boldsymbol{x} \mid \boldsymbol{\mu}_i)$ is the symmetric function

$$k(\boldsymbol{\mu}_j \mid \boldsymbol{\mu}_i) = \int_{\Re^k} p(\overline{\boldsymbol{x}} \mid \boldsymbol{\mu}_i) \log \frac{p(\overline{\boldsymbol{x}} \mid \boldsymbol{\mu}_i)}{p(\overline{\boldsymbol{x}} \mid \boldsymbol{\mu}_j)}\, d\overline{\boldsymbol{x}} = \frac{n}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'\Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j). \tag{36}$$

It follows that the intrinsic discrepancy between the null model $p(\boldsymbol{x} \mid \boldsymbol{\mu}_0)$ and the assumed model $p(\boldsymbol{x} \mid \boldsymbol{\mu})$ has the quadratic form

$$\delta(\boldsymbol{\mu}_0, \boldsymbol{\mu}) = \frac{n}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0). \tag{37}$$

The required test statistic, the intrinsic statistic, is the reference posterior expectation of $\delta(\boldsymbol{\mu}_0, \boldsymbol{\mu})$, $d(\boldsymbol{\mu}_0, \boldsymbol{x}) = \int_{\Re^k} \delta(\boldsymbol{\mu}_0, \boldsymbol{\mu})\, \pi_\delta(\boldsymbol{\mu} \mid \boldsymbol{x})\, d\boldsymbol{\mu}$.

Marginal reference prior We first make use of standard normal distribution theory to obtain the marginal reference prior distribution of $\lambda = (\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$, and hence that of $\delta = n\lambda/2$. Reference priors only depend on the asymptotic behaviour of the model and, for any regular prior, the posterior distribution of $\boldsymbol{\mu}$ is asymptotically multivariate normal $\mathrm{N}_k(\boldsymbol{\mu} \mid \overline{\boldsymbol{x}}, n^{-1}\Sigma)$. Consider $\boldsymbol{\eta} = A(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$, where $A'A = \Sigma^{-1}$, so that $\lambda = \boldsymbol{\eta}'\boldsymbol{\eta}$; the posterior distribution of $\boldsymbol{\eta}$ is asymptotically normal $\mathrm{N}_k(\boldsymbol{\eta} \mid A(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0), n^{-1}I_k)$. Hence (see *e.g.*, Rao, 1973, Ch. 3), the posterior distribution of $n\lambda = n\,\boldsymbol{\eta}'\boldsymbol{\eta} = n\,(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$ is asymptotically a non-central Chi squared with k degrees

of freedom and non-centrality parameter $n\,\hat{\lambda}$, with $\hat{\lambda} = (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)' \Sigma^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)$, and this distribution has mean $k + n\,\hat{\lambda}$ and variance $2(k + 2n\,\hat{\lambda})$. It follows that the marginal posterior distribution of λ is asymptotically normal; specifically,

$$p(\lambda \mid \boldsymbol{x}) \approx \mathrm{N}(\lambda \mid (k + n\,\hat{\lambda})/n,\ 2(k + 2n\,\hat{\lambda})/n^2) \approx \mathrm{N}(\lambda \mid \hat{\lambda},\ 4\hat{\lambda}/n). \tag{38}$$

Hence, the posterior distribution of λ has an asymptotic approximation $\hat{\pi}(\lambda \mid \hat{\lambda})$ which only depends on the data through $\hat{\lambda}$, a consistent estimator of λ. Therefore, using (19), the λ-reference prior is

$$\pi_\lambda(\lambda) \propto \hat{\pi}(\lambda \mid \hat{\lambda})\Big|_{\hat{\lambda}=\lambda} \propto \lambda^{-1/2}. \tag{39}$$

But the parameter of interest, $\delta = n\lambda/2$, is a linear transformation of λ and, therefore, the δ-reference prior is

$$\pi_\delta(\delta) \propto \pi_\lambda(\lambda) |\partial\lambda/\partial\delta| \propto \delta^{-1/2}. \tag{40}$$

Reference posterior and intrinsic statistic. Normal distribution theory may be used to derive the *exact* sampling distribution of the asymptotically sufficient estimator $\hat{\lambda} = (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)' \Sigma^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)$. Indeed, letting $\boldsymbol{y} = A(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)$, with $A'A = \Sigma^{-1}$, the sampling distribution of $\boldsymbol{y}$ is normal $\mathrm{N}_k(\boldsymbol{y} \mid A(\boldsymbol{\mu} - \boldsymbol{\mu}_0), n^{-1} I_k)$; thus, the sampling distribution of $n\,\boldsymbol{y}'\boldsymbol{y} = n\,\hat{\lambda}$ is a non-central Chi squared with k degrees of freedom and non-centrality parameter $n\,(\boldsymbol{\mu} - \boldsymbol{\mu}_0)' \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)$, which by equation (37) is precisely equal to 2δ. Thus, the asymptotic marginal posterior distribution of δ only depends on the data through the statistic,

$$z^2 = n\,\hat{\lambda} = n\,(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)' \Sigma^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0), \tag{41}$$

whose sampling distribution only depends on δ. Therefore, using (20), the reference posterior distribution of δ given z^2 is

$$\pi(\delta \mid z^2) \propto \pi(\delta)\, p(z^2 \mid \delta) = \delta^{-1/2} \chi^2(z^2 \mid k, 2\delta). \tag{42}$$

Transforming to polar coordinates it may be shown (Berger, Philippe & Robert, 1998) that (42) is actually the reference posterior distribution of δ which corresponds to the ordered parametrization $\{\delta, \boldsymbol{\omega}\}$, where $\boldsymbol{\omega}$ is the vector of the angles, so that, using such a prior, $\pi(\delta \mid \boldsymbol{x}) = \pi(\delta \mid z^2)$, and z^2 encapsulates all available information about the value of δ.

After some tedious algebra, both the missing proportionality constant, and the expected value of $\pi(\delta \mid z^2)$ may be obtained in terms of the ${}_1F_1$ confluent hypergeometric function, leading to

$$d(\boldsymbol{\mu_0}, z^2) = \mathrm{E}[\delta \mid k, z^2] = \frac{1}{2} \frac{{}_1F_1(3/2; k/2, z^2/2)}{{}_1F_1(1/2; k/2, z^2/2)}. \tag{43}$$

Moreover, the exact value for $\mathrm{E}[\delta \mid k, z^2]$ given by (43) has a simple linear approximation for large values of z^2, namely,

$$\mathrm{E}[\delta \mid k, z^2] \approx \frac{1}{2}\,(2 - k + z^2). \tag{44}$$

Notice that, in general, (44) is only appropriate for values of z^2 which are large relative to k (showing strong evidence against the null), but it is actually exact for $k = 1$, so that (43) provides a multivariate generalization of (24). Figure 4 shows the form of $\mathrm{E}[\delta \mid k, z^2]$ as a function of z^2 for different values of the dimension k.

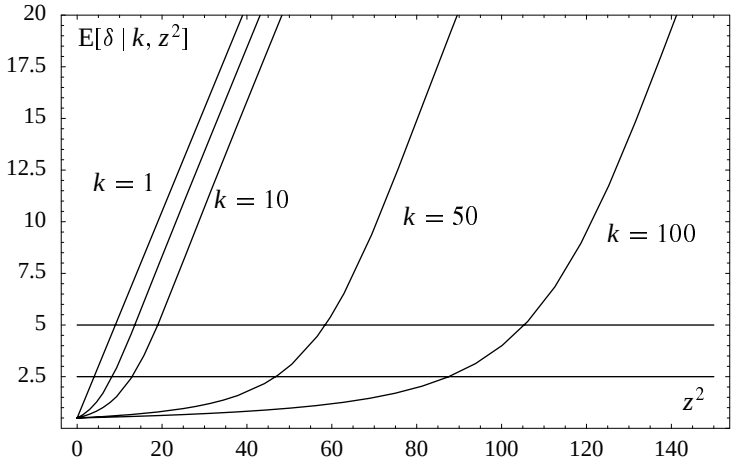

Figure 4. *Approximate behaviour of the intrinsic statistic* $d(\boldsymbol{\mu}_0, \boldsymbol{x}) \approx E[\delta \mid k, z^2]$ *as a function of* $z^2 = n\,(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)' \Sigma^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)$*, for* $k = 1, 5, 10, 50$ *and* 100*.*

Numerical Example: Rao's paradox As an illustrative numerical example, consider one observation $\boldsymbol{x} = (2.06, 2.06)$ from a bivariate normal density with variances $\sigma_1^2 = \sigma_2^2 = 1$ and correlation coefficient $\rho = 0.5$; the problem is to test whether or not the data $\boldsymbol{x}$ are compatible with the null hypothesis $\boldsymbol{\mu} = (0, 0)$. These data were used by Rao (1966) (and reassessed by Healy, 1969), to illustrate the often neglected fact that using standard significance tests, it can happen that a test for $\mu_1 = 0$ can lead to rejection at the same time as one for $\mu_2 = 0$, whereas the test for $\boldsymbol{\mu} = (0, 0)$ can result in acceptance, a clear example of frequentist incoherence, often known as Rao's paradox. Indeed, with those data, both $\mu_1 = 0$ and $\mu_2 = 0$ are rejected at the 5% level (since $x_1^2 = x_2^2 = 2.06^2 = 4.244$, larger than 3.841, the 0.95 quantile of a χ_1^2), while the same (Hotelling's T^2) test leads to acceptance of $\boldsymbol{\mu} = (0, 0)$ at the same level (since $z^2 = \boldsymbol{x}' \Sigma^{-1} \boldsymbol{x} = 5.658$, smaller than 5.991, the 0.95 quantile of a χ_2^2). However, using (43), we find,

$$\begin{cases} E[\delta \mid 1, 2.06^2] & = \frac{1}{2}(1 + 2.06^2) = 2.622, \\ E[\delta \mid 2, 5.658] & = \frac{1}{2} \frac{{}_1F_1(3/2;\, 1,\, 5.658/2)}{{}_1F_1(1/2;\, 1,\, 5.658/2)} = 2.727. \end{cases} \tag{45}$$

Thus, the BRC criterion suggests tentative rejection in both cases (since both numbers are larger than 2.5, the '2σ' rule in the canonical normal case), with some extra evidence in the bivariate case, as intuition clearly suggests.

Acknowledgements

The authors thank Professor Dennis Lindley, the Journal Editor Professor Elja Arjas, and an anonymous referee, for helpful comments on an earlier version of the paper. J.M. Bernardo was funded with grants BFM2001-2889 of the DGICYT Madrid and GV01-7 of Generalitat Valenciana (Spain). R. Rueda was funded with grant CONACyT 32256-E (Mexico).

References

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Berlin: Springer.

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

Berger, J.O. & Bernardo, J.M. (1992). On the development of reference priors. In *Bayesian Statistics 4*, Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, pp. 35–60 (with discussion). Oxford: University Press.
Berger, J.O. & Delampady, M. (1987). Testing precise hypotheses. *Statist. Sci.*, **2**, 317–352 (with discussion).
Berger, J.O., Philippe, A. & Robert, C.P. (1998). Estimation of Quadratic Functions: Noninformative priors for non-centrality parameters. *Statistica Sinica*, **8**, 359–376.
Berger, J.O. & Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of significance levels and evidence. *J. Amer. Statist. Assoc.*, **82**, 112–133 (with discussion).
Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc.*, **41**, 113–147 (with discussion). Reprinted In *Bayesian Inference*, Eds. N.G. Polson and G.C. Tiao, (1995) pp. 229–263. Brookfield, VT: Edward Elgar.
Bernardo, J.M. (1980). A Bayesian analysis of classical hypothesis testing. In *Bayesian Statistics*, Eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, pp. 605–647 (with discussion). Valencia: University Press.
Bernardo, J.M. (1982). Contraste de modelos probabilísticos desde una perspectiva Bayesiana. *Trab. Estadist.*, **33**, 16–30.
Bernardo, J.M. (1985). Análisis Bayesiano de los contrastes de hipótesis paramétricos. *Trab. Estadist.*, **36**, 45–54.
Bernardo, J.M. (1997). Noninformative priors do not exist. *J. Statist. Planning and Inference*, **65**, 159–189 (with discussion).
Bernardo, J.M. (1999). Nested hypothesis testing: The Bayesian reference criterion. In *Bayesian Statistics 6*, Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, pp. 101–130 (with discussion). Oxford: University Press.
Bernardo, J.M. & Bayarri, M.J. (1985). Bayesian model criticism. In *Model Choice*, Eds. J.-P. Florens, M. Mouchart, J.-P. Raoult and L. Simar, pp. 43–59. Brussels: Pub. Fac. Univ. Saint Louis.
Bernardo, J.M. & Ramón, J.M. (1998). An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters. *The Statistician*, **47**, 101–135.
Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*. Chichester: Wiley.
Casella, G. & Berger, R.L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc.*, **82**, 106–135 (with discussion).
Edwards, W.L., Lindman, H. & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* **70**, 193–242. Reprinted in *Robustness of Bayesian Analysis*, Ed. J.B. Kadane, (1984) pp. 1–62. Amsterdam: North Holland. Reprinted in *Bayesian Inference*, Eds. N.G. Polson and G.C. Tiao (1995) pp. 140–189. Brookfield, VT: Edward Elgar.
Ferrándiz, J.R. (1985). Bayesian inference on Mahalanobis distance: an alternative approach to Bayesian model testing. In *Bayesian Statistics 2*, Eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, pp. 645–654. Amsterdam: North-Holland.
Good, I.J. (1950). *Probability and the Weighing of Evidence*. London: Griffin; New York: Hafner Press.
Good, I.J. (1983). *Good Thinking: The Foundations of Probability and its Applications*. Minneapolis: Univ. Minnesota Press.
Gutiérrez-Peña, E. (1992). Expected logarithmic divergence for exponential families. In *Bayesian Statistics 4*, Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, pp. 669–674. Oxford: University Press.
Healy, J.R. (1969). Rao's paradox concerning multivariate tests of significance. *Biometrics*, **25**, 411–413.
Jaynes, E.T. (1980). Discussion to the session on hypothesis testing. In *Bayesian Statistics*, Eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, pp. 618–629. Valencia: University Press. Reprinted in *E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, Ed. R.D. Rosenkranz (1983) pp. 378–400. Dordrecht: Kluwer.
Jeffreys, H. (1961). *Theory of Probability*. (3rd edition) Oxford: University Press.
Jeffreys, H. (1980). Some general points in probability theory. In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys*, Ed. A. Zellner, pp. 451–453. Amsterdam: North-Holland.
Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley. Second edition in 1968, New York: Dover. Reprinted in 1978, Gloucester, MA: Peter Smith.
Kullback, S. & Leibler, R.A. (1951). On information and sufficiency. *Ann. Math. Statist.*, **22**, 79–86.
Lindley, D.V. (1957). A statistical paradox. *Biometrika*, **44**, 187–192.
Lindley, D.V. (1972). *Bayesian Statistics, a Review*. Philadelphia, PA: SIAM.
Matthews, R.A.J. (2001). Why *should* clinicians care about Bayesian methods? *J. Statist. Planning and Inference*, **94**, 43–71 (with discussion).
Rao, C.R. (1966). Covariance adjustment and related problems in multivariate analysis. In *Multivariate Analysis*, Ed. P.E. Krishnaiah, pp. 87–103. New York: Academic Press.
Rao, C.R. (1973). *Linear Statistical Inference and its Applications*. New York: Wiley
Robert, C.P. (1993). A note on Jeffreys–Lindley paradox. *Statistica Sinica*, **3**, 603–608.
Robert, C.P. (1996). Intrinsic Losses. *Theory and Decision*, **40**, 191–214.
Rueda, R. (1992). A Bayesian alternative to parametric hypothesis testing. *Test*, **1**, 61–67.
Shafer, G. (1982). Lindley's paradox. *J. Amer. Statist. Assoc.*, **77**, 325–351 (with discussion).

Résumé

Pour un modèle probabiliste $M \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega}), \boldsymbol{\theta} \in \Theta, \boldsymbol{\omega} \in \Omega\}$ censé décrire le comportement probabiliste de données $\boldsymbol{x} \in X$, nous soutenons que tester si les données sont compatibles avec une hypothèse $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ doit être considéré comme un problème décisionnel concernant l'usage du modèle $M_0 \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, avec une fonction de coût qui mesure la quantité d'information qui peut être perdue si le modèle simplifié M_0 est utilisé comme approximation du véritable modèle M. Le coût moyen, calculé par rapport à une loi a priori de référence idoine fournit une statistique de test pertinente,

la statistique intrinsèque $d(\theta_0, \boldsymbol{x})$, invariante par reparamétrisation. La statistique intrinsèque $d(\theta_0, \boldsymbol{x})$ est mesurée en unités d'information, et sa calibrage, qui est independante de la taille de l'échantillon et de la dimension du paramètre, ne dépend pas de sa distribution à l'échantillonage. La règle de Bayes correspondante, le critère de Bayes de référence (BRC), indique que H_0 doit seulement être rejeté si le coût a posteriori moyen de la perte d'information à utiliser le modèle simplifié M_0 est trop grande. Le critère BRC fournit une solution bayésienne générale et objective pour les tests d'hypothèses précises qui ne réclame pas une masse de Dirac concentrée sur M_0. Par conséquent, elle échappe au paradoxe de Lindley. Cette théorie est illustrée dans le contexte de variables normales multivariées, et on montre qu'elle évite le paradoxe de Rao sur l'inconsistence existant entre tests univariés et multivariés.

[*Received June 2001, accepted August 2002*]