Reference Analysis

José M. Bernardo [1]

Departamento de Estadística e I.0., Universitat de València, Spain

Abstract

This chapter describes reference analysis, a method to produce Bayesian inferential statements which only depend on the assumed model and the available data. Statistical information theory is used to define the *reference* prior function as a mathematical description of that situation where data would best dominate prior knowledge about the quantity of interest. Reference priors are not descriptions of personal beliefs; they are proposed as formal *consensus* prior functions to be used as standards for scientific communication. Reference posteriors are obtained by formal use of Bayes theorem with a reference prior. Reference prediction is achieved by integration with a reference posterior. Reference decisions are derived by minimizing a reference posterior expected loss. An information theory based loss function, the *intrinsic discrepancy*, may be used to derive reference procedures for conventional inference problems in scientific investigation, such as point estimation, region estimation and hypothesis testing.

Key words: Amount of information, Intrinsic discrepancy, Bayesian asymptotics, Fisher information, Objective priors, Noninformative priors, Jeffreys priors, Reference priors, Maximum entropy, Consensus priors, Intrinsic statistic, Point Estimation, Region Estimation, Hypothesis testing,

1 Introduction and notation

This chapter is mainly concerned with statistical inference problems such as occur in scientific investigation. Those problems are typically solved conditional on the assumption that a particular statistical model is an appropriate description of the probabilistic mechanism which has generated the data, and the choice of that model naturally involves an element of subjectivity. It has become standard practice, however, to describe as "objective" any statistical

Email address: jose.m.bernardo@uv.es (José M. Bernardo).
URL: www.uv.es/~bernardo (José M. Bernardo).

[1] Supported by grant BMF2001-2889 of the MCyT, Madrid, Spain

analysis which only depends on the model assumed and the data observed. In this precise sense (and only in this sense) reference analysis is a method to produce "objective" Bayesian inference.

Foundational arguments (Savage, 1954; de Finetti, 1970; Bernardo and Smith, 1994) dictate that scientists should elicit a unique (joint) prior distribution on all unknown elements of the problem on the basis of available information, and use Bayes theorem to combine this with the information provided by the data, encapsulated in the likelihood function. Unfortunately however, this elicitation is a formidable task, specially in realistic models with many nuisance parameters which rarely have a simple interpretation. Weakly informative priors have here a role to play as approximations to genuine proper prior distributions. In this context, the (unfortunately very frequent) naïve use of simple proper "flat" priors (often a limiting form of a conjugate family) as presumed "noninformative" priors often hides important unwarranted assumptions which may easily dominate, or even invalidate, the analysis: see *e.g.*, Hobert and Casella (1996, 1998), Casella (1996), Palmer and Pettit (1996), Hadjicostas and Berry (1999) or Berger (2000). The uncritical (ab)use of such "flat" priors should be strongly discouraged. An appropriate *reference* prior (see below) should instead be used. With numerical simulation techniques, where a proper prior is often needed, a proper approximation to the *reference* prior may be employed.

Prior elicitation would be even harder in the important case of scientific inference, where some sort of *consensus* on the elicited prior would obviously be required. A fairly natural candidate for such a consensus prior would be a "noninformative" prior, where prior knowledge could be argued to be dominated by the information provided by the data. Indeed, scientific investigation is seldom undertaken unless it is likely to substantially increase knowledge and, even if the scientist holds strong prior beliefs, the analysis would be most convincing to the scientific community if done with a consensus prior which is dominated by the data. Notice that the concept of a "noninformative" prior is *relative* to the information provided by the data.

As evidenced by the long list of references which concludes this chapter, there has been a considerable body of conceptual and theoretical literature devoted to identifying appropriate procedures for the formulation of "noninformative" priors. Beginning with the work of Bayes (1763) and Laplace (1825) under the name of inverse probability, the use of "noninformative" priors became central to the early statistical literature, which at that time was mainly objective Bayesian. The obvious limitations of the principle of insufficient reason used to justify the (by then) ubiquitous uniform priors, motivated the developments of Fisher and Neyman, which overshadowed Bayesian statistics during the first half of the 20th century. The work of Jeffreys (1946) prompted a strong revival of objective Bayesian statistics; the seminal books by Jeffreys (1961), Lindley (1965), Zellner (1971), Press (1972) and Box and Tiao (1973), demonstrated that the conventional textbook problems which frequentist statistics were able

to handle could better be solved from a unifying objective Bayesian perspective. Gradual realization of the fact that no *single* "noninformative" prior could possibly be always appropriate for all inference problems within a given multiparameter model (Dawid, Stone and Zidek, 1973; Efron, 1986) suggested that the long search for a *unique* "noninformative" prior representing "ignorance" within a given model was misguided. Instead, efforts concentrated in identifying, for each particular inference problem, a specific (joint) *reference prior* on all the unknown elements of the problem which would lead to a (marginal) *reference posterior* for the quantity of interest, a posterior which would always be dominated by the information provided by the data (Bernardo, 1979b). As will later be described in detail, statistical information theory was used to provide a precise meaning to this dominance requirement.

Notice that reference priors were *not* proposed as an approximation to the scientist's (unique) personal beliefs, but as a collection of formal *consensus* (not necessarily proper) prior functions which could conveniently be used as standards for scientific communication. As Box and Tiao (1973, p. 23) required, using a reference prior the scientist employs the jury principle; as the jury is carefully screened among people with no connection with the case, so that testimony may be assumed to dominate prior ideas of the members of the jury, the reference prior is carefully chosen to guarantee that the information provided by the data will not be overshadowed by the scientist's prior beliefs.

Reference posteriors are obtained by formal use of Bayes theorem with a reference prior function. If required, they may be used to provide point or region estimates, to test hypothesis, or to predict the value of future observations. This provides a unified set of objective Bayesian solutions to the conventional problems of scientific inference, objective in the precise sense that those solutions only depend on the assumed model and the observed data.

By restricting the class $\mathcal{P}$ of candidate priors, the reference algorithm makes it possible to incorporate into the analysis any genuine prior knowledge (over which scientific consensus will presumably exist). From this point of view, derivation of reference priors may be described as a new, powerful method for *prior elicitation*. Moreover, when subjective prior information is actually specified, the corresponding subjective posterior may be compared with the *reference* posterior—hence its name—to assess the relative importance of the initial opinions in the final inference.

In this chapter, it is assumed that probability distributions may be described through their probability density functions, and no notational distinction is made between a random quantity and the particular values that it may take. Bold italic roman fonts are used for observable random vectors (typically data) and bold italic greek fonts for unobservable random vectors (typically parameters); lower case is used for variables and upper case calligraphic for their dominion sets. Moreover, the standard mathematical convention of referring to functions, say $f_{\boldsymbol{x}}$ and $g_{\boldsymbol{x}}$ of $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, respectively by $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ will be

used throughout. Thus, the conditional probability density of data $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ given $\boldsymbol{\theta}$ will be represented by either $p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}$ or $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$, with $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}) \geq 0$ and $\int_{\boldsymbol{\mathcal{X}}} p(\boldsymbol{x}\,|\,\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x} = 1$, and the posterior distribution of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ given $\boldsymbol{x}$ will be represented by either $p_{\boldsymbol{\theta}\,|\,\boldsymbol{x}}$ or $p(\boldsymbol{\theta}\,|\,\boldsymbol{x})$, with $p(\boldsymbol{\theta}\,|\,\boldsymbol{x}) \geq 0$ and $\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}\,|\,\boldsymbol{x})\,\mathrm{d}\boldsymbol{\theta} = 1$. This admittedly imprecise notation will greatly simplify the exposition. If the random vectors are discrete, these functions naturally become probability mass functions, and integrals over their values become sums. Density functions of specific distributions are denoted by appropriate names. Thus, if x is an observable random variable with a normal distribution of mean μ and variance σ^2, its probability density function will be denoted $\mathrm{N}(x\,|\,\mu, \sigma)$. If the posterior distribution of μ is Student with location $\overline{x}$, scale s, and $n-1$ degrees of freedom, its probability density function will be denoted $\mathrm{St}(\mu\,|\,\overline{x},\ s,\ n-1)$.

The reference analysis argument is always defined in terms of some *parametric model* of the general form $\mathcal{M} \equiv \{p(\boldsymbol{x}\,|\,\boldsymbol{\omega}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, which describes the conditions under which data have been generated. Thus, data $\boldsymbol{x}$ are assumed to consist of one observation of the random vector $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, with probability density $p(\boldsymbol{x}\,|\,\boldsymbol{\omega})$ for some $\boldsymbol{\omega} \in \boldsymbol{\Omega}$. Often, but not necessarily, data will consist of a random sample $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ of fixed size n from some distribution with, say, density $p(\boldsymbol{y}\,|\,\boldsymbol{\omega})$, $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$, in which case $p(\boldsymbol{x}\,|\,\boldsymbol{\omega}) = \prod_{j=1}^{n} p(\boldsymbol{y}_j\,|\,\boldsymbol{\omega})$ and $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{Y}}^n$. In this case, reference priors relative to model $\mathcal{M}$ turn out to be the same as those relative to the simpler model $\mathcal{M}_{\boldsymbol{y}} \equiv \{p(\boldsymbol{y}\,|\,\boldsymbol{\omega}),\ \boldsymbol{y} \in \boldsymbol{\mathcal{Y}},\ \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$.

Let $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \boldsymbol{\Theta}$ be some vector of interest; without loss of generality, the assumed model $\mathcal{M}$ may be reparametrized in the form

$$\mathcal{M} \equiv \{\, p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\,\}, \tag{1}$$

where $\boldsymbol{\lambda}$ is some vector of nuisance parameters; this is often simply referred to as "model" $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})$. Conditional on the assumed model, all valid Bayesian inferential statements about the value of $\boldsymbol{\theta}$ are encapsulated in its posterior distribution $p(\boldsymbol{\theta}\,|\,\boldsymbol{x}) \propto \int_{\boldsymbol{\Lambda}}\ p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})\ p(\boldsymbol{\theta}, \boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda}$, which *combines* the information provided by the data $\boldsymbol{x}$ with any other information about $\boldsymbol{\theta}$ contained in the prior density $p(\boldsymbol{\theta}, \boldsymbol{\lambda})$. Intuitively, the *reference prior function* for $\boldsymbol{\theta}$, given model $\mathcal{M}$ and a class of candidate priors $\mathcal{P}$, is that (joint) prior $\pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda}\,|\,\mathcal{M}, \mathcal{P})$ which may be expected to have a minimal effect on the posterior inference about the quantity of interest $\boldsymbol{\theta}$ among the class of priors which belong to $\mathcal{P}$, *relative* to data which could be obtained from $\mathcal{M}$. The reference prior $\pi^{\boldsymbol{\theta}}(\boldsymbol{\omega}\,|\,\mathcal{M}, \mathcal{P})$ is specifically designed to be a reasonable *consensus* prior (within the class $\mathcal{P}$ of priors compatible with assumed prior knowledge) for inferences about a *particular quantity of interest* $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega})$, and it is always conditional to the *specific experimental design* $\mathcal{M} \equiv \{p(\boldsymbol{x}\,|\,\boldsymbol{\omega}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ which is assumed to have generated the data.

By definition, the reference prior $\pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda}\,|\,\mathcal{M}, \mathcal{P})$ is "objective", in the sense that it is a well-defined mathematical function of the vector of interest $\boldsymbol{\theta}$, the assumed model $\mathcal{M}$, and the class $\mathcal{P}$ of candidate priors, with no additional subjective elements. By formal use of Bayes theorem and appropriate integ-

ration (provided the integral is finite), the (joint) reference prior produces a (marginal) *reference posterior* for the vector of interest

$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}, \mathcal{M}, \mathcal{P}) \propto \int_{\boldsymbol{\Lambda}} p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})\, \pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \mathcal{M}, \mathcal{P})\, \mathrm{d}\boldsymbol{\lambda}, \tag{2}$$

which could be described as a mathematical expression of the inferential content of data $\boldsymbol{x}$ with respect to the value of $\boldsymbol{\theta}$, with no additional knowledge beyond that contained in the assumed statistical model $\mathcal{M}$ and the class $\mathcal{P}$ of candidate priors (which may well consist of the class $\mathcal{P}_0$ of *all* suitably regular priors). To simplify the exposition, the dependence of the reference prior on both the model and the class of candidate priors is frequently dropped from the notation, so that $\pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ and $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ are written instead of $\pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \mathcal{M}, \mathcal{P})$ and $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}, \mathcal{M}, \mathcal{P})$.

The reference prior function $\pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ often turns out to be an *improper* prior, *i.e.*, a positive function such that $\int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda})\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{\lambda}$ diverges and, hence, cannot be renormalized into a proper density function. Notice that this is not a problem provided the resulting posterior distribution (2) is proper for all suitable data. Indeed the declared objective of reference analysis is to provide appropriate reference *posterior* distributions; reference prior *functions* are merely useful technical devices for a simple computation (via formal use of Bayes theorem) of reference posterior *distributions*. For discussions on the axiomatic foundations which justify the use of improper prior functions, see Hartigan (1983) and references therein.

In the long quest for objective posterior distributions, several requirements have emerged which may reasonably be requested as *necessary* properties of any proposed solution:

(1) *Generality*. The procedure should be completely general, *i.e.*, applicable to any properly defined inference problem, and should produce no untenable answers which could be used as counterexamples. In particular, an objective posterior $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ must be a *proper* probability distribution for any data set $\boldsymbol{x}$ large enough to identify the unknown parameters.

(2) *Invariance*. Jeffreys (1946), Hartigan (1964), Jaynes (1968), Box and Tiao (1973, Sec. 1.3), Villegas (1977b, 1990), Dawid (1983), Yang (1995), Datta and J. K. Ghosh (1995b), Datta and M. Ghosh (1996). For any one-to-one function $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$, the posterior $\pi(\boldsymbol{\phi} \,|\, \boldsymbol{x})$ obtained from the reparametrized model $p(\boldsymbol{x} \,|\, \boldsymbol{\phi}, \boldsymbol{\lambda})$ must be coherent with the posterior $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ obtained from the original model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ in the sense that, for any data set $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, $\pi(\boldsymbol{\phi} \,|\, \boldsymbol{x}) = \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) |\, \mathrm{d}\boldsymbol{\theta} / \mathrm{d}\boldsymbol{\phi}|$. Moreover, if the model has a sufficient statistic $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$, then the posterior $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ obtained from the full model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ must be the same as the posterior $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{t})$ obtained from the equivalent model $p(\boldsymbol{t} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$.

(3) *Consistent marginalization.* Stone and Dawid (1972), Dawid, Stone and Zidek (1973), Dawid (1980). If, for all data $\boldsymbol{x}$, the posterior $\pi_1(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ obtained from model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ is of the form $\pi_1(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \pi_1(\boldsymbol{\theta} \,|\, \boldsymbol{t})$ for some statistic $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$ whose sampling distribution $p(\boldsymbol{t} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}) = p(\boldsymbol{t} \,|\, \boldsymbol{\theta})$ only depends on $\boldsymbol{\theta}$, then the posterior $\pi_2(\boldsymbol{\theta} \,|\, \boldsymbol{t})$ obtained from the marginal model $p(\boldsymbol{t} \,|\, \boldsymbol{\theta})$ must be the same as the posterior $\pi_1(\boldsymbol{\theta} \,|\, \boldsymbol{t})$ obtained from the original full model.

(4) *Consistent sampling properties.* Neyman and Scott (1948), Stein (1959), Dawid and Stone (1972, 1973), Cox and Hinkley (1974, Sec. 2.4.3), Stone (1976), Lane and Sudderth (1984). The properties under repeated sampling of the posterior distribution must be consistent with the model. In particular, the family of posterior distributions $\{\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}_j),\, \boldsymbol{x}_j \in \boldsymbol{\mathcal{X}}\}$ which could be obtained by repeated sampling from $p(\boldsymbol{x}_j \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})$ should concentrate on a region of $\boldsymbol{\Theta}$ which contains the true value of $\boldsymbol{\theta}$.

Reference analysis, introduced by Bernardo (1979b) and further developed by Berger and Bernardo (1989, 1992a,b,c), appears to be the only available method to derive objective posterior distributions which satisfy all these desiderata. This chapter describes the basic elements of reference analysis, states its main properties, and provides signposts to the huge related literature.

Section 2 summarizes some necessary concepts of discrepancy and convergence, which are based on information theory. Section 3 provides a formal definition of reference distributions, and describes their main properties. Section 4 describes an integrated approach to point estimation, region estimation, and hypothesis testing, which is derived from the joint use of reference analysis and an information-theory based loss function, the *intrinsic discrepancy*. Section 5 provides many additional references for further reading on reference analysis and related topics.

2 Intrinsic discrepancy and expected information

Intuitively, a reference prior for $\boldsymbol{\theta}$ is one which maximizes what it is *not known* about $\boldsymbol{\theta}$, *relative* to what *could* possibly be learnt from repeated observations from a particular model. More formally, a reference prior for $\boldsymbol{\theta}$ is defined to be one which maximizes—within some class of candidate priors—the *missing information* about the quantity of interest $\boldsymbol{\theta}$, defined as a limiting form of the amount of information about its value which repeated data from the assumed model could possibly provide. In this section, the notions of discrepancy, convergence, and expected information—which are required to make these ideas precise—are introduced and illustrated.

Probability theory makes frequent use of *divergence measures* between probability distributions. The total variation distance, Hellinger distance, Kullback-Leibler logarithmic divergence, and Jeffreys logarithmic divergence are fre-

quently cited; see, for example, Kullback (1968, 1983, 1987), Ibragimov and Khasminskii (1973), and Gutiérrez-Peña (1992) for precise definitions and properties. Each of those divergence measures may be used to define a type of convergence. It has been found, however, that the behaviour of many important limiting processes, in both probability theory and statistical inference, is better described in terms of another information-theory related divergence measure, the *intrinsic discrepancy* (Bernardo and Rueda, 2002), which is now defined and illustrated.

Definition 1 (Intrinsic discrepancy) *The* intrinsic discrepancy $\delta\{p_1, p_2\}$ *between two probability distributions of a random vector* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$*, specified by their density functions* $p_1(\boldsymbol{x})$*,* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_1 \subset \boldsymbol{\mathcal{X}}$*, and* $p_2(\boldsymbol{x})$*,* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_2 \subset \boldsymbol{\mathcal{X}}$*, with either identical or nested supports, is*

$$\delta\{p_1, p_2\} = \min\left\{ \int_{\boldsymbol{\mathcal{X}}_1} p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\, d\boldsymbol{x},\ \int_{\boldsymbol{\mathcal{X}}_2} p_2(\boldsymbol{x}) \log \frac{p_2(\boldsymbol{x})}{p_1(\boldsymbol{x})}\, d\boldsymbol{x} \right\}, \qquad (3)$$

provided one of the integrals (or sums) is finite. The intrinsic discrepancy between two parametric models for $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$*,* $\mathcal{M}_1 \equiv \{p_1(\boldsymbol{x} \,|\, \boldsymbol{\omega}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}_1,\ \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ *and* $\mathcal{M}_2 \equiv \{p_2(\boldsymbol{x} \,|\, \boldsymbol{\psi}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}_2,\ \boldsymbol{\psi} \in \boldsymbol{\Psi}\}$*, is the minimum intrinsic discrepancy between their elements,*

$$\delta\{\mathcal{M}_1, \mathcal{M}_2\} = \inf_{\boldsymbol{\omega} \in \boldsymbol{\Omega},\ \boldsymbol{\psi} \in \boldsymbol{\Psi}} \delta\{p_1(\boldsymbol{x} \,|\, \boldsymbol{\omega}),\ p_2(\boldsymbol{x} \,|\, \boldsymbol{\psi})\}. \qquad (4)$$

The *intrinsic discrepancy* is a new element of the class of *intrinsic loss functions* defined by Robert (1996); the concept is *not* related to the concepts of "intrinsic Bayes factors" and "intrinsic priors" introduced by Berger and Pericchi (1996), and reviewed in Pericchi (2005).

Notice that, as one would require, the intrinsic discrepancy $\delta\{\mathcal{M}_1, \mathcal{M}_2\}$ between two parametric families of distributions $\mathcal{M}_1$ and $\mathcal{M}_2$ does not depend on the particular parametrizations used to describe them. This will be crucial to guarantee the desired invariance properties of the statistical procedures described later.

It follows from Definition 1 that the intrinsic discrepancy between two probability distributions may be written in terms of their two possible Kullback-Leibler *directed divergences* as

$$\delta\{p_2, p_1\} = \min\left\{ \kappa\{p_2 \,|\, p_1\},\ \kappa\{p_1 \,|\, p_2\} \right\} \qquad (5)$$

where (Kullback and Leibler, 1951) the $\kappa\{p_j \,|\, p_i\}$'s are the non-negative invariant quantities defined by

$$\kappa\{p_j \,|\, p_i\} = \int_{\boldsymbol{\mathcal{X}}_i} p_i(\boldsymbol{x}) \log \frac{p_i(\boldsymbol{x})}{p_j(\boldsymbol{x})}\, \mathrm{d}\boldsymbol{x}, \quad \text{with } \ \boldsymbol{\mathcal{X}}_i \subseteq \boldsymbol{\mathcal{X}}_j. \qquad (6)$$

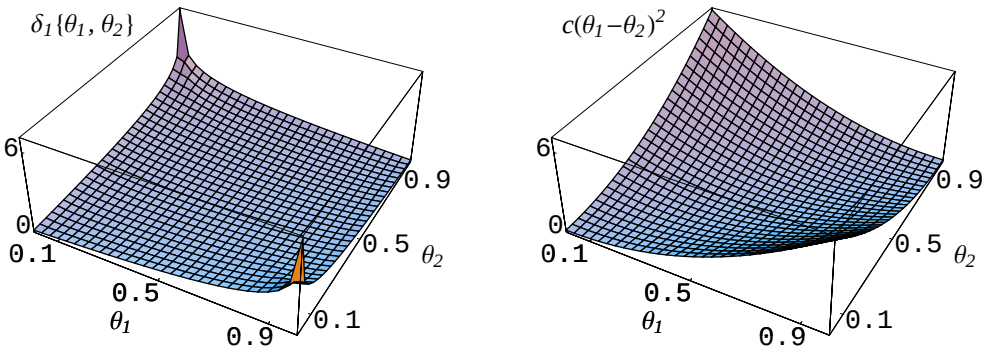
Since $\kappa\{p_j \,|\, p_i\}$ is the expected value of the logarithm of the density (or probability) ratio for p_i against p_j, when p_i is true, it also follows from Definition 1 that, if $\mathcal{M}_1$ and $\mathcal{M}_2$ describe two alternative models, one of which is assumed to generate the data, their intrinsic discrepancy $\delta\{\mathcal{M}_1, \mathcal{M}_2\}$ is the *minimum expected log-likelihood ratio in favour of the model which generates the data* (the "true" model). This will be important in the interpretation of many of the results described in this chapter.

The intrinsic discrepancy is obviously *symmetric*. It is non-negative, vanishes if (and only if) $p_1(\boldsymbol{x}) = p_2(\boldsymbol{x})$ almost everywhere, and it is invariant under one-to-one transformations of $\boldsymbol{x}$. Moreover, if $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ have strictly nested supports, one of the two directed divergences will not be finite, but their intrinsic discrepancy is still defined, and reduces to the other directed divergence. Thus, if $\boldsymbol{\mathcal{X}}_i \subset \boldsymbol{\mathcal{X}}_j$, then $\delta\{p_i, p_j\} = \delta\{p_j, p_i\} = \kappa\{p_j \,|\, p_i\}$.

The intrinsic discrepancy is *information additive*. Thus, if $\boldsymbol{x}$ consists of n independent observations, so that $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ and $p_i(\boldsymbol{x}) = \prod_{j=1}^{n} q_i(\boldsymbol{y}_j)$, then $\delta\{p_1, p_2\} = n\,\delta\{q_1, q_2\}$. This statistically important additive property is essentially unique to logarithmic discrepancies; it is basically a consequence of the fact that the joint density of independent random quantities is the product of their marginals, and the logarithm is the only analytic function which transforms products into sums.

Example 1 *Intrinsic discrepancy between Binomial distributions.* The intrinsic discrepancy $\delta\{\theta_1, \theta_2 \,|\, n\}$ between the two Binomial distributions

Figure 1 *Intrinsic discrepancy between Bernoulli variables.*



with common value for n, $p_1(r) = \mathrm{Bi}(r \,|\, n, \theta_1)$ and $p_2(r) = \mathrm{Bi}(r \,|\, n, \theta_2)$, is

$$\begin{aligned} \delta\{p_1, p_2\} &= \delta\{\theta_1, \theta_2 \,|\, n\} = n\,\delta_1\{\theta_1, \theta_2\}, \\ \delta_1\{\theta_1, \theta_2\} &= \min[\,\kappa\{\theta_1 \,|\, \theta_2\},\, \kappa\{\theta_2 \,|\, \theta_1\}\,] \\ \kappa(\theta_i \,|\, \theta_j) &= \theta_j \log[\theta_j/\theta_i] + (1-\theta_j)\log[(1-\theta_j)/(1-\theta_i)], \end{aligned} \tag{7}$$

where $\delta_1\{\theta_1, \theta_2\}$ (represented in the left panel of Figure 1) is the intrinsic discrepancy $\delta\{q_1, q_2\}$ between the corresponding Bernoulli distributions,

$q_i(y) = \theta_i^y(1-\theta_i)^{1-y}$, $y \in \{0, 1\}$. It may be appreciated that, specially near the extremes, the behaviour of the intrinsic discrepancy is rather different from that of the conventional quadratic loss $c\,(\theta_1 - \theta_2)^2$ (represented in the right panel of Figure 1 with c chosen to preserve the vertical scale).

As a direct consequence of the information-theoretical interpretation of the Kullback-Leibler directed divergences (Kullback, 1968, Ch. 1), the intrinsic discrepancy $\delta\{p_1, p_2\}$ is a measure, in natural information units or *nits* (Boulton and Wallace, 1970), of the *minimum* amount of expected information, in Shannon (1948) sense, required to discriminate between p_1 and p_2. If base 2 logarithms were used instead of natural logarithms, the intrinsic discrepancy would be measured in binary units of information (*bits*).

The quadratic loss $\ell\{\theta_1, \theta_2\} = (\theta_1 - \theta_2)^2$, often (over)used in statistical inference as measure of the discrepancy between two distributions $p(\boldsymbol{x} \,|\, \theta_1)$ and $p(\boldsymbol{x} \,|\, \theta_2)$ of the same parametric family $\{p(\boldsymbol{x} \,|\, \theta), \theta \in \Theta\}$, heavily depends on the parametrization chosen. As a consequence, the corresponding point estimate, the posterior expectation is not coherent under one-to-one transformations of the parameter. For instance, under quadratic loss, the "best" estimate of the logarithm of some positive physical magnitude is *not* the logarithm of the "best" estimate of such magnitude, a situation hardly acceptable by the scientific community. In sharp contrast to conventional loss functions, the intrinsic discrepancy is invariant under one-to-one reparametrizations. Some important consequences of this fact are summarized below.

Let $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be a family of probability densities, with no nuisance parameters, and let $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ be a possible point estimate of the quantity of interest $\boldsymbol{\theta}$. The intrinsic discrepancy $\delta\{\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\} = \delta\{p_{\boldsymbol{x} \,|\, \tilde{\boldsymbol{\theta}}},\, p_{\boldsymbol{x} \,|\, \boldsymbol{\theta}}\}$ between the estimated model and the true model measures, as a function of $\boldsymbol{\theta}$, the loss which would be suffered if model $p(\boldsymbol{x} \,|\, \tilde{\boldsymbol{\theta}})$ were used as a proxy for model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$. Notice that this directly measures how different the two *models* are, as opposed to measuring how different their *labels* are, which is what conventional loss functions—like the quadratic loss—typically do. As a consequence, the resulting discrepancy measure is independent of the particular parametrization used; indeed, $\delta\{\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\}$ provides a natural, *invariant* loss function for estimation, the *intrinsic loss*. The *intrinsic estimate* is that value $\boldsymbol{\theta}^*$ which minimizes $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}) = \int_{\boldsymbol{\Theta}} \delta\{\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\}\, p(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta}$, the posterior expected intrinsic loss, among all $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$. Since $\delta\{\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\}$ is invariant under reparametrization, the intrinsic estimate of any one-to-one transformation of $\boldsymbol{\theta}$, $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$, is simply $\boldsymbol{\phi}^* = \boldsymbol{\phi}(\boldsymbol{\theta}^*)$ (Bernardo and Juárez, 2003).

The posterior expected loss function $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x})$ may further be used to define posterior *intrinsic p-credible regions* $R_p = \{\tilde{\boldsymbol{\theta}};\ d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}) < d_p^*\}$, where d_p^* is chosen such that $\Pr[\boldsymbol{\theta} \in R_p \,|\, \boldsymbol{x}] = p$. In contrast to conventional highest posterior density (HPD) credible regions, which do *not* remain HPD under one-to-one transformations of $\boldsymbol{\theta}$, these *lowest posterior loss* (LPL) credible regions *remain* LPL under those transformations.

Similarly, if $\boldsymbol{\theta}_0$ is a parameter value of special interest, the intrinsic discrepancy $\delta\{\boldsymbol{\theta}_0, \boldsymbol{\theta}\} = \delta\{p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}_0}, p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}\}$ provides, as a function of $\boldsymbol{\theta}$, a measure of how far the particular density $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_0)$ (often referred to as the *null* model) is from the assumed model $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$, suggesting a natural invariant loss function for precise hypothesis testing. The null model $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_0)$ will be rejected if the corresponding posterior expected loss (called the *intrinsic statistic*) $d(\boldsymbol{\theta}_0\,|\,\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} \delta\{\boldsymbol{\theta}_0, \boldsymbol{\theta}\}\, p(\boldsymbol{\theta}\,|\,\boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta}$, is too large. As one should surely require, for any one-to-one transformation $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$, testing whether of not data are compatible with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ yields precisely the same result as testing $\boldsymbol{\phi} = \boldsymbol{\phi}_0 = \boldsymbol{\phi}(\boldsymbol{\theta}_0)$ (Bernardo and Rueda, 2002).

These ideas, extended to include the possible presence of nuisance parameters, will be further analyzed in Section 4.

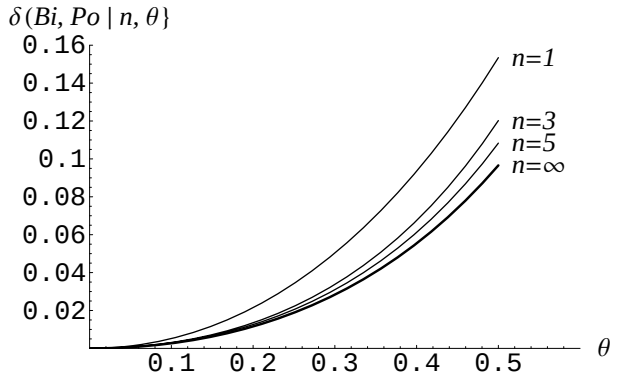
Definition 2 (Intrinsic convergence) *A sequence of probability distributions specified by their density functions* $\{p_i(\boldsymbol{x})\}_{i=1}^{\infty}$ *is said to* converge intrinsically *to a probability distribution with density* $p(\boldsymbol{x})$ *whenever the sequence of their intrinsic discrepancies* $\{\delta(p_i,\, p)\}_{i=1}^{\infty}$ *converges to zero.*

Example 2 *Poisson approximation to a Binomial distribution.* The intrinsic discrepancy between a Binomial distribution with probability function $\mathrm{Bi}(r\,|\,n, \theta)$ and its Poisson approximation $\mathrm{Po}(r\,|\,n\theta)$, is

$$\delta\{\mathrm{Bi}, \mathrm{Po}\,|\,n, \theta\} = \sum_{r=0}^{n} \mathrm{Bi}(r\,|\,n,\, \theta) \log \frac{\mathrm{Bi}(r\,|\,n,\, \theta)}{\mathrm{Po}(r\,|\,n\theta)} \,,$$

since the second sum in Definition 1 diverges. It may easily be verified that $\lim_{n\to\infty} \delta\{\mathrm{Bi}, \mathrm{Po}\,|\,n, \lambda/n\} = 0$ and $\lim_{\theta\to 0} \delta\{\mathrm{Bi}, \mathrm{Po}\,|\,\lambda/\theta,\, \theta\} = 0$; thus, as one would expect from standard probability theory, the sequences of Binomials $\mathrm{Bi}(r\,|\,n, \lambda/n)$ and $\mathrm{Bi}(r\,|\,\lambda/\theta_i,\, \theta_i)$ both intrinsically converge to a Poisson $\mathrm{Po}(r\,|\,\lambda)$ when $n \to \infty$ and $\theta_i \to 0$, respectively.

Figure 2 *Intrinsic discrepancy* $\delta\{\mathrm{Bi}, \mathrm{Po}\,|\,n, \theta\}$ *between a Binomial* $\mathrm{Bi}(r\,|\,n, \theta)$ *and a Poisson* $\mathrm{Po}(r\,|\,n\theta)$ *as a function of* θ*, for* $n = 1, 3, 5$ *and* ∞.



However, if one is interest in approximatiing a Binomial $\mathrm{Bi}(r\,|\,n, \theta)$ by a

Poisson $\mathrm{Po}(r \,|\, n\theta)$ the rôles of n and θ are far from similar: the important condition for the Poisson approximation to the Binomial to work is that the value of θ must be small, while the value of n is largely irrelevant. Indeed, (see Figure 2), $\lim_{\theta \to 0} \delta\{\mathrm{Bi}, \mathrm{Po} \,|\, n, \theta\} = 0$, for all $n > 0$, but $\lim_{n \to \infty} \delta\{\mathrm{Bi}, \mathrm{Po} \,|\, n, \theta\} = \frac{1}{2}[-\theta - \log(1-\theta)]$ for all $\theta > 0$. Thus, arbitrarily good approximations are possible with any n, provided θ is sufficiently small. However, for fixed θ, the quality of the approximation cannot improve over a certain limit, no matter how large n might be. For example, $\delta\{\mathrm{Bi}, \mathrm{Po} \,|\, 3, 0.05\} = 0.00074$ and $\delta\{\mathrm{Bi}, \mathrm{Po} \,|\, 5000, 0.05\} = 0.00065$, both yielding an expected log-probability ratio of about 0.0007. Thus, for all $n \geq 3$ the Binomial distribution $\mathrm{Bi}(r \,|\, n, 0.05)$ is quite well approximated by the Poisson distribution $\mathrm{Po}(r \,|\, 0.05n)$, and the quality of the approximation is very much the same for any value n.

Many standard approximations in probability theory may benefit from an analysis similar to that of Example 2. For instance, the sequence of Student distributions $\{\mathrm{St}(x \,|\, \mu, \sigma, \nu)\}_{\nu=1}^{\infty}$ converges intrinsically to the normal distribution $\mathrm{N}(x \,|\, \mu, \sigma)$ with the same location and scale parameters, and the discrepancy $\delta(\nu) = \delta\{\mathrm{St}(x \,|\, \mu, \sigma, \nu), \mathrm{N}(x \,|\, \mu, \sigma)\}$ (which only depends on the degrees of freedom ν) is smaller than 0.001 when $\nu > 40$. Thus approximating a Student with more than 40 degrees of freedom by a normal yields an expected log-density ratio smaller than 0.001, suggesting quite a good approximation.

As mentioned before, a reference prior is often an improper prior function. Justification of its use as a *formal* prior in Bayes theorem to obtain a reference posterior necessitates proving that the reference posterior thus obtained is an appropriate limit of a sequence of posteriors obtained from proper priors.

Theorem 1 *Consider a model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$. *If* $\pi(\boldsymbol{\omega})$ *is a strictly positive improper prior,* $\{\boldsymbol{\Omega}_i\}_{i=1}^{\infty}$ *is an increasing sequence of subsets of the parameter space which converges to* $\boldsymbol{\Omega}$ *and such that* $\int_{\boldsymbol{\Omega}_i} \pi(\boldsymbol{\omega})\, d\boldsymbol{\omega} < \infty$, *and* $\pi_i(\boldsymbol{\omega})$ *is the renormalized proper density obtained by restricting* $\pi(\boldsymbol{\omega})$ *to* $\boldsymbol{\Omega}_i$, *then, for any data set* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, *the sequence of the corresponding posteriors* $\{\pi_i(\boldsymbol{\omega} \,|\, \boldsymbol{x})\}_{i=1}^{\infty}$ *converges intrinsically to the posterior* $\pi(\boldsymbol{\omega} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, \pi(\boldsymbol{\omega})$ *obtained by formal use of Bayes theorem with the improper prior* $\pi(\boldsymbol{\omega})$.

However, to avoid possible pathologies, a stronger form of convergence is needed; for a sequence of proper priors $\{\pi_i\}_{i=1}^{\infty}$ to converge to a (possibly improper) prior function π, it will further be required that the *predicted* intrinsic discrepancy between the corresponding posteriors converges to zero. For a motivating example, see Berger and Bernardo (1992c, p. 43), where the model

$$\left\{ p(x \,|\, \theta) = \tfrac{1}{3}, \quad x \in \{[\tfrac{\theta}{2}], 2\theta, 2\theta + 1\}, \quad \theta \in \{1, 2, \ldots\} \right\},$$

where $[u]$ denotes the integer part of u (and $[\frac{1}{2}]$ is separately defined as 1), originally proposed by Fraser, Monette and Ng (1985), is reanalysed.

Definition 3 (Permissible prior function) *A positive function* $\pi(\boldsymbol{\omega})$ *is an* permissible *prior function for model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ *if for all* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ *one has* $\int_{\boldsymbol{\Omega}} p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, \pi(\boldsymbol{\omega})\, d\boldsymbol{\omega} < \infty$*, and for some increasing sequence* $\{\boldsymbol{\Omega}_i\}_{i=1}^{\infty}$ *of subsets of* $\boldsymbol{\Omega}$*, such that* $\lim_{i\to\infty} \boldsymbol{\Omega}_i = \boldsymbol{\Omega}$*, and* $\int_{\boldsymbol{\Omega}_i} \pi(\boldsymbol{\omega})\, d\boldsymbol{\omega} < \infty$*,*

$$\lim_{i\to\infty} \int_{\boldsymbol{\mathcal{X}}} p_i(\boldsymbol{x})\; \delta\{\pi_i(\boldsymbol{\omega} \,|\, \boldsymbol{x}),\, \pi(\boldsymbol{\omega} \,|\, \boldsymbol{x})\}\, d\boldsymbol{x} = 0,$$

where $\pi_i(\boldsymbol{\omega})$ *is the renormalized restriction of* $\pi(\boldsymbol{\omega})$ *to* $\boldsymbol{\Omega}_i$*,* $\pi_i(\boldsymbol{\omega} \,|\, \boldsymbol{x})$ *is the corresponding posterior,* $p_i(\boldsymbol{x}) = \int_{\boldsymbol{\Omega}_i} p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, \pi_i(\boldsymbol{\omega})\, d\boldsymbol{\omega}$ *is the corresponding predictive, and* $\pi(\boldsymbol{\omega} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, \pi(\boldsymbol{\omega})$*.*

In words, $\pi(\boldsymbol{\omega})$ is a permissible prior function for model $\mathcal{M}$ if it always yields proper posteriors, and the sequence of the *predicted* intrinsic discrepancies between the corresponding posterior $\pi(\boldsymbol{\omega} \,|\, \boldsymbol{x})$ and its renormalized restrictions to $\boldsymbol{\Omega}_i$ converges to zero for some suitable approximating sequence of the parameter space. All proper priors are permissible in the sense of Definition 3, but improper priors may or may not be permissible, even if they seem to be arbitrarily close to proper priors.

Example 3 *Exponential model.* Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from $p(x \,|\, \theta) = \theta e^{-\theta\, x}$, $\theta > 0$, so that $p(\boldsymbol{x} \,|\, \theta) = \theta^n e^{-\theta\, t}$, with sufficient statistic $t = \sum_{j=1}^{n} x_j$. Consider a positive function $\pi(\theta) \propto \theta^{-1}$, so that $\pi(\theta \,|\, t) \propto \theta^{n-1} e^{-\theta\, t}$, a gamma density $\mathrm{Ga}(\theta \,|\, n, t)$, which is a proper distribution for all possible data sets. Take now some sequence of pairs of positive real numbers $\{a_i, b_i\}$, with $a_i < b_i$, and let $\Theta_i = (a_i, b_i)$; the intrinsic discrepancy between $\pi(\theta \,|\, t)$ and its renormalized restriction to Θ_i, denoted $\pi_i(\theta \,|\, t)$, is $\delta_i(n, t) = \kappa\{\pi(\theta \,|\, t) \,|\, \pi_i(\theta \,|\, t)\} = \log[c_i(n, t)]$, where $c_i(n, t) = \Gamma(n)/\{\Gamma(n, a_i\, t) - \Gamma(n, b_i\, t)\}$. The renormalized restriction of $\pi(\theta)$ to Θ_i is $\pi_i(\theta) = \theta^{-1}/\log[b_i/a_i]$, and the corresponding (prior) predictive of t is $p_i(t \,|\, n) = c_i^{-1}(n, t)\; t^{-1}/\log[b_i/a_i]$. It may be verified that, for all $n \geq 1$, the expected intrinsic discrepancy $\int_0^{\infty} p_i(t \,|\, n)\, \delta_i(n, t)\, \mathrm{d}t$ converges to zero as $i \to \infty$. Hence, all positive functions of the form $\pi(\theta) \propto \theta^{-1}$ are permissible priors for the parameter of an exponential model.

Example 4 *Mixture model.* Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from $\mathcal{M} \equiv \{\frac{1}{2}\mathrm{N}(x \,|\, \theta, 1) + \frac{1}{2}\mathrm{N}(x \,|\, 0, 1),\, x \in \mathbb{R},\, \theta \in \mathbb{R}\}$. It is easily verified that the likelihood function $p(\boldsymbol{x} \,|\, \theta) = \prod_{j=1}^{n} p(x_j \,|\, \theta)$ is always bounded below by a strictly positive function of $\boldsymbol{x}$. Hence, $\int_{-\infty}^{\infty} p(\boldsymbol{x} \,|\, \theta)\, \mathrm{d}\theta = \infty$ for all $\boldsymbol{x}$, and the "natural" objective uniform prior function $\pi(\theta) = 1$ is obviously *not* permissible, although it may be pointwise arbitrarily well approximated by a sequence of proper "flat" priors.

Definition 4 (Intrinsic association) *The* intrinsic association $\alpha_{\boldsymbol{x}\boldsymbol{y}}$ *between two random vectors* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ *and* $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$ *with joint density* $p(\boldsymbol{x}, \boldsymbol{y})$ *and marginals* $p(\boldsymbol{x})$ *and* $p(\boldsymbol{y})$ *is the intrinsic discrepancy* $\alpha_{\boldsymbol{x}\boldsymbol{y}} = \delta\{p_{\boldsymbol{x}\,\boldsymbol{y}},\, p_{\boldsymbol{x}} p_{\boldsymbol{y}}\}$ *between their joint density and the product of their marginals. The intrinsic coefficient of association* $\rho^2_{\boldsymbol{x}\,\boldsymbol{y}} = 1 - \exp\{-2\alpha_{\boldsymbol{x}\,\boldsymbol{y}}\}$ *rescales the intrinsic association to* $[0, 1]$*.*

The intrinsic association is a non-negative invariant measure of association between two random vectors, which vanishes if they are independent, and tends to infinity as $\boldsymbol{y}$ and $\boldsymbol{x}$ approach a functional relationship. If their joint distribution is bivariate normal, then $\alpha_{\boldsymbol{x}\,\boldsymbol{y}} = -\frac{1}{2}\log(1-\rho^2)$, and $\rho^2_{\boldsymbol{x}\,\boldsymbol{y}} = \rho^2$, the square of their coefficient of correlation ρ.

The concept of intrinsic association extends that of *mutual information*; see *e.g.*, Cover and Thomas (1991), and references therein. Important differences arise in the context of contingency tables, where both $\boldsymbol{x}$ and $\boldsymbol{y}$ are discrete random variables which may only take a finite number of different values.

Definition 5 (Expected intrinsic information) *The* expected intrinsic information $I\{p_{\boldsymbol{\omega}} \,|\, \mathcal{M}\}$ *from one observation of* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ *about the value of* $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ *when the prior density is* $p(\boldsymbol{\omega})$*, is the intrinsic association* $\alpha_{\boldsymbol{x}\boldsymbol{\omega}} = \delta\{p_{\boldsymbol{x}\,\boldsymbol{\omega}},\; p_{\boldsymbol{x}}\, p_{\boldsymbol{\omega}}\}$ *between* $\boldsymbol{x}$ *and* $\boldsymbol{\omega}$*, where* $p(\boldsymbol{x}, \boldsymbol{\omega}) = p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, p(\boldsymbol{\omega})$*, and* $p(\boldsymbol{x}) = \int_{\boldsymbol{\Omega}} p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, p(\boldsymbol{\omega})\, d\boldsymbol{\omega}$.

For a fixed model $\mathcal{M}$, the expected intrinsic information $I\{p_{\boldsymbol{\omega}} \,|\, \mathcal{M}\}$ is a concave, positive functional of the prior $p(\boldsymbol{\omega})$. Under appropriate regularity conditions, in particular when data consists of a *large* random sample $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ from some model $\{p(\boldsymbol{y} \,|\, \boldsymbol{\omega}),\, \boldsymbol{y} \in \boldsymbol{\mathcal{Y}},\, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, one has

$$\int \int_{\boldsymbol{\mathcal{X}} \times \boldsymbol{\Omega}} [p(\boldsymbol{x})p(\boldsymbol{\omega}) + p(\boldsymbol{x}, \boldsymbol{\omega})] \, \log \frac{p(\boldsymbol{x})\, p(\boldsymbol{\omega})}{p(\boldsymbol{x}, \boldsymbol{\omega})} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{\omega} \geq 0 \tag{8}$$

so that $\kappa\{p_{\boldsymbol{x}}\, p_{\boldsymbol{\omega}} \,|\, p_{\boldsymbol{x}\,\boldsymbol{\omega}}\} \leq \kappa\{p_{\boldsymbol{x}\,\boldsymbol{\omega}} \,|\, p_{\boldsymbol{x}}\, p_{\boldsymbol{\omega}}\}$. If this is the case,

$$\begin{aligned} I\{p_{\boldsymbol{\omega}} \,|\, \mathcal{M}\} &= \delta\{p_{\boldsymbol{x}\,\boldsymbol{\omega}},\; p_{\boldsymbol{x}}\, p_{\boldsymbol{\omega}}\} = \kappa\{p_{\boldsymbol{x}}\, p_{\boldsymbol{\omega}} \,|\, p_{\boldsymbol{x}\,\boldsymbol{\omega}}\} \\ &= \int \int_{\boldsymbol{\mathcal{X}} \times \boldsymbol{\Omega}} p(\boldsymbol{x}, \boldsymbol{\omega}) \, \log \frac{p(\boldsymbol{x}, \boldsymbol{\omega})}{p(\boldsymbol{x})\, p(\boldsymbol{\omega})} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{\omega} \qquad (9) \\ &= \int_{\boldsymbol{\Omega}} p(\boldsymbol{\omega}) \int_{\boldsymbol{\mathcal{X}}} p(\boldsymbol{x} \,|\, \boldsymbol{\omega}) \, \log \frac{p(\boldsymbol{\omega} \,|\, \boldsymbol{x})}{p(\boldsymbol{\omega})} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{\omega} \qquad (10) \\ &= H[p_{\boldsymbol{\omega}}] - \int_{\boldsymbol{\mathcal{X}}} p(\boldsymbol{x}) \, H[p_{\boldsymbol{\omega} \,|\, \boldsymbol{x}}] \, \mathrm{d}\boldsymbol{x}, \qquad (11) \end{aligned}$$

where $H[p_{\boldsymbol{\omega}}] = -\int_{\boldsymbol{\Omega}} p(\boldsymbol{\omega}) \log p(\boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega}$ is the *entropy* of $p_{\boldsymbol{\omega}}$, and the expected intrinsic information reduces to the Shannon's expected information (Shannon, 1948; Lindley, 1956; Stone, 1959; de Waal and Groenewald, 1989; Clarke and Barron, 1990).

For any fixed model $\mathcal{M}$, the expected intrinsic information $I\{p_{\boldsymbol{\omega}} \,|\, \mathcal{M}\}$ measures, as a functional of the prior $p_{\boldsymbol{\omega}}$, the amount of information about the value of $\boldsymbol{\omega}$ which one observation $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ may be expected to provide. The stronger the prior knowledge described by $p_{\boldsymbol{\omega}}$, the smaller the information the data may be expected to provide; conversely, weak initial knowledge about $\boldsymbol{\omega}$ will correspond to large expected information from the data. This is the intuitive basis for the definition of a reference prior.

3 Reference distributions

Let $\boldsymbol{x}$ be one observation from model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, and let $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \boldsymbol{\Theta}$ be some vector of interest, whose posterior distribution is required. Notice that $\boldsymbol{x}$ represents the *complete* available data; often, but not always, this will consist of a random sample $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ of fixed size n from some simpler model. Let $\mathcal{P}$ be the *class of candidate priors* for $\boldsymbol{\omega}$, defined as those sufficiently regular priors which are compatible with whatever agreed "objective" initial information about the value of $\boldsymbol{\omega}$ one is willing to assume. A permissible prior function $\pi^{\boldsymbol{\theta}}(\boldsymbol{\omega} \,|\, \mathcal{M}, \mathcal{P})$ is desired which may be expected to have a minimal effect (in a sense to be made precise) among all priors in $\mathcal{P}$, on the posterior inferences about $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega})$ which could be derived given data generated from $\mathcal{M}$. This will be named a *reference prior function* of $\boldsymbol{\omega}$ for the quantity of interest $\boldsymbol{\theta}$, relative to model $\mathcal{M}$ and class $\mathcal{P}$ of candidate priors, and will be denoted by $\pi^{\boldsymbol{\theta}}(\boldsymbol{\omega} \,|\, \mathcal{M}, \mathcal{P})$. The reference prior function $\pi^{\boldsymbol{\theta}}(\boldsymbol{\omega} \,|\, \mathcal{M}, \mathcal{P})$ will then be used as a formal prior density to derive the required *reference posterior* distribution of the quantity of interest, $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}, \mathcal{M}, \mathcal{P})$, via Bayes theorem and the required probability operations.

This section contains the definition and basic properties of reference distributions. The ideas are first formalized in one-parameter models, and then extended to multiparameter situations. Special attention is devoted to *restricted* reference distributions, where the class of candidate priors $\mathcal{P}$ consists of those which satisfy some set of assumed conditions. This provides a continuous collection of solutions, ranging from situations with no assumed prior information on the quantity of interest, when $\mathcal{P}$ is the class $\mathcal{P}_0$ of *all* sufficiently regular priors, to situations where accepted prior knowledge is sufficient to specify a unique prior $p_0(\boldsymbol{\omega})$, so that $\pi^{\boldsymbol{\theta}}(\boldsymbol{\omega} \,|\, \mathcal{M}, \mathcal{P}) = p_0(\boldsymbol{\theta})$, the situation commonly assumed in Bayesian subjective analysis.

3.1 One parameter models

Let $\theta \in \Theta \subset \mathbb{R}$ be a real-valued quantity of interest, and let available data $\boldsymbol{x}$ consist of one observation from model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta\}$, so that there are no nuisance parameters. A permissible prior function $\pi(\theta) = \pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ in a class $\mathcal{P}$ is desired with a minimal expected effect on the posteriors of θ which could be obtained after data $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ generated from $\mathcal{M}$ have been observed.

Let $\boldsymbol{x}^{(k)} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ consist of k conditionally independent (given θ) observations from $\mathcal{M}$, so that $\boldsymbol{x}^{(k)}$ consists of one observation from the product model $\mathcal{M}^k = \{ \prod_{j=1}^k p(\boldsymbol{x}_j \,|\, \theta),\, \boldsymbol{x}_j \in \boldsymbol{\mathcal{X}},\, \theta \in \Theta \}$. Let p_θ be a prior distribution for the quantity of interest, and consider the intrinsic information about θ, $I\{p_\theta \,|\, \mathcal{M}^k\}$, which could be expected from the vector $\boldsymbol{x}^{(k)} \in \boldsymbol{\mathcal{X}}^k$. For any sufficiently regular prior p_θ, the posterior distribution of θ would concentrate on its true value as k increases and therefore, as $k \to \infty$, the true value of θ

would get to be precisely known. Thus, as $k \to \infty$, the functional $I\{p_\theta \,|\, \mathcal{M}^k\}$ will approach a precise measure of the amount of *missing information* about θ which corresponds to the prior p_θ. It is natural to define the reference prior as that prior function $\pi^\theta = \pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ which *maximizes the missing information* about the value of θ within the class $\mathcal{P}$ of candidate priors.

Under regularity conditions, the expected intrinsic information $I\{p_\theta \,|\, \mathcal{M}^k\}$ becomes, for large k, Shannon's expected information and hence, using (11),

$$I\{p_\theta \,|\, \mathcal{M}^k\} = H[p_\theta] - \int_{\mathcal{X}^k} p(\boldsymbol{x}^{(k)})\, H[p_{\theta \,|\, \boldsymbol{x}^{(k)}}]\, \mathrm{d}\boldsymbol{x}^{(k)}, \tag{12}$$

where $H[p_\theta] = -\int_\Theta p(\theta) \log p(\theta)\, \mathrm{d}\theta$, is the *entropy* of p_θ. It follows that, when the parameter space $\Theta = \{\theta_1, \ldots, \theta_m\}$ is finite, the missing information which corresponds to any strictly positive prior p_θ is, *for any model* $\mathcal{M}$,

$$\lim_{k \to \infty} I\{p_\theta \,|\, \mathcal{M}^k\} = H[p_\theta] = -\sum_{j=1}^m p(\theta_j) \log p(\theta_j), \tag{13}$$

since, as $k \to \infty$, the discrete posterior probability function $p(\theta \,|\, \boldsymbol{x}^{(k)})$ converges to a degenerate distribution with probability one on the true value of θ and zero on all others, and thus, the posterior entropy $H[p_{\theta \,|\, \boldsymbol{x}^{(k)}}]$ converges to zero. Hence, in finite parameter spaces, the reference prior for the parameter does not depend on the precise form of the model, and it is precisely that which *maximizes the entropy* within the class $\mathcal{P}$ of candidate priors. This was the solution proposed by Jaynes (1968), and it is often used in mathematical physics. In particular, if the class of candidate priors is the class $\mathcal{P}_0$ of *all* strictly positive probability distributions, the reference prior for θ is a uniform distribution over Θ, the "noninformative" prior suggested by the old insufficient reason argument (Laplace, 1812). For further information on the concept of *maximum entropy*, see Jaynes (1968, 1982, 1985, 1989), Akaike (1977), Csiszár (1985, 1991), Clarke and Barron (1994), Grünwald and Dawid (2004), and references therein.

In the continuous case, however, $I\{p_\theta \,|\, \mathcal{M}^k\}$ typically diverges as $k \to \infty$, since an infinite amount of information is required to know exactly the value of a real number. A general definition of the reference prior (which includes the finite case as a particular case), is nevertheless possible as an appropriate limit, when $k \to \infty$, of the sequence of priors maximizing $I\{p_\theta \,|\, \mathcal{M}^k\}$ within the class $\mathcal{P}$. Notice that this limiting procedure is *not* some kind of asymptotic approximation, but an essential element of the *concept* of a reference prior. Indeed, the reference prior is defined to maximize the *missing information* about the quantity of interest which *could* be obtained by repeated sampling from $\mathcal{M}$ (not just the information expected from a finite data set), and this is precisely achieved by maximizing the expected information from the arbitrarily large data set which could be obtained by unlimited repeated sampling from the assumed model.

Since $I\{p_\theta \,|\, \mathcal{M}^k\}$ is only defined for proper priors, and $I\{p_\theta \,|\, \mathcal{M}^k\}$ is not guaranteed to attain its maximum at a proper prior, the formal definition of a reference prior is stated as a limit, as $i \to \infty$, of the sequence of solutions obtained for restrictions $\{\Theta_i\}_{i=1}^{\infty}$ of the parameter space chosen to ensure that the maximum of $I\{p_\theta \,|\, \mathcal{M}^k\}$ is actually obtained at a proper prior. The definition below (Berger, Bernardo and Sun, 2005) generalizes those in Bernardo (1979b) and Berger and Bernardo (1992c), and addresses the problems described in Berger, Bernardo and Mendoza (1989).

Definition 6 (One-parameter reference priors) *Consider the one-parameter model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \theta \in \Theta \subset \mathbb{R}\}$*, and let* $\mathcal{P}$ *be a class of candidate priors for* θ*. The positive function* $\pi(\theta) = \pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ *is a* reference prior *for model* $\mathcal{M}$ *given* $\mathcal{P}$ *if it is a permissible prior function such that, for some increasing sequence* $\{\Theta_i\}_{i=1}^{\infty}$ *with* $\lim_{i\to\infty} \Theta_i = \Theta$ *and* $\int_{\Theta_i} \pi(\theta)\, d\theta < \infty$*,*

$$\lim_{k\to\infty} \{I\{\pi_i \,|\, \mathcal{M}^k\} - I\{p_i \,|\, \mathcal{M}^k\}\} \geq 0, \quad \textit{for all } \Theta_i, \textit{ for all } p \in \mathcal{P},$$

where $\pi_i(\theta)$ *and* $p_i(\theta)$ *are the renormalized restrictions of* $\pi(\theta)$ *and* $p(\theta)$ *to* Θ_i*.*

Notice that Definition 6 involves two rather different limiting processes. The limiting process of the Θ_i's towards the whole parameter space Θ is only required to guarantee the existence of the expected informations; this may often (but not always) be avoided if the parameter space is (realistically) chosen to be some finite interval $[a, b]$. On the other hand, the limiting process as $k \to \infty$ is an *essential* part of the definition. Indeed, the reference prior is *defined* as that prior function which maximizes the *missing* information, which is the expected discrepancy between prior knowledge and *perfect* knowledge; but perfect knowledge is only approached *asymptotically*, as $k \to \infty$.

Definition 6 implies that reference priors *only* depend on the *asymptotic behaviour* of the assumed model, a feature which greatly simplifies their actual derivation; to obtain a reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ for the parameter θ of model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \theta \in \Theta\}$, it is both necessary and sufficient to establish the asymptotic behaviour of its posterior distribution under (conceptual) repeated sampling from $\mathcal{M}$, that is the limiting form, as $k \to \infty$, of the posterior density (or probability function) $\pi(\theta \,|\, \boldsymbol{x}^{(k)}) = \pi(\theta \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$.

As one would hope, Definition 6 yields the maximum entropy result in the case where the parameter space is finite and the quantity of interest is the actual value of the parameter:

Theorem 2 (Reference priors with finite parameter space) *Consider a model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \theta \in \Theta\}$*, with a* finite *parameter space* $\Theta = \{\theta_1, \ldots, \theta_m\}$ *and such that, for all pairs* θ_i *and* θ_j*,* $\delta\{p_{x\,|\,\theta_i},\ p_{x\,|\,\theta_j}\} > 0$*, and let* $\mathcal{P}$ *be a class of probability distributions over* Θ*. Then the reference prior for the parameter* θ *is*

$$\pi^\theta(\theta \,|\, \mathcal{M}, \mathcal{P}) = \arg\max_{p_\theta \in \mathcal{P}} H\{p_\theta\},$$

where $p_\theta = \{p(\theta_1), p(\theta_2), \ldots, p(\theta_m)\}$ *and* $H\{p_\theta\} = -\sum_{j=1}^{m} p(\theta_j) \log p(\theta_j)$ *is the* entropy *of* p_θ*. In particular, if the class of candidate priors for* θ *is the set* $\mathcal{P}_0$ *of all strictly positive probability distributions over* Θ*, then the reference prior is the uniform distribution* $\pi^\theta(\theta \,|\, \mathcal{M}, \mathcal{P}_0) = \{1/m, \ldots, 1/m\}$.

Theorem 2 follows immediately from the fact that, if the intrinsic discrepancies $\delta\{p_{x\,|\,\theta_i},\ p_{x\,|\,\theta_j}\}$ are all positive (and hence the m models $p(x \,|\, \theta_i)$ are all distinguishable from each other), then the posterior distribution of θ asymptotically converges to a degenerate distribution with probability one on the true value of θ (see *e.g.*, Bernardo and Smith (1994, Sec. 5.3) and references therein). Such asymptotic posterior has zero entropy and thus, by Equation 12, the missing information about θ when the prior is p_θ does not depend on $\mathcal{M}$, and is simply given by the prior entropy, $H\{p_\theta\}$. □

Consider now a model $\mathcal{M}$ indexed by a continuous parameter $\theta \in \Theta \subset \mathbb{R}$. If the family of candidate priors consist of the class $\mathcal{P}_0$ of *all* continuous priors with support Θ, then the reference prior, $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$ may be obtained as the result of an explicit limit. This provides a relatively simple procedure to obtain reference priors in models with one continuous parameter. Moreover, this analytical procedure may easily be converted into a programmable algorithm for numerical derivation of reference distributions. The results may conveniently be described in terms of any *asymptotically sufficient* statistic, *i.e.*, a function $\boldsymbol{t}_k = \boldsymbol{t}_k(\boldsymbol{x}^{(k)})$ such that, for all θ and for all $\boldsymbol{x}^{(k)}$, $\lim_{k\to\infty}[p(\theta \,|\, \boldsymbol{x}^{(k)})/p(\theta \,|\, \boldsymbol{t}_k)] = 1$.

Obviously, the entire sample $\boldsymbol{x}^{(k)}$ is sufficient (and hence asymptotically sufficient), so there is no loss of generality in framing the results in terms of asymptotically sufficient statistics.

Theorem 3 (Explicit form of the reference prior) *Consider the model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \theta \in \Theta \subset \mathbb{R}\}$*, and let* $\mathcal{P}_0$ *be the class of all continuous priors with support* Θ*. Let* $\boldsymbol{x}^{(k)} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ *consist of* k *independent observations from* $\mathcal{M}$*, so that* $p(\boldsymbol{x}^{(k)} \,|\, \theta) = \prod_{j=1}^{k} p(\boldsymbol{x}_j \,|\, \theta)$*, and let* $\boldsymbol{t}_k = \boldsymbol{t}_k(\boldsymbol{x}^{(k)}) \in \boldsymbol{\mathcal{T}}$ *be any asymptotically sufficient statistic. Let* $h(\theta)$ *be a continuous strictly positive function such that, for sufficiently large* k*,* $\int_\Theta p(\boldsymbol{t}_k \,|\, \theta)\, h(\theta)\, d\theta < \infty$*, and define*

$$f_k(\theta) = \exp\left\{ \int_{\boldsymbol{\mathcal{T}}} p(\boldsymbol{t}_k \,|\, \theta)\ \log\left(\frac{p(\boldsymbol{t}_k \,|\, \theta)\, h(\theta)}{\int_\Theta p(\boldsymbol{t}_k \,|\, \theta)\, h(\theta)\, d\theta} \right) d\boldsymbol{t}_k \right\}, \quad \textit{and} \tag{14}$$

$$f(\theta) = \lim_{k\to\infty} \frac{f_k(\theta)}{f_k(\theta_0)}\ , \tag{15}$$

where θ_0 *is any interior point of* Θ*. If* $f(\theta)$ *is a permissible prior function then, for any* $c > 0$*,* $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0) = c\, f(\theta)$ *is a reference prior function.*

Intuitively, Theorem 3 states that the reference prior $\pi(\theta \,|\, \mathcal{M})$ relative to

model $\mathcal{M}$ only depends on the asymptotic behaviour of the model and that, with no additional information to restrict the class of candidate priors, it has (from Equation 14), the form

$$\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0) \propto \exp\Big\{ E_{\boldsymbol{t}_k \,|\, \theta} \Big[\log\, p(\theta \,|\, \boldsymbol{t}_k) \Big] \Big\}, \tag{16}$$

where $p(\theta \,|\, \boldsymbol{t}_k)$ is any asymptotic approximation to the posterior distribution of θ, and the expectation is taken with respect to the sampling distribution of the relevant asymptotically sufficient statistic $\boldsymbol{t}_k = \boldsymbol{t}_k(\boldsymbol{x}^{(k)})$. A heuristic derivation of Theorem 3 is provided below. For a precise statement of the regularity conditions and a formal proof, see Berger, Bernardo and Sun (2005).

Under fairly general regularity conditions, the intrinsic expected information reduces to Shannon's expected information when $k \to \infty$. Thus, starting from (10), the amount of information about θ to be expected from $\mathcal{M}^k$ when the prior is $p(\theta)$ may be rewritten as $I\{p_\theta \,|\, \mathcal{M}^k\} = \int_\Theta p(\theta)\, \log[h_k(\theta)/p(\theta)]\, \mathrm{d}\theta$, where $h_k(\theta) = \exp\{\int_{\boldsymbol{\mathcal{T}}} p(\boldsymbol{t}_k \,|\, \theta)\, \log p(\theta \,|\, \boldsymbol{t}_k)\, \mathrm{d}\boldsymbol{t}_k\}$. If $c_k = \int_\Theta h_k(\theta)\, \mathrm{d}\theta < \infty$, then $h_k(\theta)$ may be renormalized to get the proper density $h_k(\theta)/c_k$, and $I\{p_\theta \,|\, \mathcal{M}^k\}$ may be rewritten as

$$I\{p_\theta \,|\, \mathcal{M}^k\} = \log c_k - \int_\Theta p(\theta) \log \frac{p(\theta)}{h_k(\theta)/c_k}\, \mathrm{d}\theta. \tag{17}$$

But the integral in (17) is the Kullback-Leibler directed divergence of $h_k(\theta)/c_k$ from $p(\theta)$, which is non-negative, and it is zero iff $p(\theta) = h_k(\theta)/c_k$ almost everywhere. Thus, $I\{p_\theta \,|\, \mathcal{M}^k\}$ would be maximized by a prior $\pi_k(\theta)$ which satisfies the functional equation

$$\pi_k(\theta) \propto h_k(\theta) = \exp\Big\{ \int_{\boldsymbol{\mathcal{T}}} p(\boldsymbol{t}_k \,|\, \theta)\, \log \pi_k(\theta \,|\, \boldsymbol{t}_k)\, \mathrm{d}\boldsymbol{t}_k \Big\}, \tag{18}$$

where $\pi_k(\theta \,|\, \boldsymbol{t}_k) \propto p(\boldsymbol{t}_k \,|\, \theta)\, \pi_k(\theta)$ and, therefore, the reference prior should be a limiting form, as $k \to \infty$ of the sequence of proper priors given by (18). This only provides an implicit solution, since the posterior density $\pi_k(\theta \,|\, \boldsymbol{t}_k)$ in the right hand side of (18) obviously depends on the prior $\pi_k(\theta)$; however, as $k \to \infty$, the posterior $\pi_k(\theta \,|\, \boldsymbol{t}_k)$ will approach its asymptotic form which, under the assumed conditions, is independent of the prior. Thus, the posterior density in (18) may be replaced by the posterior $\pi^0(\theta \,|\, \boldsymbol{t}_k) \propto p(\boldsymbol{t}_k \,|\, \theta)\, h(\theta)$ which corresponds to any fixed prior, say $\pi^0(\theta) = h(\theta)$, to obtain an explicit expression for a sequence of priors,

$$\pi_k(\theta) \propto f_k(\theta) = \exp\Big\{ \int_{\boldsymbol{\mathcal{T}}} p(\boldsymbol{t}_k \,|\, \theta)\, \log \pi^0(\theta \,|\, \boldsymbol{t}_k)\, \mathrm{d}\boldsymbol{t}_k \Big\}, \tag{19}$$

whose limiting form will still maximize the missing information about θ. The preceding argument rests however on the assumption that (at least for sufficiently large k) the integrals in Θ of $f_k(\theta)$ are finite, but those integrals may

well diverge. The problem is solved by considering an increasing sequence $\{\Theta_i\}_{i=1}^{\infty}$ of subsets of Θ which converges to Θ and such that, for all i and sufficiently large k, $c_{ik} = \int_{\Theta_i} f_k(\theta)\,\mathrm{d}\theta < \infty$, so that the required integrals are finite. An appropriate limiting form of the double sequence $\pi_{ik}(\theta) = f_k(\theta)/c_{ik}$, $\theta \in \Theta_i$ will then approach the required reference prior.

Such a limiting form is easily established; indeed, let $\pi_{ik}(\theta \,|\, \boldsymbol{x})$, $\theta \in \Theta_i$ be the posterior which corresponds to $\pi_{ik}(\theta)$ and, for some interior point θ_0 of all the Θ_i's, consider the limit

$$\lim_{k\to\infty} \frac{\pi_{ik}(\theta \,|\, \boldsymbol{x})}{\pi_{ik}(\theta_0 \,|\, \boldsymbol{x})} = \lim_{k\to\infty} \frac{p(\boldsymbol{x} \,|\, \theta)\, f_k(\theta)}{p(\boldsymbol{x} \,|\, \theta_0)\, f_k(\theta_0)} \propto p(\boldsymbol{x} \,|\, \theta)\, f(\theta), \tag{20}$$

where $f(\theta) = \lim_{k\to\infty} f_k(\theta)/f_k(\theta_0)$, which does not depend on the initial function $h(\theta)$ (and therefore $h(\theta)$ may be chosen by mathematical convenience). It follows from (20) that, for any data $\boldsymbol{x}$, the sequence of posteriors $\pi_{ik}(\theta \,|\, \boldsymbol{x})$ which maximize the missing information will approach the posterior $\pi(\theta \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta)\, f(\theta)$ obtained by formal use of Bayes theorem, using $f(\theta)$ as the prior. This completes the heuristic justification of Theorem 3. □

3.2 Main properties

Reference priors enjoy many attractive properties, as stated below. For detailed proofs, see Bernardo and Smith (1994, Secs. 5.4 and 5.6).

In the frequently occurring situation where the available data consist of a random sample of fixed size n from some model $\mathcal{M}$ (so that the assumed model is $\mathcal{M}^n$), the reference prior relative to $\mathcal{M}^n$ is independent of n, and may simply be obtained as the reference prior relative to $\mathcal{M}$, assuming the later exists.

Theorem 4 (Independence of sample size) *If data $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ consists of a random sample of size n from model $\mathcal{M} \equiv \{p(\boldsymbol{y} \,|\, \theta), \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}, \theta \in \Theta\}$, with reference prior $\pi^\theta(\theta \,|\, \mathcal{M}, \mathcal{P})$ relative to the class of candidate priors $\mathcal{P}$, then, for any fixed sample size n, the reference prior for θ relative to $\mathcal{P}$ is $\pi^\theta(\theta \,|\, \mathcal{M}^n, \mathcal{P}) = \pi^\theta(\theta \,|\, \mathcal{M}, \mathcal{P})$.*

This follows from the additivity of the information measure. Indeed, for any sample size n and number of replicates k, $I\{p_\theta \,|\, \mathcal{M}^{nk}\} = n\, I\{p_\theta \,|\, \mathcal{M}^k\}$. □

Note, however, that Theorem 4 requires $\boldsymbol{x}$ to be a random sample from the assumed model. If the model entails dependence between the observations (as in time series, or in spatial models) the reference prior may well depend on the sample size; see, for example, Berger and Yang (1994), and Berger, de Oliveira and Sansó (2001).

The possible dependence of the reference prior on the sample size and, more generally, on the design of the experiment highlights the fact that a reference prior is *not* a description of (personal) prior beliefs, but a possible *consensus*

prior for a particular problem of scientific inference. Indeed, genuine prior beliefs about some quantity of interest should not depend on the design of the experiment performed to learn about its value (although they will typically influence the choice of the design), but a prior function to be used as a *consensus* prior to analyse the results of an experiment may be expected to depend on its design. Reference priors, which by definition maximize the missing information which repeated observations from a *particular* experiment could possibly provide, generally depend on the design of that experiment.

As one would hope, if the assumed model $\mathcal{M}$ has a sufficient statistic $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$, the reference prior relative to $\mathcal{M}$ is the same as the reference prior relative to the equivalent model derived from the sampling distribution of $\boldsymbol{t}$:

Theorem 5 (Compatibility with sufficient statistics) *Consider a model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \theta \in \Theta\}$ *with sufficient statistic* $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x}) \in \boldsymbol{\mathcal{T}}$, *and let* $\mathcal{M}_{\boldsymbol{t}} \equiv \{p(\boldsymbol{t} \,|\, \theta), \boldsymbol{t} \in \boldsymbol{\mathcal{T}},\ \theta \in \Theta\}$ *be the corresponding model in terms of* $\boldsymbol{t}$. *Then, for any class of candidate priors* $\mathcal{P}$, *the reference prior for* θ *relative to model* $\mathcal{M}$ *is* $\pi^{\theta}(\theta \,|\, \mathcal{M}, \mathcal{P}) = \pi^{\theta}(\theta \,|\, \mathcal{M}_{\boldsymbol{t}}, \mathcal{P})$.

Theorem 5 follows from the fact that the expected information is invariant under such transformation, so that, for all k, $I\{p_\theta \,|\, \mathcal{M}^k\} = I\{p_\theta \,|\, \mathcal{M}^k_{\boldsymbol{t}}\}$. □

When data consist of a random sample of fixed size from some model, and there exists a sufficient statistic of fixed dimensionality, Theorems 3, 4 and 5 may be combined for an easy, direct derivation of the reference prior, as illustrated below.

Example 5 *Exponential model, continued.* Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample of size n from an exponential distribution. By Theorem 4, to obtain the corresponding reference prior it suffices to analyse the behaviour, as $k \to \infty$, of k replications of the model which corresponds to a single observation, $\mathcal{M} \equiv \{\theta\, e^{-\theta\, y},\ y > 0,\ \theta > 0\}$, as opposed to k replications of the actual model for data $\boldsymbol{x}$, $\mathcal{M}^n \equiv \{\prod_{j=1}^n \theta\, e^{-\theta\, x_j},\ x_j > 0,\ \theta > 0\}$.

Thus, consider $\boldsymbol{y}^{(k)} = \{y_1, \ldots, y_k\}$, a random sample of size k from the single observation model $\mathcal{M}$; clearly $t_k = \sum_{j=1}^k y_j$ is sufficient, and the sampling distribution of t_k has a gamma density $p(t_k \,|\, \theta) = \mathrm{Ga}(t_k \,|\, k, \theta)$. Using a constant for the arbitrary function $h(\theta)$ in Theorem 3, the corresponding posterior has a gamma density $\mathrm{Ga}(\theta \,|\, k+1, t_k)$ and, thus,

$$f_k(\theta) = \exp\left[\int_0^\infty \mathrm{Ga}(t_k \,|\, k, \theta) \log\left\{\mathrm{Ga}(\theta \,|\, k+1, t_k)\right\} \mathrm{d}t_k\right] = c_k\, \theta^{-1},$$

where c_k is a constant which does not contain θ. Therefore, using (15), $f(\theta) = \theta_0/\theta$ and, since this is a permissible prior function (see Example 3), the unrestricted reference prior (for both the single observation model $\mathcal{M}$ and the actual model $\mathcal{M}^n$) is $\pi(\theta \,|\, \mathcal{M}^n, \mathcal{P}_0) = \pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0) = \theta^{-1}$.

Parametrizations are essentially arbitrary. As one would hope, reference priors are coherent under reparametrization in the sense that if $\phi = \phi(\theta)$ is a one-to-one mapping of Θ into $\Phi = \phi(\Theta)$ then, for all $\phi \in \Phi$,

(i) $\pi^{\phi}(\phi) = \pi^{\theta}\{\theta(\phi)\}$, if Θ is discrete;

(ii) $\pi^{\phi}(\phi) = \pi^{\theta}\{\theta(\phi)\}\,|\,\partial\theta(\phi)/\partial\phi\,|$, if Θ is continuous;

More generally, reference posteriors are coherent under piecewise invertible transformations $\phi = \phi(\theta)$ of the parameter θ in the sense that, for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, the reference posterior for ϕ derived from first principles, $\pi(\phi \,|\, \boldsymbol{x})$, is precisely the same as that which could be obtained from $\pi(\theta \,|\, \boldsymbol{x})$ by standard probability calculus:

Theorem 6 (Consistency under reparametrization) *Consider a model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta\}$ and let $\phi(\theta)$ be a piecewise invertible transformation of θ. For any data $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, the reference posterior density of ϕ, $\pi(\phi \,|\, \boldsymbol{x})$, is that induced by the reference posterior density of θ, $\pi(\theta \,|\, \boldsymbol{x})$.*

If $\phi(\theta)$ is one-to-one, Theorem 6 follows immediately from the fact that the expected information is also invariant under such transformation, so that, for all k, $I\{p_\theta \,|\, \mathcal{M}_\theta^k\} = I\{p_\psi \,|\, \mathcal{M}_\psi^k\}$; this may also be directly verified using Theorems 2 and 3. Suppose now that $\phi(\theta) = \phi_j(\theta)$, $\theta \in \Theta_j$, where the Θ_j's form a partition of Θ, such that each of the $\phi_j(\theta)$'s is one-to-one in Θ_j. The reference prior for θ only depends on the asymptotic posterior of θ which, for sufficiently large samples, will concentrate on that subset Θ_j of the parameter space Θ to which the true value of θ belongs. Since $\phi(\theta)$ is one-to-one within Θ_j, and reference priors are coherent under one-to-one parametrizations, the general result follows. □

An important consequence of Theorem 6 is that the reference prior of any location parameter, and the reference prior of the logarithm of any scale parameter are both uniform:

Theorem 7 (Location models and scale models) *Consider a* location *model $\mathcal{M}_1$, so that for some function f_1, $\mathcal{M}_1 \equiv \{f_1(x-\mu), x \in \mathbb{R}, \mu \in \mathbb{R}\}$, and let $\mathcal{P}_0$ be the class of all continuous strictly positive priors on $\mathbb{R}$; then, if it exists, a reference prior for μ is of the form $\pi(\mu \,|\, \mathcal{M}_1, \mathcal{P}_0) = c$. Moreover, if $\mathcal{M}_2$ is a* scale *model, $\mathcal{M}_2 \equiv \{\sigma^{-1} f_2(x/\sigma), x > 0, \sigma > 0\}$, and $\mathcal{P}_0$ is the class of all continuous strictly positive priors on $(0, \infty)$, then a reference prior for σ, if it exists, is of the form $\pi(\sigma \,|\, \mathcal{M}_2, \mathcal{P}_0) = c\,\sigma^{-1}$.*

Let $\pi(\mu)$ be the reference prior which corresponds to model $\mathcal{M}_1$; the changes $y = x + \alpha$ and $\theta = \mu + \alpha$ produce $\{f_1(y-\theta), y \in \mathcal{Y}, \theta \in \mathbb{R}\}$, which is again model $\mathcal{M}_1$. Hence, using Theorem 6, $\pi(\mu) = \pi(\mu + \alpha)$ for all α and, therefore, $\pi(\mu)$ must be constant. Moreover, the obvious changes $y = \log x$ and $\phi = \log \sigma$ transform the scale model $\mathcal{M}_2$ into a location model; hence, $\pi(\phi) = c$ and, therefore, $\pi(\sigma) \propto \sigma^{-1}$. □

Example 6 *Cauchy data.* Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from a Cauchy distribution with unknown location μ and known scale $\sigma = 1$, so that $p(x_j \,|\, \mu) \propto [1 + (x_j - \mu)^2]^{-1}$. Since this is a location model, the reference prior is uniform and, by Bayes theorem, the corresponding reference posterior is

$$\pi(\mu \,|\, \boldsymbol{x}) \propto \prod_{j=1}^{n} \left[1 + (x_j - \mu)^2\right]^{-1}, \qquad \mu \in \mathbb{R}.$$

Using the change of variable theorem, the reference posterior of (say) the one-to-one transformation $\phi = e^{\mu}/(1+e^{\mu})$ mapping the original parameter space $\mathbb{R}$ into $(0, 1)$, is $\pi(\phi \,|\, \boldsymbol{x}) = \pi(\mu(\phi) \,|\, \boldsymbol{x})|\partial\mu/\partial\phi|$, $\phi \in (0, 1)$. Similarly, the reference posterior $\pi(\psi \,|\, \boldsymbol{x})$ of (say) $\psi = \mu^2$ may be derived from $\pi(\mu \,|\, \boldsymbol{x})$ using standard change of variable techniques, since $\psi = \mu^2$ is a piecewise invertible function of μ, and Theorem 6 may therefore be applied.

3.3 Approximate location parametrization

Another consequence of Theorem 6 is that, for any model with one continuous parameter $\theta \in \Theta$, there is a parametrization $\phi = \phi(\theta)$ (which is unique up to a largely irrelevant proportionality constant), for which the reference prior is uniform. By Theorem 6 this may be obtained from the reference prior $\pi(\theta)$ in the original parametrization as a function $\phi = \phi(\theta)$ which satisfies the differential equation $\pi(\theta)|\partial\phi(\theta)/\partial\theta|^{-1} = 1$, that is, any solution to the indefinite integral $\phi(\theta) = \int \pi(\theta)\, \mathrm{d}\theta$. Intuitively, $\phi = \phi(\theta)$ may be expected to behave as an *approximate* location parameter; this links reference priors with the concept data translated likelihood inducing priors introduced by Box and Tiao (1973, Sec. 1.3). For many models, good simple approximations to the posterior distribution may be obtained in terms of this parametrization, which often yields an *exact* location model.

Definition 7 (Approximate location parametrization) *Consider the model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta \subset \mathbb{R}\}$. *An* approximate location parametrization $\phi = \phi(\theta)$ *for model* $\mathcal{M}$ *is one for which the reference prior is uniform. In continuous regular models, this is given by any solution to the indefinite integral* $\phi(\theta) = \int \pi(\theta)\, d\theta$, *where* $\pi(\theta) = \pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$ *is the (unrestricted) reference prior for the original parameter.*

Example 7 *Exponential model, continued.* Consider again the exponential model $\mathcal{M} \equiv \{\theta\, e^{-\theta\, x},\, x > 0, \theta > 0\}$. The reference prior for θ is (see Example 5) $\pi(\theta) = \theta^{-1}$; thus an approximate location parameter is $\phi = \phi(\theta) = \int \pi(\theta)\, \mathrm{d}\theta = \log \theta$. Using $y = -\log x$, this yields

$$\mathcal{M}_y \equiv \left\{ \exp\left[-(y - \phi) + e^{-(y-\phi)} \right], \quad y \in \mathbb{R}, \quad \phi \in \mathbb{R} \right\},$$

where ϕ is an (actually exact) location parameter.

Example 8 *Uniform model on* $(0, \theta)$. Let $\boldsymbol{x} = \{x_1, \ldots, x_k\}$ be a random sample from the uniform model $\mathcal{M} \equiv \{p(x \,|\, \theta) = \theta^{-1},\ 0 < x < \theta,\ \theta > 0\}$, so that $t_k = \max_{j=1}^{k} x_j$ is sufficient, and the sampling distribution of t_k is the inverted Pareto $p(t_k \,|\, \theta) = \mathrm{IPa}(t_k \,|\, k, \theta^{-1}) = k\,\theta^{-k}\,t_k^{k-1}$, if $0 < t_k < \theta$, and zero otherwise. Using a uniform prior for the arbitrary function $h(\theta)$ in Theorem 3, the corresponding posterior distribution has the Pareto density $\mathrm{Pa}(\theta \,|\, k-1, t_k) = (k-1)\,t_k^{k-1}\,\theta^{-k}$, $\theta > t_k$, and (14) becomes

$$f_k(\theta) = \exp\left[\int_0^{\theta} \mathrm{IPa}(t_k \,|\, k, \theta^{-1}) \log \mathrm{Pa}(\theta \,|\, k-1, t_k)\,\mathrm{d}t_k\right] = c_k\,\theta^{-1},$$

where c_k is a constant which does not contain θ. Therefore, using (15), $f(\theta) = \theta_0/\theta$, $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0) = \theta^{-1}$.

By Theorem 4, this is also the reference prior for samples of any size; hence, by Bayes theorem, the reference posterior density of θ after, say, a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of size n has been observed is

$$\pi(\theta \,|\, \boldsymbol{x}) \propto \textstyle\prod_{j=1}^{n} p(x_j \,|\, \theta)\,\pi(\theta) = \theta^{-(n+1)}, \quad \theta > t_n,$$

where $t_n = \max\{x_1, \ldots, x_n\}$, which is a kernel of the Pareto density $\pi(\theta \,|\, \boldsymbol{x}) = \pi(\theta \,|\, t_n) = \mathrm{Pa}(\theta \,|\, n, t_n) = n\,(t_n)^n\,\theta^{-(n+1)}$, $\theta > t_n$.

The approximate location parameter is $\phi(\theta) = \int \theta^{-1}\,\mathrm{d}\theta = \log\theta$. The sampling distribution of the sufficient statistic $s_n = \log t_n$ in terms of the new parameter is the reversed exponential $p(s_n \,|\, n, \phi) = n\,e^{-n(\phi - s_n)}$, $s_n < \phi$, which explicitly shows ϕ as an (exact) location parameter. The reference prior of ϕ is indeed uniform, and the reference posterior after $\boldsymbol{x}$ has been observed is the shifted exponential $\pi(\phi \,|\, \boldsymbol{x}) = n\,e^{-n(\phi - s_n)}$, $\phi > s_n$, which may also be obtained by changing variables in $\pi(\theta \,|\, \boldsymbol{x})$.

3.4 Numerical reference priors

Analytical derivation of reference priors may be technically demanding in complex models. However, Theorem 3 may also be used to obtain a numerical approximation to the reference prior which corresponds to any one-parameter model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \theta \in \Theta\}$ from which random observations may be efficiently simulated.

The proposed algorithm requires a numerical evaluation of Equation (14). This is relatively straightforward, for simulation from the assumed model may be used to approximate by Monte Carlo the integral in (14), and the evaluation of its integrand for each simulated set of data only requires (cheap) one-dimensional numerical integration. Moderate values of k (to simulate the asymptotic posterior) are typically sufficient to obtain a good approximation to the reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$ (up to an irrelevant proportionality constant). The appropriate pseudo code is quite simple:

(1) Starting values:
Choose a moderate value for k,
Choose an arbitrary positive function $h(\theta)$, say $h(\theta) = 1$.
Choose the number m of samples to be simulated,

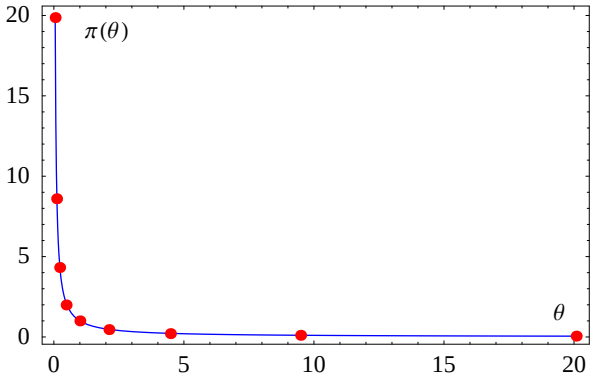
(2) For any given θ value, **repeat**, for $j = 1, \ldots, m$:
Simulate a random sample $\{\boldsymbol{x}_{1j}, \ldots, \boldsymbol{x}_{kj}\}$ of size k from $p(\boldsymbol{x} \,|\, \theta)$.
Compute numerically the integral $c_j = \int_\Theta \prod_{i=1}^k p(\boldsymbol{x}_{ij} \,|\, \theta)\, h(\theta)\, \mathrm{d}\theta$.
Evaluate $r_j(\theta) = \log[\, \prod_{i=1}^k p(\boldsymbol{x}_{ij} \,|\, \theta)\, h(\theta)/c_j \,]$.

(3) Compute $\pi(\theta) = \exp[\, m^{-1} \sum_{j=1}^m r_j(\theta) \,]$ and **store** the pair $\{\theta, \pi(\theta)\}$.

(4) **Repeat** routines (2) and (3) for all θ values for which the pair $\{\theta, \pi(\theta)\}$ is required.

Example 9 *Exponential data, continued.* Figure 3 represents the exact reference prior for the exponential model $\pi(\theta) = \theta^{-1}$ (continuous line) and the reference prior numerically calculated with the algorithm above for nine θ values, ranging from e^{-3} to e^3, uniformly log-spaced and rescaled to have $\pi(1) = 1$; $m = 500$ samples of $k = 25$ observations were used to compute each of the nine $\{\theta_i, \pi(\theta_i)\}$ points.

Figure 3 *Numerical reference prior for the exponential model*



If required, a continuous approximation to $\pi(\theta)$ may easily be obtained from the computed points using standard interpolation techniques.

An educated choice of the arbitrary function $h(\theta)$ often leads to an analytical form for the required posterior, $p(\theta \,|\, \boldsymbol{x}_{1j}, \ldots, \boldsymbol{x}_{kj}) \propto \prod_{i=1}^k p(\boldsymbol{x}_{ij} \,|\, \theta)\, h(\theta)$; for instance, this is the case in Example 9 if $h(\theta)$ is chosen to be of the form $h(\theta) = \theta^a$, for some $a \geq -1$. If the posterior may be analytically computed, then the values of the $r_j(\theta) = \log[\, p(\theta \,|\, \boldsymbol{x}_{1j}, \ldots, \boldsymbol{x}_{kj}) \,]$ are immediately obtained, and the numerical algorithm reduces to only one Monte Carlo integration for each desired pair $\{\theta_i, \pi(\theta_i)\}$.

For an alternative, MCMC based, numerical computation method of reference priors, see Lafferty and Wasserman (2001).

If data consist of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of a model with one continuous parameter θ, it is often possible to find an *asymptotically sufficient* statistic $\tilde{\theta}_n = \tilde{\theta}_n(x_1, \ldots, x_n)$ which is also a *consistent* estimator of θ; for example, under regularity conditions, the maximum likelihood estimator (mle) $\hat{\theta}_n$ is consistent and asymptotically sufficient. In that case, the reference prior may easily be obtained in terms of *either* (i) an asymptotic approximation $\pi(\theta \,|\, \tilde{\theta}_n)$ to the posterior distribution of θ, or (ii) the sampling distribution $p(\tilde{\theta}_n \,|\, \theta)$ of the asymptotically sufficient consistent estimator $\tilde{\theta}_n$.

Theorem 8 (Reference priors under regularity conditions) *Let available data $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ consist of a random sample of any size from a one-parameter model $\mathcal{M} \equiv \{p(x \,|\, \theta), x \in \mathcal{X}, \theta \in \Theta\}$. Let $\boldsymbol{x}^{(k)} = \{x_1, \ldots, x_k\}$ be a random sample of size k from model $\mathcal{M}$, let $\tilde{\theta}_k = \tilde{\theta}_k(\boldsymbol{x}^{(k)}) \in \Theta$ be an asymptotically sufficient statistic which is a consistent estimator of θ, and let $\mathcal{P}_0$ be the class of all continuous priors with support Θ. Let $\pi_k(\theta \,|\, \tilde{\theta}_k)$ be any asymptotic approximation (as $k \to \infty$) to the posterior distribution of θ, let $p(\tilde{\theta}_k \,|\, \theta)$ be the sampling distribution of $\tilde{\theta}_k$, and define*

$$f_k^a(\theta) = \pi_k(\theta \,|\, \tilde{\theta}_k)\Big|_{\tilde{\theta}_k = \theta}, \quad f^a(\theta) = \lim_{k \to \infty} \frac{f_k^a(\theta)}{f_k^a(\theta_0)} \tag{21}$$

$$f_k^b(\theta) = p(\tilde{\theta}_k \,|\, \theta)\Big|_{\tilde{\theta}_k = \theta}, \quad f^b(\theta) = \lim_{k \to \infty} \frac{f_k^b(\theta)}{f_k^b(\theta_0)}\,, \tag{22}$$

where θ_0 is any interior point of Θ. Then, under frequently occurring additional technical conditions, $f^a(\theta) = f^b(\theta) = f(\theta)$ and, if $f(\theta)$ is a permissible prior, any function of the form $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0) \propto f(\theta)$ is a reference prior for θ.

Since $\tilde{\theta}_k$ is asymptotically sufficient, Equation (14) in Theorem 3 becomes

$$f_k(\theta) = \exp\Big\{ \int_\Theta p(\tilde{\theta}_k \,|\, \theta) \log \pi_k(\theta \,|\, \tilde{\theta}_k)\, \mathrm{d}\tilde{\theta}_k \Big\}.$$

Moreover, since $\tilde{\theta}_k$ is consistent, the sampling distribution of $\tilde{\theta}_k$ will concentrate on θ as $k \to \infty$, $f_k(\theta)$ will converge to $f_k^a(\theta)$, and Equation (21) will have the same limit as Equation (15). Moreover, for any formal prior function $h(\theta)$,

$$\pi(\theta \,|\, \tilde{\theta}_k) = \frac{p(\tilde{\theta}_k \,|\, \theta)\, h(\theta)}{\int_\Theta p(\tilde{\theta}_k \,|\, \theta)\, h(\theta)\, \mathrm{d}\theta}\,.$$

As $k \to \infty$, the integral in the denominator converges to $h(\tilde{\theta}_k)$ and, therefore, $f_k^a(\theta) = \pi(\theta \,|\, \tilde{\theta}_k)\,|_{\tilde{\theta}_k = \theta}$ converges to $p(\tilde{\theta}_k \,|\, \theta)|_{\tilde{\theta}_k = \theta} = f_k^b(\theta)$. Thus, both limits in Equations (21) and (22) yield the same result, and their common value provides an explicit expression for the reference prior. For details, and precise technical conditions, see Berger, Bernardo and Sun (2005). □

Example 10 *Exponential model, continued.* Let $\boldsymbol{x} = \{x_1, \ldots, x_k\}$ be a random sample of k exponential observations from $\mathrm{Ex}(x \,|\, \theta)$. The mle is $\hat{\theta}_k(\boldsymbol{x}) = 1/\overline{x}$, a sufficient, consistent estimator of θ whose sampling distribution is the inverted gamma $p(\hat{\theta}_k \,|\, \theta) = \mathrm{IGa}(\hat{\theta}_k \,|\, k\theta, k)$. Therefore, $f_k^b(\theta) = p(\hat{\theta}_k \,|\, \theta)|_{\hat{\theta}_k = \theta} = c_k/\theta$, where $c_k = e^{-k}k^k/\Gamma(k)$ and, using Theorem 8, the reference prior is $\pi(\theta) = \theta^{-1}$.

Alternatively, the likelihood function is $\theta^n e^{-k\theta/\hat{\theta}_k}$; hence, for any positive function $h(\theta)$, $\pi_k(\theta \,|\, \hat{\theta}_k) \propto \theta^k e^{-k\theta/\hat{\theta}_k}\, h(\theta)$ is an asymptotic approximation to the posterior distribution of θ. Taking, for instance, $h(\theta) = 1$, this yields the gamma posterior $\pi_k(\theta \,|\, \hat{\theta}_k) = \mathrm{Ga}(\theta \,|\, k+1, k/\hat{\theta}_k)$. Consequently, $f_k^a(\theta) = \pi_k(\theta \,|\, \hat{\theta}_k)\,|_{\hat{\theta}_k = \theta} = c_k/\theta$, and $\pi(\theta) = \theta^{-1}$ as before.

Example 11 *Uniform model, continued.* Let $\boldsymbol{x} = \{x_1, \ldots, x_k\}$ be a random sample of k uniform observations from $\mathrm{Un}(x \,|\, 0, \theta)$. The mle is $\hat{\theta}_k(\boldsymbol{x}) = \max\{x_1, \ldots, x_k\}$, a sufficient, consistent estimator of θ whose sampling distribution is the inverted Pareto $p(\hat{\theta}_k \,|\, \theta) = \mathrm{IPa}(\hat{\theta}_k \,|\, k, \theta^{-1})$. Therefore, $f_k^b(\theta) = p(\hat{\theta}_k \,|\, \theta)|_{\hat{\theta}_k = \theta} = k/\theta$ and, using Theorem 8, the reference prior is $\pi(\theta) = \theta^{-1}$.

Alternatively, the likelihood function is θ^{-k}, $\theta > \hat{\theta}_k$; hence, taking for instance a uniform prior, the Pareto $\pi^k(\theta \,|\, \hat{\theta}_k) = \mathrm{Pa}(\theta \,|\, k-1, \hat{\theta}_k)$ is found to be a particular asymptotic approximation of the posterior of θ; thus, $f_k^a(\theta) = \pi_k(\theta \,|\, \hat{\theta}_k)\,|_{\hat{\theta}_k = \theta} = (k-1)/\theta$, and $\pi(\theta) = \theta^{-1}$ as before.

The posterior distribution of the parameter is often asymptotically normal (see *e.g.*, Bernardo and Smith (1994, Sec. 5.3), and references therein). In this case, the reference prior is easily derived. The result includes (univariate) Jeffreys (1946) and Perks (1947) rules as a particular cases:

Theorem 9 (Reference priors under asymptotic normality) *Let data consist of a random sample from model $\mathcal{M} \equiv \{p(\boldsymbol{y} \,|\, \theta), \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}, \theta \in \Theta \subset \mathbb{R}\}$, and let $\mathcal{P}_0$ be the class of all continuous priors with support Θ. If the posterior distribution of θ, $\pi(\theta \,|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k)$, is asymptotically normal with standard deviation $s(\tilde{\theta}_k)/\sqrt{k}$, where $\tilde{\theta}_k$ is a consistent estimator of θ, and $s(\theta)^{-1}$ is a permissible prior function, then any function of the form*

$$\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0) \propto s(\theta)^{-1} \tag{23}$$

is a reference prior. Under appropriate regularity conditions the posterior distribution of θ is asymptotically normal with variance $i(\hat{\theta}_k)^{-1}/k$, where $\hat{\theta}_k$ is the mle of θ and

$$i(\theta) = -\int_{\mathcal{Y}} p(\boldsymbol{y} \,|\, \theta)\, \frac{\partial^2}{\partial \theta^2} \log p(\boldsymbol{y} \,|\, \theta)\, d\boldsymbol{y} \tag{24}$$

is Fisher's information function. If this is the case, and $i(\theta)^{1/2}$ is a permissible prior function, the reference prior is Jeffreys prior, $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0) \propto i(\theta)^{1/2}$.

The result follows directly from Theorem 8 since, uder the assumed conditions, $f_k^a(\theta) = \pi(\theta \,|\, \hat{\theta}_k) \,|_{\hat{\theta}_k=\theta} = c_k s(\theta)^{-1}$. Jeffreys prior is the particular case which obtains when $s(\theta) = i(\theta)^{-1/2}$. □

Jeffreys (1946, 1961) prior, independently rediscovered by Perks (1947), was central in the early objective Bayesian reformulation of standard textbook problems of statistical inference (Lindley, 1965; Zellner, 1971; Press, 1972; Box and Tiao, 1973). By Theorem 9, this is also the reference prior in regular models with one continuous parameter, whose posterior distribution is asymptotically normal. By Theorem 6, reference priors are coherently transformed under one-to-one reparametrizations; hence, Theorem 9 may be typically applied with any mathematically convenient (re)parametrization. For conditions which preserve asymptotic normality under transformations see Mendoza (1994).

The posterior distribution of the exponential parameter in Example 10 is asymptotically normal; thus the corresponding reference prior may also be obtained using Theorem 9; the reference prior for the uniform parameter in Example 11 cannot be obtained however in this way, since the relevant posterior distribution is *not* asymptotically normal. Notice that, even under conditions which guarantee asymptotic normality, Jeffreys formula is not necessarily the easiest way to derive a reference prior; indeed, Theorem 8 often provides a simpler alternative.

3.6 Reference priors and the likelihood principle

By definition, reference priors are a function of the *entire* statistical model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta\}$, not of the *observed* likelihood. Indeed, the reference prior $\pi(\theta \,|\, \mathcal{M})$ is a mathematical description of lack of information about θ *relative* to the information about θ which could be obtained by repeated sampling from a particular experimental design $\mathcal{M}$. If the design is changed, the reference prior may be expected to change accordingly. This is now illustrated by comparing the reference priors which correspond to direct and inverse sampling of Bernoulli observations.

Example 12 *Binomial and negative Binomial data.* Let available data $\boldsymbol{x} = \{r, m\}$ consist of m Bernoulli trials (with m fixed in advance) which contain r successes, so that the assumed model is Binomial $\text{Bi}(r \,|\, m, \theta)$:

$$\mathcal{M}_1 \equiv \{p(r \,|\, m, \theta) = \binom{m}{r} \theta^r (1-\theta)^{m-r},\ r = 0, 1, \ldots, m, \quad 0 < \theta < 1\}$$

Using Theorem 9, with $n = 1$, m fixed, and $\boldsymbol{y} = r$, the reference prior for θ is the (proper) prior $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$; Bayes theorem yields the Beta reference posterior $\pi(\theta \,|\, \boldsymbol{x}) = \text{Be}(\theta \,|\, r + 1/2,\ m - r + 1/2)$. Notice that $\pi(\theta \,|\, \boldsymbol{x})$ is proper, for all values of r; in particular, if $r = 0$, the reference posterior is $\pi(\theta \,|\, \boldsymbol{x}) = \text{Be}(\theta \,|\, 1/2,\ m + 1/2)$, from which sensible

conclusions may be reached, even though there are no observed successes. This may be compared with the Haldane (1948) prior, also proposed by Jaynes (1968), $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$, which produces an improper posterior until at least one success and one failure are observed.

Consider, however, that data $\boldsymbol{x} = \{r, m\}$ consist of the sequence of Bernoulli trials observed until r successes are obtained (with $r \geq 1$ fixed in advance), so that the assumed model is negative Binomial:

$$\mathcal{M}_2 \equiv \{p(m \,|\, r, \theta) = \binom{m-1}{r-1} \theta^r (1-\theta)^{m-r},\ m = r, r+1, \ldots \quad 0 < \theta < 1\}$$

Using Theorem 9, with $n = 1$ and $\boldsymbol{y} = m$, the reference prior for θ is the (improper) prior $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$, and Bayes theorem yields the Beta reference posterior $\pi(\theta \,|\, \boldsymbol{x}) = \text{Be}(\theta \,|\, r,\, m - r + 1/2)$, which is proper whatever the number of observations m required to obtain r successes. Notice that $r = 0$ is *not* possible under this model: inverse Binomial sampling implicitly assumes that $r \geq 1$ successes will occur for sure.

In reporting results, scientists are typically required to specify not only the observed data but also the conditions under which those where obtained, the *design* of the experiment, so that the data analyst has available the full specification of the model, $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$. To carry out a reference analysis of the data, such a full specification (that is, including the experimental design) is indeed required. The reference prior $\pi(\boldsymbol{\omega} \,|\, \mathcal{M}, \mathcal{P})$ is proposed as a *consensus* prior to analyse data *associated to a particular design* $\mathcal{M}$ (and under any agreed assumptions about the value of $\boldsymbol{\omega}$ which might be encapsulated in the choice of $\mathcal{P}$).

The *likelihood principle* (Berger and Wolpert, 1988) says that all evidence about an unknown quantity $\boldsymbol{\omega}$, which is obtained from an experiment which has produced data $\boldsymbol{x}$, is contained in the likelihood function $p(\boldsymbol{x} \,|\, \boldsymbol{\omega})$ of $\boldsymbol{\omega}$ for the *observed* data $\boldsymbol{x}$. In particular, for any *specific* prior beliefs (described by a *fixed* prior), proportional likelihoods should produce the same posterior distribution.

As Example 12 demonstrates, it may be argued that formal use of reference priors is not compatible with the likelihood principle. However, the likelihood principle applies *after* data have been observed while reference priors are derived *before* the data are observed. Reference priors are a (limiting) form of rather specific beliefs, namely those which would maximize the missing information (about the quantity or interest) *associated to a particular design*, and thus depend on the particular design considered. There is no claim that these particular beliefs describe (or even approximate) those of any particular individual; instead, they are precisely defined as possible *consensus* prior functions, presumably useful as a *reference* for scientific communication. Notice that reference prior *functions* (often improper) should *not* be interpreted

as prior probability *distributions*: they are merely technical devices to facilitate the derivation of reference posteriors, and only reference posteriors support a probability interpretation.

Any statistical analysis should include an evaluation of the sensitivity of the results to accepted assumptions. In particular, any Bayesian analysis should include some discussion of the sensitivity of the results to the choice of the prior, and reference priors are better viewed as a useful tool for this important aspect of *sensitivity analysis*. The analyst is supposed to have a unique (often subjective) prior $p(\boldsymbol{\omega})$, independent of the design of the experiment, but the scientific community will presumably be interested in comparing the corresponding analyst's personal posterior with the *reference* (consensus) posterior associated to the published experimental design. To report reference posteriors (possibly for a range of alternative designs) should be seen as part of this sensitivity analysis. Indeed, reference analysis provides an answer to an important *conditional* question in scientific inference: the reference posterior encapsulates what *could* be said about the quantity of interest *if* prior information about its value were minimal *relative* to the information which repeated data from an specific experimental design $\mathcal{M}$ could possibly provide.

3.7 Restricted reference priors

The reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ is that which maximizes the missing information about θ relative to model $\mathcal{M}$ among the priors which belong to $\mathcal{P}$, the class of all sufficiently regular priors which are compatible with available knowledge (Definition 6). By restricting the class $\mathcal{P}$ of candidate priors to those which satisfy specific restrictions (derived from assumed knowledge) one may use the reference prior algorithm as an effective tool for *prior elicitation*: the corresponding reference prior will incorporate the accepted restrictions, but no other information.

Under regularity conditions, Theorems 3, 8 and 9, make it relatively simple to obtain the unrestricted reference prior $\pi(\theta) = \pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$ which corresponds to the case where the class of candidate priors is the class $\mathcal{P}_0$ of all continuous priors with support Θ. Hence, it is useful to be able to express a general reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ in terms of the corresponding unrestricted reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$, and the set of restrictions which define the class $\mathcal{P}$ of candidate priors.

If the unrestricted reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$ is proper, then $\pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ is the closest prior in $\mathcal{P}$ to $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$, in the sense of minimizing the intrinsic discrepancy (see Definition 1) between them, so that

$$\pi(\theta \,|\, \mathcal{M}, \mathcal{P}) = \arg \inf_{p(\theta) \in \mathcal{P}} \delta\{\, p(\theta),\ \pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)\,\}$$

If $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$ is not proper it may be necessary to derive $\pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ from its definition. However, in the rather large class of problems where the conditions

which define $\mathcal{P}$ may all be expressed in the general form $\int_\Theta g_i(\theta)\, p(\theta)\, \mathrm{d}\theta = \beta_i$, for appropriately chosen functions $g_i(\theta)$, (*i.e.*, as a collection of expected values which the prior $p(\theta)$ must satisfy), an explicit solution is available in terms of the unrestricted reference prior:

Theorem 10 (Explicit form of restricted reference priors) *Consider a model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta\}$, let $\mathcal{P}$ be the class of continuous proper priors with support Θ*

$$\mathcal{P} = \Big\{ p_\theta; \quad \int_\Theta p(\theta)\, d\theta = 1, \quad \int_\Theta g_i(\theta)\, p(\theta)\, d\theta = \beta_i, \quad i = 1, \ldots, m \Big\}$$

which satisfies the restrictions imposed by the expected values $\mathrm{E}[g_i(\theta)] = \beta_i$, and let $\mathcal{P}_0$ be the class of all continuous priors with support Θ. The reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P})$, if it exists, is then of the form

$$\pi(\theta \,|\, \mathcal{M}, \mathcal{P}) \propto \pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)\ \exp\Big\{ \textstyle\sum_{i=1}^m \lambda_i\, g_i(\theta) \Big\}$$

where the λ_i's are constants determined by the conditions which define $\mathcal{P}$.

Theorem 10 may be proven using a standard calculus of variations argument. If $m = 0$, so that one only has the constraint that the prior is proper, then there typically is no restricted reference prior. For details, see Bernardo and Smith (1994, p. 316). □

Example 13 *Location models, continued.* Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from a location model $\mathcal{M} \equiv \{f(x - \mu), x \in \mathcal{X}, \mu \in \mathbb{R}\}$, and suppose that the prior mean and variance of μ are restricted to be $\mathrm{E}[\mu] = \mu_0$, and $\mathrm{Var}[\mu] = \sigma_0^2$. By Theorem 7, the unrestricted reference prior $\pi(\mu \,|\, \mathcal{M}, \mathcal{P}_0)$ is uniform; hence, using Theorem 10, the (restricted) reference prior must be of the form

$$\pi(\mu \,|\, \mathcal{M}, \mathcal{P}) \propto \exp\{\lambda_1 \mu + \lambda_2 (\mu - \mu_0)^2\}$$

with $\int_{-\infty}^{\infty} \mu\, \pi(\mu \,|\, \mathcal{M}, \mathcal{P})\, \mathrm{d}\mu = \mu_0$ and $\int_{-\infty}^{\infty} (\mu - \mu_0)^2\, \pi(\mu \,|\, \mathcal{M}, \mathcal{P})\, \mathrm{d}\mu = \sigma_0^2$. It follows that $\lambda_1 = 0$ and $\lambda_2 = -1/(2\sigma_0^2)$ and, substituting above, the restricted reference prior is $\pi(\mu \,|\, \mathcal{M}, \mathcal{P}) \propto \exp\{-(\mu - \mu_0)^2/(2\sigma_0^2)\}$, which is the *normal* distribution $\mathrm{N}(\mu \,|\, \mu_0, \sigma_0)$ with the specified mean and variance. This provides a very powerful argument for the choice of a normal density to describe prior information in location models, when prior knowledge about the location parameter is *limited* to its first two moments.

3.8 One nuisance parameter

Consider now the case where the statistical model $\mathcal{M}$ contains one nuisance parameter, so that $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta, \lambda), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta, \lambda \in \Lambda\}$, the quantity of

interest is $\theta \in \Theta \subset \mathbb{R}$, and the nuisance parameter is $\lambda \in \Lambda \subset \mathbb{R}$. To obtain the required reference posterior for θ, $\pi(\theta \,|\, \boldsymbol{x})$, an appropriate *joint* reference prior $\pi^{\theta}(\theta, \lambda)$ is obviously needed: by Bayes theorem, the corresponding joint posterior is $\pi^{\theta}(\theta, \lambda \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta, \lambda)\, \pi^{\theta}(\theta, \lambda)$ and, integrating out the nuisance parameter, the (marginal) reference posterior for the parameter of interest is

$$\pi(\theta \,|\, \boldsymbol{x}) = \int_{\Lambda} \pi^{\theta}(\theta, \lambda \,|\, \boldsymbol{x})\, \mathrm{d}\lambda \propto \int_{\Lambda} p(\boldsymbol{x} \,|\, \theta, \lambda)\, \pi^{\theta}(\theta, \lambda)\, \mathrm{d}\lambda.$$

The extension of the reference prior algorithm to the case of two parameters follows the usual mathematical procedure of reducing the two parameter problem to a sequential application of the established procedure for the single parameter case. Thus, the reference algorithm proceeds by combining the results obtained in two successive applications of the one-parameter solution:

(1) Conditional on θ, $p(\boldsymbol{x} \,|\, \theta, \lambda)$ only depends on the nuisance parameter λ and, hence, the one-parameter algorithm may be used to obtain the *conditional* reference prior $\pi(\lambda \,|\, \theta) = \pi(\lambda \,|\, \theta, \mathcal{M}, \mathcal{P})$.

(2) If $\pi(\lambda \,|\, \theta)$ has a finite integral in Λ (so that, when normalized, yields a proper density with $\int_{\Lambda} \pi(\lambda \,|\, \theta)\, \mathrm{d}\lambda = 1$), the conditional reference prior $\pi(\lambda \,|\, \theta)$ may be used to integrate out the nuisance parameter and derive the one-parameter integrated model,

$$p(\boldsymbol{x} \,|\, \theta) = \int_{\Lambda} p(\boldsymbol{x} \,|\, \theta, \lambda)\, \pi(\lambda \,|\, \theta)\, d\lambda, \tag{25}$$

to which the one-parameter algorithm may be applied again to obtain the *marginal* reference prior $\pi(\theta) = \pi(\theta \,|\, \mathcal{M}, \mathcal{P})$.

(3) The desired θ-reference prior is then $\pi^{\theta}(\theta, \lambda) = \pi(\lambda \,|\, \theta)\, \pi(\theta)$, and the required reference posterior is

$$\pi(\theta \,|\, \boldsymbol{x}) \propto \int_{\Lambda} p(\boldsymbol{x} \,|\, \theta, \lambda)\, \pi^{\theta}(\theta, \lambda)\, \mathrm{d}\lambda = p(\boldsymbol{x} \,|\, \theta)\, \pi(\theta). \tag{26}$$

Equation (25) suggests that conditional reference priors provides a general procedure to eliminate nuisance parameters, a major problem within the frequentist paradigm. For a review of this important topic, see Liseo (2005), in this volume.

If the conditional reference prior $\pi(\lambda \,|\, \theta)$ is *not* proper, Equation (25) does not define a valid statistical model and, as a consequence, a more subtle approach is needed to provide a general solution; this will be described later. Nevertheless, the simple algorithm described above may be used to obtain appropriate solutions to a number of interesting problems which serve to illustrate the crucial need to identify the quantity of interest, as is the following two examples.

Example 14 *Induction.* Consider a finite population of (known) size N, all of whose elements may or may not have a specified property. A random sample of size n is taken without replacement, and all the elements in the sample turn out to have that property. Scientific interest often centres in the probability that all the N elements in the population have the property under consideration (natural induction). It has often been argued that for relatively large n values, this should be close to one whatever might be the population size N (typically much larger than the sample size n). Thus, if all the $n = 225$ randomly chosen turtles in an isolated volcanic island are found to show a particular difference with respect to those in the mainland, zoologists would tend to believe that all the turtles in the island share that property. Formally, if r and R respectively denote the number of elements in the sample and in the population which have the property under study, the statistical model is

$$\mathcal{M} \equiv \Big\{ p(r \,|\, n, R, N), \quad r \in \{0, \ldots, n\}, \quad R \in \{0, \ldots, N\} \Big\},$$

where R is the unknown parameter, and $p(r \,|\, n, R, N) = \binom{R}{r}\binom{N-R}{n-r} / \binom{N}{n}$ is the relevant hypergeometric distribution. The required result,

$$p(R = N \,|\, r = n, N) = \frac{p(r = n \,|\, n, R, N)\, p(R = N)}{\sum_{R=0}^{N} p(r = n \,|\, n, R, N)\, p(R)} \,. \tag{27}$$

may immediately be obtained from Bayes theorem, once a prior $p(R)$ for the unknown number R of elements in the population which have the property has been established. If the parameter of interest were R itself, the reference prior would be uniform over its range (Theorem 2), so that $p(R) = (N + 1)^{-1}$; using (27) this would lead to the posterior probability $p(R = N \,|\, r = n, N) = (n + 1)/(N + 1)$ which will be small when (as it is usually the case) the sampling fraction n/N is small. However, the quantity of interest here is *not* the value of R but whether or not $R = N$, and a reference prior is desired which maximizes the missing information about this *specific* question. Rewriting the unknown parameter as $R = (\theta, \lambda)$, where $\theta = 1$ if $R = N$ and $\theta = 0$ otherwise, and $\lambda = 1$ if $R = N$ and $\lambda = R$ otherwise (so that the quantity of interest θ is explicitly shown), and using Theorem 2 and the argument above, one gets $\pi(\lambda \,|\, \theta = 1) = 1$, $\pi(\lambda \,|\, \theta = 0) = N^{-1}$, and $\pi(\theta = 0) = \pi(\theta = 1) = 1/2$, so that the θ-reference prior is $\pi^{\theta}(R) = 1/2$ if $R = N$ and $\pi^{\theta}(R) = 1/(2N)$ if $R \neq N$. Using (27), this leads to

$$p(R = N \,|\, r = n, N) = \left[1 + \frac{1}{n+1} \left(1 - \frac{n}{N} \right) \right]^{-1} \approx \frac{n+1}{n+2} \tag{28}$$

which, as expected, clearly displays the irrelevance of the sampling fraction, and the approach to unity for large n. In the turtles example (a real question posed to the author at the Galápagos Islands in the eighties), this

yields $p(R = N \,|\, r = n = 225,\, N) \approx 0.995$ for all large N. The *reference* result (28) does not necessarily represents any personal scientist's beliefs (although apparently it may approach actual scientists's beliefs in many situations), but the conclusions which should be reached from a situation where the missing information about the quantity of interest (whether or not $R = N$) is maximized, a situation mathematically characterized by the θ-reference prior described above. For further discussion of this problem (with important applications in philosophy of science, physical sciences and reliability), see Jeffreys (1961, pp. 128–132), Geisser (1984), Bernardo (1985b) and Singpurwalla and Wilson (2004).

Example 15 *Ratio of multinomial parameters.* Let data $\boldsymbol{x} = \{r_1, r_2, n\}$ consist of the result of n trinomial observations, with parameters α_1, α_2 and $\alpha_3 = 1 - \alpha_1 - \alpha_2$, so that, for $0 < \alpha_i < 1$, $\alpha_1 + \alpha_2 < 1$,

$$p(r_1, r_2 \,|\, n, \alpha_1, \alpha_2) = c(r_1, r_2, n)\, \alpha_1^{r_1} \alpha_2^{r_2} (1 - \alpha_1 - \alpha_2)^{n - r_1 - r_2},$$

where $c(r_1, r_2, n) = (n!)/(r_1!\, r_2!\, (n - r_1 - r_2)!)$, and suppose that the quantity of interest is the *ratio* $\theta = \alpha_1/\alpha_2$ of the first two original parameters. Reparametrization in terms of θ and (say) $\lambda = \alpha_2$ yields

$$p(r_1, r_2 \,|\, n, \theta, \lambda) = c(r_1, r_2, n)\, \theta^{r_1}\, \lambda^{r_1 + r_2}\, \{1 - \lambda(1 + \theta)\}^{n - r_1 - r_2},$$

for $\theta > 0$ and, given θ, $0 < \lambda < (1 + \theta)^{-1}$. Conditional on θ, this is a model with one continuous parameter λ, and the corresponding Fisher information function is $i(\lambda \,|\, \theta) = n(1 + \theta)/\{\lambda(1 - \lambda(1 + \theta))\}$; using Theorem 9 the conditional reference prior of the nuisance parameter is $\pi(\lambda \,|\, \theta) \propto i(\lambda \,|\, \theta)^{1/2}$ which is the *proper* beta-like prior $\pi(\lambda \,|\, \theta) \propto \lambda^{-1/2} \{1 - \lambda(1 + \theta)\}^{-1/2}$, with support on $\lambda \in [0, (1 + \theta)^{-1}]$ (which depends on θ). Integration of the full model $p(r_1, r_2 \,|\, n, \theta, \lambda)$ with the conditional reference prior $\pi(\lambda \,|\, \theta)$ yields $p(r_1, r_2 \,|\, n, \theta) = \int_0^{(1+\theta)^{-1}} p(r_1, r_2 \,|\, n, \theta, \lambda)\, \pi(\lambda \,|\, \theta)\, \mathrm{d}\lambda$, the *integrated* one-parameter model

$$p(r_1, r_2 \,|\, n, \theta) = \frac{\Gamma(r_1 + r_2 + \frac{1}{2})\, \Gamma(n - r_1 - r_2 + \frac{1}{2})}{r_1!\, r_2!\, (n - r_1 - r_2)!}\; \frac{\theta^{r_1}}{(1 + \theta)^{r_1 + r_2}}\;.$$

The corresponding Fisher information function is $i(\theta) = n/\{2\theta(1 + \theta)^2\}$; using again Theorem 9 the reference prior of the parameter of interest is $\pi(\theta) \propto i(\theta)^{1/2}$ which is the proper prior $\pi(\theta) \propto \theta^{-1/2}(1 + \theta)^{-1}$, $\theta > 0$. Hence, by Bayes theorem, the reference posterior of the quantity of interest is $\pi(\theta \,|\, r_1, r_2, n) \propto p(r_1, r_2 \,|\, n, \theta)\, \pi(\theta)$; this yields

$$\pi(\theta \,|\, r_1, r_2) = \frac{\Gamma(r_1 + r_2 + 1)}{\Gamma(r_1 + \frac{1}{2})\, \Gamma(r_2 + \frac{1}{2})}\; \frac{\theta^{r_1 - 1/2}}{(1 + \theta)^{r_1 + r_2 + 1}}\;, \quad \theta > 0.$$

Notice that $\pi(\theta \,|\, r_1, r_2)$ does *not* depend on n; to draw conclusions about the value of $\theta = \alpha_1/\alpha_2$ only the numbers r_1 and r_2 observed in the first

two classes matter: a result $\{55, 45, 100\}$ carries precisely the same information about the *ratio* α_1/α_2 than a result $\{55, 45, 10000\}$. For instance, if an electoral survey of size n yields r_1 voters for party A and r_2 voters for party B, the reference posterior distribution of the *ratio* θ of the proportion of A voters to B voters in the population only depends on their respective number of voters in the sample, r_1 and r_2, whatever the size and political intentions of the other $n - r_1 - r_2$ citizens in the sample. In particular, the reference posterior probability that party A gets better results than party B is $\Pr[\theta > 1 \,|\, r_1, r_2] = \int_1^\infty \pi(\theta \,|\, r_1, r_2)\, \mathrm{d}\theta$. As one would expect, this is precisely equal to $1/2$ if, and only if, $r_1 = r_2$; one-dimensional numerical integration (or use of the incomplete beta function) is required to compute other values. For instance, whatever the total sample size n in each case, this yields $\Pr[\theta > 1 \,|\, r_1 = 55, r_2 = 45] = 0.841$ (with $r_1 + r_2 = 100$) and $\Pr[\theta > 1 \,|\, r_1 = 550, r_2 = 450] = 0.999$ (with the same ratio r_1/r_2, but $r_1 + r_2 = 1000$).

As illustrated by the preceding examples, in a multiparameter model, say $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ the required (joint) reference prior $\pi^\theta(\boldsymbol{\omega})$ may depend on the quantity of interest, $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega})$ (although, as one would certainly expect, and will later be demonstrated, this will *not* be the case if the new quantity of interest $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\omega})$ say, is a one-to-one function of $\boldsymbol{\theta}$). Notice that this does *not* mean that the analyst's beliefs should depend on his or her interests; as stressed before, reference priors are not meant to describe the analyst's beliefs, but the mathematical formulation of a particular type of prior beliefs—those which would maximize the expected missing information about the quantity of interest—which could be adopted by consensus as a standard for scientific communication.

If the conditional reference prior $\pi(\lambda \,|\, \theta)$ is *not* proper, so that Equation (25) does not define a valid statistical model, then integration may be performed within each of the elements of an increasing sequence $\{\Lambda_i\}_{i=1}^\infty$ of subsets of Λ converging to Λ over which $\pi(\lambda \,|\, \theta)$ is integrable. Thus, Equation (25) is to be replaced by

$$p_i(\boldsymbol{x} \,|\, \theta) = \int_{\Lambda_i} p(\boldsymbol{x} \,|\, \theta, \lambda)\, \pi_i(\lambda \,|\, \theta)\, d\lambda, \tag{29}$$

where $\pi_i(\lambda \,|\, \theta)$ is the renormalized proper restriction of $\pi(\lambda \,|\, \theta)$ to Λ_i, from which the reference posterior $\pi_i(\theta \,|\, \boldsymbol{x}) = \pi(\theta \,|\, \mathcal{M}_i, \mathcal{P})$, which corresponds to model $\mathcal{M}_i \equiv \{p(\boldsymbol{x} \,|\, \theta, \lambda), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta, \lambda \in \Lambda_i\}$ may be derived.

The use of the sequence $\{\Lambda_i\}_{i=1}^\infty$ makes it possible to obtain a corresponding sequence of θ-reference posteriors $\{\pi_i(\theta \,|\, \boldsymbol{x})\}_{i=1}^\infty$ for the quantity of interest θ which corresponds to the sequence of integrated models (29); the required reference posterior may then be found as the corresponding intrinsic limit $\pi(\theta \,|\, \boldsymbol{x}) = \lim_{i \to \infty} \pi_i(\theta \,|\, \boldsymbol{x})$. A θ-reference prior is then defined as any positive function $\pi^\theta(\theta, \lambda)$ which may formally be used in Bayes' theorem to directly

obtain the reference posterior, so that for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, the posterior density satisfies $\pi(\theta \,|\, \boldsymbol{x}) \propto \int_\Lambda p(\boldsymbol{x} \,|\, \theta, \lambda)\, \pi^\theta(\theta, \lambda)\, d\lambda$.

The approximating sequences should be *consistently* chosen within the same model: given a statistical model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ an appropriate approximating sequence $\{\boldsymbol{\Omega}_i\}$ should be chosen for the whole parameter space $\boldsymbol{\Omega}$. Thus, if the analysis is done in terms of $\psi = \{\psi_1, \psi_2\} \in \Psi(\boldsymbol{\Omega})$, the approximating sequence should be chosen such that $\Psi_i = \psi(\boldsymbol{\Omega}_i)$. A very natural approximating sequence in location-scale problems is $\{\mu, \log \sigma\} \in [-i, i]^2$; reparametrization to asymptotically independent parameters and approximate location reparametrizations (Definition 7) may be combined to choose appropriate approximating sequences in more complex situations. A formal definition of reference prior functions in multiparameter problems is possible along the lines of Definition 6.

As one would hope, the θ-reference prior does *not* depend on the choice of the nuisance parameter λ; thus, for any $\psi = \psi(\theta, \lambda)$ such that (θ, ψ) is a one-to-one function of (θ, λ), the θ-reference prior in terms of (θ, ψ) is simply $\pi^\theta(\theta, \psi) = \pi^\theta(\theta, \lambda)/|\partial(\theta, \psi)/\partial(\theta, \lambda)|$, the appropriate probability transformation of the θ-reference prior in terms of (θ, λ). Notice however that, as mentioned before, the reference prior *may* depend on the parameter of interest; thus, the θ-reference prior may differ from the ϕ-reference prior unless either ϕ is a one-to-one transformation of θ, or ϕ is asymptotically independent of θ. This is an expected consequence of the mathematical fact that the prior which maximizes the missing information about θ is not generally the same as the prior which maximizes the missing information about any function $\phi = \phi(\theta, \lambda)$.

The *non-existence* of a unique "noninformative" prior for all inference problems within a given model was established by Dawid, Stone and Zidek (1973) when they showed that this is incompatible with *consistent marginalization*. Indeed, given a two-parameter model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta, \lambda), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta, \lambda \in \Lambda\}$, if the reference posterior of the quantity of interest θ, $\pi(\theta \,|\, \boldsymbol{x}) = \pi(\theta \,|\, \boldsymbol{t})$, only depends on the data through a statistic $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$ whose sampling distribution, $p(\boldsymbol{t} \,|\, \theta, \lambda) = p(\boldsymbol{t} \,|\, \theta)$, only depends on θ, one would expect the reference posterior to be of the form $\pi(\theta \,|\, \boldsymbol{t}) \propto p(\boldsymbol{t} \,|\, \theta)\, \pi(\theta)$ for some prior $\pi(\theta)$. However, examples were found where this *cannot* be the case if a *unique* joint "noninformative" prior were to be used for all possible quantities of interest within the same statistical model $\mathcal{M}$.

By definition, a reference prior must be a *permissible* prior function. In particular (Definition 3), it must yield *proper posteriors* for all data sets large enough to identify the parameters. For instance, if data $\boldsymbol{x}$ consist of a random sample of fixed size n from a normal $\mathrm{N}(x \,|\, \mu, \sigma)$ distribution, so that, $\mathcal{M} \equiv \{\prod_{j=1}^n \mathrm{N}(x_j \,|\, \mu, \sigma),\ x_j \in \mathbb{R},\ \sigma > 0\}$, the function $\pi^\mu(\mu, \sigma) = \sigma^{-1}$ is only a permissible (joint) prior for μ if $n \geq 2$ (and, without restrictions in the class $\mathcal{P}$ of candidate priors, a reference prior function *does not exist* for $n = 1$).

Under posterior asymptotic normality, reference priors are easily obtained in terms of the relevant Fisher information matrix. The following result extends Theorem 9 to models with two continuous parameters:

Theorem 11 (Reference priors under asymptotic binormality) *Let data* $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ *consist of* n *conditionally independent (given* θ*) observations from a model* $\mathcal{M} \equiv \{p(\boldsymbol{y} \,|\, \theta, \lambda), \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}, \theta \in \Theta, \lambda \in \Lambda\}$*, and let* $\mathcal{P}_0$ *be the class of all continuous (joint) priors with support* $\Theta \times \Lambda\}$*. If the posterior distribution of* $\{\theta, \lambda\}$ *is asymptotically normal with dispersion matrix* $V(\hat{\theta}_n, \hat{\lambda}_n)/n$*, where* $\{\hat{\theta}_n, \hat{\lambda}_n\}$ *is a consistent estimator of* $\{\theta, \lambda\}$*, define*

$$V(\theta, \lambda) = \begin{pmatrix} v_{\theta\theta}(\theta, \lambda) & v_{\theta\lambda}(\theta, \lambda) \\ v_{\theta\lambda}(\theta, \lambda) & v_{\lambda\lambda}(\theta, \lambda) \end{pmatrix}, \quad H(\theta, \lambda) = V^{-1}(\theta, \lambda), \quad \textit{and}$$

$$\pi(\lambda \,|\, \theta) \propto h^{1/2}_{\lambda\lambda}(\theta, \lambda), \quad \lambda \in \Lambda, \tag{30}$$

and, if $\pi(\lambda \,|\, \theta)$ *is proper,*

$$\pi(\theta) \propto \exp\Big\{ \int_\Lambda \pi(\lambda \,|\, \theta) \, \log[v^{-1/2}_{\theta\theta}(\theta, \lambda)] \, d\lambda \Big\}, \quad \theta \in \Theta. \tag{31}$$

Then, if $\pi(\lambda \,|\, \theta)\, \pi(\theta)$ *is a permissible prior function, the* θ*-reference prior is*

$$\pi(\theta \,|\, \mathcal{M}^n, \mathcal{P}_0) \propto \pi(\lambda \,|\, \theta)\, \pi(\theta).$$

If $\pi(\lambda \,|\, \theta)$ *is not proper, integration in* (31) *is performed on elements of an increasing sequence* $\{\Lambda_i\}_{i=1}^\infty$ *such that* $\int_{\Lambda_i} \pi(\lambda \,|\, \theta)\, d\lambda < \infty$*, to obtain the sequence* $\{\pi_i(\lambda \,|\, \theta)\, \pi_i(\theta)\}_{i=1}^\infty$*, where* $\pi_i(\lambda \,|\, \theta)$ *is the renormalization of* $\pi(\lambda \,|\, \theta)$ *to* Λ_i*, and the* θ*-reference prior* $\pi^\theta(\theta, \lambda)$ *is defined as its corresponding intrinsic limit.*

A heuristic justification of Theorem 11 is now provided. Under the stated conditions, given k independent observations from model $\mathcal{M}$, the *conditional* posterior distribution of λ given θ is asymptotically normal with precision $k\, h_{\lambda\lambda}(\theta, \hat{\lambda}_k)$, and the *marginal* posterior distribution of θ is asymptotically normal with precision $k\, v^{-1}_{\theta\theta}(\hat{\theta}_k, \hat{\lambda}_k)$; thus, using Theorem 9, $\pi(\lambda \,|\, \theta) \propto h^{1/2}_{\lambda\lambda}(\theta, \lambda)$, which is Equation (30). Moreover, using Theorem 3,

$$\pi_k(\theta) \propto \exp\Big\{ \iint p(\hat{\theta}_k, \hat{\lambda}_k \,|\, \theta) \log[\mathrm{N}\{\theta \,|\, \hat{\theta}_k, k^{-1/2} v^{1/2}_{\theta\theta}(\hat{\theta}_k, \hat{\lambda}_k)\}] \, \mathrm{d}\hat{\theta}_k \, \mathrm{d}\hat{\lambda}_k \Big\} \tag{32}$$

where, if $\pi(\lambda \,|\, \theta)$ is proper, the integrated model $p(\hat{\theta}_k, \hat{\lambda}_k \,|\, \theta)$ is given by

$$p(\hat{\theta}_k, \hat{\lambda}_k \,|\, \theta) = \int_\Lambda p(\hat{\theta}_k, \hat{\lambda}_k \,|\, \theta, \lambda)\, \pi(\lambda \,|\, \theta) \, \mathrm{d}\lambda. \tag{33}$$

Introducing (33) into (32) and using the fact that $(\hat{\theta}_k, \hat{\lambda}_k)$ is a consistent estimator of (θ, λ)—so that as $k \to \infty$ integration with $p(\hat{\theta}_k, \hat{\lambda}_k \,|\, \theta, \lambda)$ reduces

to substitution of $(\hat{\theta}_k, \hat{\lambda}_k)$ by (θ, λ)—directly leads to Equation (31). If $\pi(\lambda \,|\, \theta)$ is not proper, it is necessary to integrate in an increasing sequence $\{\Lambda_i\}_{i=1}^{\infty}$ of subsets of Λ such that the restriction $\pi_i(\lambda \,|\, \theta)$ of $\pi(\lambda \,|\, \theta)$ to Λ_i is proper, obtain the sequence of reference priors which correspond to these restricted models, and then take limits to obtain the required result. □

Notice that under appropriate regularity conditions (see *e.g.*, Bernardo and Smith (1994, Sec. 5.3) and references therein) the joint posterior distribution of $\{\theta, \lambda\}$ is asymptotically normal with precision matrix $n\, I(\hat{\theta}_n, \hat{\lambda}_n)$, where $I(\theta)$ is Fisher information matrix; in that case, the asymptotic dispersion matrix in Theorem 11 is simply $V(\theta, \lambda) = I^{-1}(\theta, \lambda)/n$.

Theorem 12 (Reference priors under factorization) *In the conditions of Theorem 11, if* (i) *θ and λ are variation independent—so that Λ does not depend on θ—and* (ii) *both $h_{\lambda\lambda}(\theta, \lambda)$ and $v_{\theta\theta}(\theta, \lambda)$ factorize, so that*

$$v_{\theta\theta}^{-1/2}(\theta, \lambda) \propto f_\theta(\theta)\, g_\theta(\lambda), \quad h_{\lambda\lambda}^{1/2}(\theta, \lambda) \propto f_\lambda(\theta)\, g_\lambda(\lambda), \tag{34}$$

then the θ-reference prior is simply $\pi^\theta(\theta, \lambda) = f_\theta(\theta)\, g_\lambda(\lambda)$, even if the conditional reference prior $\pi(\lambda \,|\, \theta) = \pi(\lambda) \propto g_\lambda(\lambda)$ is improper.

If $h_{\lambda\lambda}^{1/2}(\theta, \lambda)$ factorizes as $h_{\lambda\lambda}^{1/2}(\theta, \lambda) = f_\lambda(\theta) g_\lambda(\lambda)$, then the conditional reference prior is $\pi(\lambda \,|\, \theta) \propto f_\lambda(\theta) g_\lambda(\lambda)$ and, normalizing, $\pi(\lambda \,|\, \theta) = c_1\, g_\lambda(\lambda)$, which does not depend on θ. If, furthermore, $v_{\theta\theta}^{-1/2}(\theta, \lambda) = f_\theta(\theta) g_\theta(\lambda)$ and Λ does not depend on θ, Equation (31) reduces to

$$\pi(\theta) \propto \exp\Big\{ \int_\Lambda c_1\, g_\lambda(\lambda) \log[f_\theta(\theta) g_\theta(\lambda)]\, \mathrm{d}\lambda \Big\} = c_2\, f_\theta(\theta)$$

and, hence, the reference prior is $\pi^\theta(\theta, \lambda) = \pi(\lambda \,|\, \theta)\, \pi(\theta) = c\, f_\theta(\theta)\, g_\lambda(\lambda)$. □

Example 16 *Inference on the univariate normal parameters.* Let data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ consist of a random sample of fixed size n from a normal distribution $\mathrm{N}(x \,|\, \mu, \sigma)$. The information matrix $I(\mu, \sigma)$ and its inverse matrix are respectively

$$I(\mu, \sigma) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{pmatrix}, \quad V(\mu, \sigma) = I^{-1}(\mu, \sigma) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{1}{2}\sigma^2 \end{pmatrix}.$$

Hence, $i_{\sigma\sigma}^{1/2}(\mu, \sigma) = \sqrt{2}\, \sigma^{-1} = f_\sigma(\mu)\, g_\sigma(\sigma)$, with $g_\sigma(\sigma) = \sigma^{-1}$, so that $\pi(\sigma \,|\, \mu) = \sigma^{-1}$. Similarly, $v_{\mu\mu}^{-1/2}(\mu, \sigma) = \sigma^{-1} = f_\mu(\mu)\, g_\sigma(\sigma)$, with $f_\mu(\mu) = 1$, and thus $\pi(\mu) = 1$. Therefore, using Theorem 11 the μ-reference prior is $\pi^\mu(\mu, \sigma) = \pi(\sigma \,|\, \mu)\, \pi(\mu) = \sigma^{-1}$ for all $n \geq 2$. For $n = 1$ the posterior distribution is not proper, the function $h(\mu, \sigma) = \sigma^{-1}$ is *not* a permissible prior, and a reference prior does not exist. Besides, since $I(\mu, \sigma)$ is diagonal, the σ-reference prior is $\pi^\sigma(\mu, \sigma) = f_\sigma(\sigma)\, g_\mu(\mu) = \sigma^{-1}$, the same as $\pi^\mu(\mu, \sigma)$.

Consider now the case where the quantity of interest is *not* the mean μ or the standard deviation σ, but the *standardized* mean $\phi = \mu/\sigma$ (or, equivalently, the coefficient of variation σ/μ). Fisher's matrix in terms of the parameters ϕ and σ is $I(\phi, \sigma) = J^t\, I(\mu, \sigma)\, J$, where $J = (\partial(\mu, \sigma)/\partial(\phi, \sigma))$ is the Jacobian of the inverse transformation, and this yields

$$I(\phi, \sigma) = \begin{pmatrix} 1 & \phi\sigma^{-1} \\ \phi\sigma^{-1} & \sigma^{-2}(2+\phi^2) \end{pmatrix}, \quad V(\phi, \sigma) = \begin{pmatrix} 1+\frac{1}{2}\phi^2 & -\frac{1}{2}\phi\sigma \\ -\frac{1}{2}\phi\sigma & \frac{1}{2}\sigma^2 \end{pmatrix}.$$

Thus, $i^{1/2}_{\sigma\sigma}(\phi, \sigma) = \sigma^{-1}(2+\phi^2)^{1/2}$, and $v^{-1/2}_{\phi\phi}(\phi, \sigma) = (1+\frac{1}{2}\phi^2)^{-1/2}$. Hence, using Theorem 11, $\pi^{\phi}(\phi, \sigma) = (1+\frac{1}{2}\phi^2)^{-1/2}\sigma^{-1}$ $(n \geq 2)$. In the original parametrization, this is $\pi^{\phi}(\mu, \sigma) = (1+\frac{1}{2}(\mu/\sigma)^2)^{-1/2}\sigma^{-2}$, which is *very* different from $\pi^{\mu}(\mu, \sigma) = \pi^{\sigma}(\mu, \sigma) = \sigma^{-1}$. The reference posterior of the quantity of interest ϕ after data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ have been observed is

$$\pi(\phi \,|\, \boldsymbol{x}) \propto (1+\tfrac{1}{2}\phi^2)^{-1/2}\, p(t \,|\, \phi) \tag{35}$$

where $t = (\sum x_j)/(\sum x_j^2)^{1/2}$, a one-dimensional statistic whose sampling distribution, $p(t \,|\, \mu, \sigma) = p(t \,|\, \phi)$, only depends on ϕ. Thus, the reference prior algorithm is seen to be consistent under marginalization.

The reference priors $\pi^{\mu}(\mu, \sigma) = \sigma^{-1}$ and $\pi^{\sigma}(\mu, \sigma) = \sigma^{-1}$ for the normal location and scale parameters obtained in the first part of Example 16 are just a particular case of a far more general result:

Theorem 13 (Location-scale models) *If $\mathcal{M}$ is a* location-scale *model, so that for some function f, $\mathcal{M} \equiv \sigma^{-1} f\{(x-\mu)/\sigma\}$, $x \in \mathcal{X}$, $\mu \in \mathbb{R}, \sigma > 0\}$, and $\mathcal{P}_0$ is the class of all continuous, strictly positive (joint) priors for (μ, σ), then a reference prior for either μ or σ, if it exists, is of the form*

$$\pi^{\mu}(\mu, \sigma \,|\, \mathcal{M}, \mathcal{P}_0) = \pi^{\sigma}(\mu, \sigma \,|\, \mathcal{M}, \mathcal{P}_0) \propto \sigma^{-1}.$$

For a proof, which is based on the form of the relevant Fisher matrix, see Fernández and Steel (1999b). □

When the quantity of interest and the nuisance parameter are *not* variation independent, derivation of the reference prior requires special care. This is illustrated in the example below:

Example 17 *Product of positive normal means.* Let data consist of two independent random samples $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ and $\boldsymbol{y} = \{y_1, \ldots, y_m\}$ from $N(x \,|\, \alpha, 1)$ and $N(y \,|\, \beta, 1)$, $\alpha > 0$, $\beta > 0$, so that the assumed model is

$$p(\boldsymbol{x}, \boldsymbol{y} \,|\, \alpha, \beta) = \textstyle\prod_{i=1}^{n} N(x_i \,|\, \alpha, 1) \prod_{j=1}^{m} N(y_j \,|\, \beta, 1), \quad \alpha > 0,\ \beta > 0,$$

and suppose that the quantity of interest is the product of the means,

$\theta = \alpha\beta$, a frequent situation in physics and engineering. Reparametrizing in terms of the one-to-one transformation $(\theta, \lambda) = (\alpha\beta, \alpha/\beta)$, Fisher matrix $I(\theta, \lambda)$ and its inverse matrix $V(\theta, \lambda)$ are,

$$I = \begin{pmatrix} \frac{m+n\lambda^2}{4\theta\lambda} & \frac{1}{4}\left(n - \frac{m}{\lambda^2}\right) \\ \frac{1}{4}\left(n - \frac{m}{\lambda^2}\right) & \frac{\theta\,(m+n\lambda^2)}{4\lambda^3} \end{pmatrix}, \quad V = \begin{pmatrix} \theta(\frac{1}{n\lambda} + \frac{\lambda}{m}) & \frac{1}{n} - \frac{\lambda^2}{m} \\ \frac{1}{n} - \frac{\lambda^2}{m} & \frac{\lambda(m+n\lambda^2)}{nm\theta} \end{pmatrix}.$$

and, therefore, using (30),

$$\pi(\lambda \,|\, \theta) \propto I_{22}(\theta, \lambda)^{1/2} \propto \theta^{1/2}(m + n\lambda^2)^{1/2}\lambda^{-3/2}. \tag{36}$$

The natural increasing sequence of subsets of the original parameter space, $\Omega_i = \{(\alpha, \beta);\ 0 < \alpha < i,\ 0 < \beta < i\}$, transforms, in the parameter space of λ, into the sequence $\Lambda_i(\theta) = \{\lambda;\ \theta\, i^{-2} < \lambda < i^2\, \theta^{-1}\}$. Notice that this depends on θ, so that θ and λ are *not* variation independent and, hence, Theorem 12 *cannot* be applied. Renormalizing (36) in $\Lambda_i(\theta)$ and using (31), it is found that, for large i,

$$\begin{aligned} \pi_i(\lambda \,|\, \theta) &= c_i(m, n)\, \theta^{1/2}(m + n\lambda^2)^{1/2}\lambda^{-3/2} \\ \pi_i(\theta) &= c_i(m, n) \int_{\Lambda_i(\theta)} (m + n\lambda^2)^{1/2}\lambda^{-3/2}\ \log\left(\frac{\lambda}{m} + \frac{1}{n\lambda}\right)^{-1/2} d\lambda, \end{aligned}$$

where $c_i(m, n) = i^{-1}\sqrt{nm}/(\sqrt{m} + \sqrt{n})$, which leads to the θ-reference prior $\pi^\theta(\theta, \lambda) \propto \theta^{1/2}\lambda^{-1}\left(\frac{\lambda}{m} + \frac{1}{n\lambda}\right)^{1/2}$. In the original parametrization, this corresponds to

$$\pi^\theta(\alpha, \beta) \propto (n\alpha^2 + m\beta^2)^{1/2}, \quad n \geq 1,\ m \geq 1 \tag{37}$$

which depends on the sample sizes through the ratio m/n. It has already been stressed that the reference prior depends on the experimental design. It is therefore not surprising that, if the design is unbalanced, the reference prior depends on the ratio m/n which controls the level of balance. Notice that the reference prior (37) is very different from the uniform prior $\pi^\alpha(\alpha, \beta) = \pi^\beta(\alpha, \beta) = 1$, which should be used to make reference inferences about either α or β.

It will later be demonstrated (Example 22) that the prior $\pi^\theta(\alpha, \beta)$ found above provides approximate agreement between Bayesian credible regions and frequentist confidence intervals for θ (Berger and Bernardo, 1989); indeed, this prior was originally suggested by Stein (1986) (who only considered the case $m = n$) to obtain such approximate agreement. Efron (1986) used this problem as an example in which *conventional* objective Bayesian theory encounters difficulties since, even within a fixed model $\mathcal{M} \equiv \{p(\boldsymbol{y} \,|\, \boldsymbol{\theta}), \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, the "correct" objective prior depends on the particular function $\phi = \phi(\boldsymbol{\theta})$ one

desires to estimate. For the reference priors associated to generalizations of the product of normal means problem, see Sun and Ye (1995, 1999).

3.9 Many parameters

Theorems 11 and 12 may easily be extended to any number of nuisance parameters. Indeed, let data $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ consist of a random sample of size n from a model $\mathcal{M} \equiv \{p(\boldsymbol{y} \,|\, \boldsymbol{\omega}), \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}, \boldsymbol{\omega} = \{\omega_1, \ldots, \omega_m\}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, let ω_1 be the quantity of interest, assume regularity conditions to guarantee that, as $n \to \infty$, the joint posterior distribution of $\boldsymbol{\omega}$ is asymptotically normal with mean $\hat{\boldsymbol{\omega}}_n$ and dispersion matrix $V(\hat{\boldsymbol{\omega}}_n)/n$, and let $H(\boldsymbol{\omega}) = V^{-1}(\boldsymbol{\omega})$. It then follows that, if $V_j(\boldsymbol{\omega})$ is the $j \times j$ upper matrix of $V(\boldsymbol{\omega})$, $j = 1, \ldots, m$, $H_j(\boldsymbol{\omega}) = V_j^{-1}(\boldsymbol{\omega})$ and $h_{jj}(\boldsymbol{\omega})$ is the lower right (j, j) element of $H_j(\boldsymbol{\omega})$, then

(1) the *conditional* posterior distribution of ω_j given $\{\omega_1, \ldots, \omega_{j-1}\}$, is asymptotically normal with precision $n\, h_{jj}(\hat{\boldsymbol{\omega}}_n)$, $(j = 2, \ldots, m)$ and
(2) the *marginal* posterior distribution of ω_1 is asymptotically normal with precision $n\, h_{11}(\hat{\boldsymbol{\omega}}_n)$.

This may be used to extend the algorithm described in Theorem 11 to sequentially derive $\pi(\omega_m \,|\, \omega_1, \ldots, \omega_{m-1})$, $\pi(\omega_{m-1} \,|\, \omega_1, \ldots, \omega_{m-2}), \ldots, \pi(\omega_2 \,|\, \omega_1)$ and $\pi(\omega_1)$; their product yields the reference prior associated to the particular ordering $\{\omega_1, \omega_2, \ldots, \omega_m\}$. Intuitively, this is a mathematical description of a situation where, relative to the particular design considered $\mathcal{M}$, one maximizes the missing information about the parameter ω_1 (that of higher inferential importance), but also the missing information about ω_2 given ω_1, that of ω_3 given ω_1 and ω_2,... and that of ω_m given ω_1 to ω_{m-1}. As in sequential decision theory, this must be done backwards. In particular, to maximize the missing information about ω_1, the prior which maximizes the missing information about ω_2 given ω_1 has to be derived first.

The choice of the ordered parametrization, say $\{\theta_1(\boldsymbol{\omega}), \theta_2(\boldsymbol{\omega}), \ldots, \theta_m(\boldsymbol{\omega})\}$, precisely describes the particular prior required, namely that which sequentially maximizes the missing information about the θ_j's in order of inferential interest. Indeed, "diffuse" prior knowledge about a particular sequence $\{\theta_1(\boldsymbol{\omega}), \theta_2(\boldsymbol{\omega}), \ldots, \theta_m(\boldsymbol{\omega})\}$ may be very "precise" knowledge about another sequence $\{\phi_1(\boldsymbol{\omega}), \phi_2(\boldsymbol{\omega}), \ldots, \phi_m(\boldsymbol{\omega})\}$ unless, *for all* j, $\phi_j(\boldsymbol{\omega})$ is a one-to-one function of $\theta_j(\boldsymbol{\omega})$. Failure to recognize this fact is known to produce untenable results; famous examples are the paradox of Stein (1959) (see Example 19 below) and the marginalization paradoxes (see Example 16).

Theorem 14 (Reference priors under asymptotic normality) *Let data* $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ *consist of a random sample of size* n *from a statistical model* $\mathcal{M} \equiv \{p(\boldsymbol{y} \,|\, \boldsymbol{\theta}), \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}, \boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}, \boldsymbol{\theta} \in \boldsymbol{\Theta} = \prod_{j=1}^{m} \Theta_j\}$, *and let* $\mathcal{P}_0$ *be the class of all continuous priors with support* $\boldsymbol{\Theta}$. *If the posterior distribution of* $\boldsymbol{\theta}$ *is asymptotically normal with dispersion matrix* $V(\hat{\boldsymbol{\theta}}_n)/n$, *where* $\hat{\boldsymbol{\theta}}_n$ *is a consistent estimator of* $\boldsymbol{\theta}$, $H(\boldsymbol{\theta}) = V^{-1}(\boldsymbol{\theta})$, V_j *is the upper* $j \times j$ *submatrix of* V,

$H_j = V_j^{-1}$, and $h_{jj}(\boldsymbol{\theta})$ is the lower right element of H_j, then the $\boldsymbol{\theta}$-reference prior, associated to the ordered parametrization $\{\theta_1, \ldots, \theta_m\}$, is

$$\pi(\boldsymbol{\theta} \,|\, \mathcal{M}^n, \mathcal{P}_0) = \pi(\theta_m \,|\, \theta_1, \ldots, \theta_{m-1}) \times \cdots \times \pi(\theta_2 \,|\, \theta_1)\, \pi(\theta_1),$$

with $\pi(\theta_m \,|\, \theta_1, \ldots, \theta_{m-1}) = h_{mm}^{1/2}(\boldsymbol{\theta})$ and, for $i = 1, \ldots, m-1$,

$$\pi(\theta_j \,|\, \theta_1, \ldots, \theta_{j-1}) \propto \exp\left\{ \int_{\boldsymbol{\Theta}^{j+1}} \prod_{l=j+1}^{m} \pi(\theta_l \,|\, \theta_1, \ldots, \theta_{l-1}) \log[h_{jj}^{1/2}(\boldsymbol{\theta})]\, d\boldsymbol{\theta}^{j+1} \right\}$$

with $\boldsymbol{\theta}^{j+1} = \{\theta_{j+1}, \ldots, \theta_m\}$, provided $\pi(\theta_j \,|\, \theta_1, \ldots, \theta_{j-1})$ is proper for all j.
If the conditional reference priors $\pi(\theta_j \,|\, \theta_1, \ldots, \theta_{j-1})$ are not all proper, integration is performed on elements of an increasing sequence $\{\boldsymbol{\Theta}_i\}_{i=1}^{\infty}$ such that $\int_{\Theta_{ij}} \pi(\theta_j \,|\, \theta_1, \ldots, \theta_{j-1})\, d\theta_j$ is finite, to obtain the corresponding sequence $\{\pi_i(\boldsymbol{\theta})\}_{i=1}^{\infty}$ of reference priors for the restricted models. The $\boldsymbol{\theta}$-reference prior is then defined as their intrinsic limit.
If, moreover, (i) *Θ_j does not depend on $\{\theta_1, \ldots, \theta_{j-1}\}$, and* (ii) *the functions $h_{jj}(\theta, \lambda)$ factorize in the form*

$$h_{jj}^{1/2}(\boldsymbol{\theta}) \propto f_j(\theta_j)\, g_j(\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_m), \quad j = 1, \ldots, m,$$

then the $\boldsymbol{\theta}$-reference prior is simply $\pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \prod_{j=1}^{m} f_j(\theta_j)$, even if the conditional reference priors are improper.

Under appropriate regularity conditions—see *e.g.*, Bernardo and Smith (1994, Theo. 5.14)—the posterior distribution of $\boldsymbol{\theta}$ is asymptotically normal with mean the mle $\hat{\boldsymbol{\theta}}_n$ and precision matrix $n\, I(\hat{\boldsymbol{\theta}}_n)$, where $I(\boldsymbol{\theta})$ is Fisher matrix,

$$i_{ij}(\boldsymbol{\theta}) = -\int_{\boldsymbol{\mathcal{Y}}} p(\boldsymbol{y} \,|\, \boldsymbol{\theta})\, \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log[p(\boldsymbol{y} \,|\, \boldsymbol{\theta})]\, \mathrm{d}\boldsymbol{y};$$

in that case, $H(\boldsymbol{\theta}) = n\, I(\boldsymbol{\theta})$, and the reference prior may be computed from the elements of Fisher matrix $I(\boldsymbol{\theta})$. Notice, however, that in the multivariate case, the reference prior does *not* yield Jeffreys multivariate rule (Jeffreys, 1961), $\pi^J(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}$. For instance, in location-scale models, the (μ, σ)-reference prior and the (σ, μ)-reference prior are both $\pi^R(\mu, \sigma) = \sigma^{-1}$ (Theorem 13), while Jeffreys multivariate rule yields $\pi^J(\mu, \sigma) = \sigma^{-2}$. As a matter of fact, Jeffreys himself criticised his own multivariate rule. This is known, for instance, to produce both marginalization paradoxes Dawid, Stone and Zidek (1973), and strong inconsistencies (Eaton and Freedman, 2004). See, also, Stein (1962) and Example 23.

Theorem 14 provides a procedure to obtain the reference prior $\pi^{\boldsymbol{\theta}}(\boldsymbol{\theta})$ which corresponds to any *ordered parametrization* $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$. Notice that, within any particular multiparameter model

$$\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}), \quad \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \quad \boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\} \in \boldsymbol{\Theta} \subset \mathbb{R}^k\},$$

the reference algorithm provides a (possibly different) joint reference prior

$$\pi^{\boldsymbol{\phi}}(\boldsymbol{\phi}) = \pi(\phi_m \,|\, \phi_1, \ldots, \phi_{m-1}) \times \cdots \times \pi(\phi_2 \,|\, \phi_1)\, \pi(\phi_1),$$

for each possible ordered parametrization $\{\phi_1(\boldsymbol{\theta}), \phi_2(\boldsymbol{\theta}), \ldots, \phi_m(\boldsymbol{\theta})\}$. However, as one would hope, the results are coherent under monotone transformations of each of the $\phi_i(\boldsymbol{\theta})$'s in the sense that, in that case, $\pi^{\boldsymbol{\phi}}(\boldsymbol{\phi}) = \pi^{\boldsymbol{\theta}}[\,\boldsymbol{\theta}(\boldsymbol{\phi})\,]|J(\boldsymbol{\phi})|$, where $J(\boldsymbol{\phi})$ is the Jacobian of the inverse transformation $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\phi})$, of general element $j_{ij}(\boldsymbol{\phi}) = \partial\theta_i(\boldsymbol{\phi})/\partial\phi_j$. This property of coherence under appropriate reparametrizations may be very useful in choosing a particular parametrization (for instance one with orthogonal parameters, or one in which the relevant $h_{jj}(\boldsymbol{\theta})$ functions factorize) which simplifies the implementation of the algorithm.

Starting with Jeffreys (1946) pioneering work, the analysis of the invariance properties under reparametrization of multiparameter objective priors has a very rich history. Relevant pointers include Hartigan (1964), Stone (1965, 1970), Zidek (1969), Florens (1978, 1982), Dawid (1983), Consonni and Veronese (1989b), Chang and Eaves (1990), George and McCulloch (1993), Datta and J. K. Ghosh (1995b), Yang (1995), Datta and M. Ghosh (1996), Eaton and Sudderth (1999, 2002, 2004) and Severini, Mukerjee and Ghosh (2002). In particular, Datta and J. K. Ghosh (1995b), Yang (1995) and Datta and M. Ghosh (1996) are specifically concerned with the invariance properties of reference distributions.

Example 18 *Multivariate normal data.* Let data consist of a size n random sample $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$, $n \geq 2$, from an m-variate normal distribution with mean $\boldsymbol{\mu}$, and covariance matrix $\sigma^2\,\mathbf{I}_m$, $m \geq 1$, so that

$$I(\boldsymbol{\mu}, \sigma) = \begin{pmatrix} \sigma^{-2}\,\mathbf{I}_m & 0 \\ 0 & (2/m)\,\sigma^{-2} \end{pmatrix}$$

It follows from Theorem 14 that the reference prior relative to the natural parametrization $\boldsymbol{\theta} = \{\mu_1, \ldots, \mu_m, \sigma\}$ is $\pi^{\boldsymbol{\theta}}(\mu_1, \ldots, \mu_m, \sigma) \propto \sigma^{-1}$, and also that the result does not depend on the order in which the parametrization is taken, since their asymptotic covariances are zero. Hence, $\pi^{\boldsymbol{\theta}}(\mu_1, \ldots, \mu_m, \sigma) \propto \sigma^{-1}$ is the appropriate prior function to obtain the reference posterior of any piecewise invertible function $\phi(\mu_j)$ of μ_j, and also to obtain the reference posterior of any piecewise invertible function $\phi(\sigma)$ of σ. In particular, the corresponding reference posterior for any of the μ_j's is easily shown to be the Student density

$$\pi(\mu_j \,|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \mathrm{St}\left\{\mu_j \;\middle|\; \overline{y}_j,\, s/\sqrt{(n-1)},\, m(n-1)\right\}$$

with $n\overline{y}_j = \sum_{i=1}^{n} y_{ij}$, and $nms^2 = \sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \overline{y}_j)^2$, which agrees with the standard argument according to which one degree of freedom should

be lost by each of the unknown means. Similarly, the reference posterior of σ^2 is the inverted Gamma

$$\pi(\sigma^2 \,|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \mathrm{IGa}\{\sigma^2 \,|\, n(m-1)/2,\, nms^2/2\}$$

When $m = 1$, these results reduce to those obtained in Example 16.

Example 19 *Stein's paradox.* Let $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ be a random sample of size n from a m-variate normal $\mathrm{N}_m(\boldsymbol{x} \,|\, \boldsymbol{\mu}, I_m)$ with mean $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_m\}$ and unitary dispersion matrix. The reference prior which corresponds to any permutation of the μ_i's is uniform, and this uniform prior leads indeed to appropriate reference posterior distributions for any of the μ_j's, given by $\pi(\mu_j \,|\, \boldsymbol{x}) = \mathrm{N}(\mu_j \,|\, \overline{x}_j, 1/\sqrt{n})$. Suppose, however, that the quantity of interest is $\theta = \sum_j \mu_j^2$, the squared distance of $\boldsymbol{\mu}$ from the origin. As shown by Stein (1959), the posterior distribution of θ based on the uniform prior (or indeed any "flat" *proper* approximation) has very undesirable properties; this is due to the fact that a uniform (or nearly uniform) prior, although "noninformative" with respect to each of the individual μ_j's, is actually highly informative on the sum of their squares, introducing a severe bias towards large values of θ (Stein's paradox). However, the reference prior which corresponds to a parametrization of the form $\{\theta, \boldsymbol{\lambda}\}$ produces, for any choice of the nuisance parameter vector $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\mu})$, the reference posterior for the quantity of interest $\pi(\theta \,|\, \boldsymbol{x}) = \pi(\theta \,|\, t) \propto \theta^{-1/2} \chi^2(n\, t \,|\, m, n\, \theta)$, where $t = \sum_i \overline{x}_i^2$, and this posterior is shown to have the appropriate consistency properties. For further details see Ferrándiz (1985).

If the μ_i's were known to be related, so that they could be assumed to be exchangeable, with $p(\boldsymbol{\mu}) = \prod_{i=1}^m p(\mu_i \,|\, \boldsymbol{\phi})$, for some $p(\mu \,|\, \boldsymbol{\phi})$, one would have a (very) different (hierarchical) model. Integration of the μ_i's with $p(\boldsymbol{\mu})$ would then produce a model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\phi}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\phi} \in \boldsymbol{\Phi}\}$ parametrized by $\boldsymbol{\phi}$, and only the corresponding reference prior $\pi(\boldsymbol{\phi} \,|\, \mathcal{M})$ would be required. See below (Subsection 3.12) for further discussion on reference priors in hierarchical models.

Far from being specific to Stein's example, the inappropriate behaviour in problems with many parameters of specific marginal posterior distributions derived from multivariate "flat" priors (proper or improper) is very frequent. Thus, as indicated in the introduction, uncritical use of "flat" priors (rather than the relevant reference priors), should be very strongly discouraged.

3.10 Discrete parameters taking an infinity of values

Due to the non-existence of an asymptotic theory comparable to that of the continuous case, the infinite discrete case presents special problems. However, it is often possible to obtain an approximate reference posterior by embedding the discrete parameter space within a continuous one.

Example 20 *Discrete parameters taking an infinite of values.* In the context of capture-recapture models, it is of interest to make inferences about the population size $\theta \in \{1, 2, \ldots\}$ on the basis of data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, which are assumed to consist of a random sample from

$$p(x \,|\, \theta) = \frac{\theta(\theta+1)}{(x+\theta)^2}\,, \quad 0 \le x \le 1.$$

This arises, for instance, in software reliability, when the unknown number θ of bugs is assumed to be a continuous mixture of Poisson distributions. Goudie and Goldie (1981) concluded that, in this problem, all standard non-Bayesian methods are liable to fail; Raftery (1988) finds that, for several plausible "diffuse looking" prior distributions for the discrete parameter θ, the corresponding posterior virtually ignores the data; technically, this is due to the fact that, for most samples, the corresponding likelihood function $p(\boldsymbol{x} \,|\, \theta)$ tends to one (rather than to zero) as $\theta \to \infty$. Embedding the discrete parameter space $\Theta = \{1, 2, \ldots\}$ into the continuous space $\Theta = (0, \infty)$ (since, for each $\theta > 0$, $p(x|\theta)$ is still a probability density for x), and using Theorem 9, the appropriate reference prior is

$$\pi(\theta) \propto i(\theta)^{1/2} \propto (\theta+1)^{-1}\theta^{-1},$$

and it is easily verified that this prior leads to a posterior in which the data are no longer overwhelmed. If the problem requires the use of discrete θ values, the discrete approximation $\Pr(\theta = 1 \,|\, \boldsymbol{x}) = \int_0^{3/2} \pi(\theta \,|\, \boldsymbol{x})\, \mathrm{d}\theta$, and $\Pr(\theta = j \,|\, \boldsymbol{x}) = \int_{j-1/2}^{j+1/2} \pi(\theta \,|\, \boldsymbol{x})\, \mathrm{d}\theta$, $j > 1$, may be used as an approximate discrete reference posterior, specially when interest mostly lies on large θ values, as it is typically the case.

3.11 Behaviour under repeated sampling

The frequentist coverage probabilities of the different types of credible intervals which may be derived from reference posterior distributions are sometimes identical, and usually very close, to their posterior probabilities; this means that, even for moderate samples, an interval with reference posterior probability q may often be interpreted as an *approximate* frequentist confidence interval with significance level $1 - q$.

Example 21 *Coverage in simple normal problems.* Consider again inferences about the mean μ and the variance σ^2 of a normal $\mathrm{N}(x \,|\, \mu, \sigma)$ model. Using the reference prior $\pi^\mu(\mu, \sigma) \propto \sigma^1$ derived in Example 16, the reference posterior distribution of μ after a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ has been observed, $\pi(\mu \,|\, \boldsymbol{x}) \propto \int_0^\infty \prod_{j=1}^n \mathrm{N}(x_j \,|\, \mu, \sigma)\, \pi^\mu(\mu, \sigma)\, \mathrm{d}\sigma$, is the Student density $\pi(\mu \,|\, \boldsymbol{x}) = \mathrm{St}(\mu \,|\, \overline{x}, s/\sqrt{n-1}, n-1) \propto [s^2 + (\overline{x} - \mu)^2]^{-n/2}$, where $\overline{x} = \sum_j x_j/n$, and $s^2 = \sum_j (x_j - \overline{x})^2/n$. Hence, the reference *pos-*

terior of the standardized function of μ, $\phi(\mu) = \sqrt{n-1}\,(\mu - \overline{x})/s$ is standard Student with $n-1$ degrees of freedom. But, conditional on μ, the *sampling* distribution of $t(\boldsymbol{x}) = \sqrt{n-1}\,(\mu - \overline{x})/s$ is *also* standard Student with $n-1$ degrees of freedom. It follows that, for all sample sizes, posterior reference credible intervals for μ will numerically be identical to frequentist confidence intervals based on the sampling distribution of t. Similar results are obtained concerning inferences about σ: the reference posterior distribution of $\psi(\sigma) = ns^2/\sigma^2$ is a χ^2 with $n-1$ degrees of freedom but, conditional on σ, this is also the sampling distribution of $r(\boldsymbol{x}) = ns^2/\sigma^2$.

The *exact* numerical agreement between reference posterior credible intervals and frequentist confidence intervals shown in Example 21 is however the exception, not the norm. Nevertheless, for *large* sample sizes, reference credible intervals are always *approximate* confidence intervals.

More precisely, let data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ consist of n independent observations from $\mathcal{M} = \{p(x \,|\, \theta), x \in \mathcal{X}, \theta \in \Theta\}$, and let $\theta_q(\boldsymbol{x},\, p_\theta)$ denote the q quantile of the posterior $p(\theta \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta)\, p(\theta)$ which corresponds to the prior $p(\theta)$, so that

$$\Pr\left[\theta \leq \theta_q(\boldsymbol{x},\, p_\theta) \,|\, \boldsymbol{x}\right] = \int_{\theta \leq \theta_q(\boldsymbol{x},\, p_\theta)} p(\theta \,|\, \boldsymbol{x})\, \mathrm{d}\theta = q.$$

Standard asymptotic theory may be used to establish that, for any sufficiently regular pair $\{p_\theta,\, \mathcal{M}\}$ of prior p_θ and model $\mathcal{M}$, the *coverage* probability of the region thus defined, $R_q(\boldsymbol{x},\, \theta,\, p_\theta) = \{\boldsymbol{x};\, \theta \leq \theta_q(\boldsymbol{x},\, p_\theta)\}$, converges to q as $n \to \infty$. Specifically, for all sufficiently regular priors,

$$\Pr\left[\theta_q(\boldsymbol{x}, p_\theta) \geq \theta \,|\, \theta\right] = \int_{R_q(\boldsymbol{x},\, \theta,\, p_\theta)} p(\boldsymbol{x} \,|\, \theta)\, \mathrm{d}\boldsymbol{x} = q + O(n^{-1/2}).$$

It has been found however that, when there are no nuisance parameters, the reference prior π^θ typically satisfies

$$\Pr\left[\theta_q(\boldsymbol{x}, \pi^\theta) \geq \theta \,|\, \theta\right] = q + O(n^{-1});$$

this means that the reference prior is often a *probability matching* prior, that is, a prior for which the coverage probabilities of *one-sided* posterior credible intervals are asymptotically closer to their posterior probabilities. Hartigan (1966) showed that the coverage probabilities of *two-sided* Bayesian posterior credible intervals satisfy this type of approximation to $O(n^{-1})$ for *all* sufficiently regular prior functions.

In a pioneering paper, Welch and Peers (1963) established that in the case of the one-parameter regular continuous models Jeffreys prior (which in this case, Theorem 9, is also the reference prior), is the only probability matching prior. Hartigan (1983, p. 79) showed that this result may be extended

to one-parameter discrete models by using continuity corrections. Datta and J. K. Ghosh (1995a) derived a differential equation which provides a necessary and sufficient condition for a prior to be probability matching in the multi-parameter continuous regular case; this has been used to verify that reference priors are typically probability matching priors.

In the nuisance parameter setting, reference priors are sometimes matching priors fro the parameter of interest, but in this general situation, matching priors may not always exist or be unique (Welch, 1965; Ghosh and Mukerjee, 1998). For a review of probability matching priors, see Datta and Sweeting (2005), in this volume.

Although the results described above only justify an *asymptotic* approximate frequentist interpretation of reference posterior probabilities, the coverage probabilities of reference posterior credible intervals derived from *relatively small samples* are also found to be typically close to their posterior probabilities. This is now illustrated within the product of positive normal means problem, already discussed in Example 17.

Example 22 *Product of normal means, continued.* Let available data $\boldsymbol{x} = \{x, y\}$ consist of one observation x from $\mathrm{N}(x \,|\, \alpha, 1)$, $\alpha > 0$, and another observation y from $\mathrm{N}(y \,|\, \beta, 1)$, $\beta > 0$, and suppose that the quantity of interest is the product of the means $\theta = \alpha\,\beta$. The behaviour under repeated sampling of the posteriors which correspond to both the conventional uniform prior $\pi^u(\alpha, \beta) = 1$, and the reference prior $\pi^\theta(\alpha, \beta) = (\alpha^2 + \beta^2)^{1/2}$ (see Example 17) is analyzed by computing the coverage probabilities $\Pr[R_q \,|\, \theta, p_\theta] = \int_{R_q(\boldsymbol{x}, \theta, p_\theta)} p(\boldsymbol{x} \,|\, \theta)\, \mathrm{d}\boldsymbol{x}$ associated to the regions $R_q(\boldsymbol{x}, \theta, p_\theta) = \{\boldsymbol{x};\, \theta \leq \theta_q(\boldsymbol{x}, p_\theta)\}$ defined by their corresponding quantiles, $\theta_q(\boldsymbol{x}, \pi^u)$ and $\theta_q(\boldsymbol{x}, \pi^\theta)$. Table 1 contains the coverage probabilities of the regions defined by the 0.05 posterior quantiles. These have been numerically computed by simulating $4,000$ pairs $\{x, y\}$ from $\mathrm{N}(x \,|\, \alpha, 1)\mathrm{N}(y \,|\, \beta, 1)$ for each of the $\{\alpha, \beta\}$ pairs listed in the first column of the table.

Table 1 *Coverage probabilities of* 0.05*-credible regions for* $\theta = \alpha\,\beta$.

$\{\alpha, \beta\}$	$\Pr[R_{0.05} \,|\, \theta, \pi^u]$	$\Pr[R_{0.05} \,|\, \theta, \pi^\theta]$
$\{1, 1\}$	0.024	0.047
$\{2, 2\}$	0.023	0.035
$\{3, 3\}$	0.028	0.037
$\{4, 4\}$	0.033	0.048
$\{5, 5\}$	0.037	0.046

The standard error of the entries in the table is about 0.0035. It may be observed that the estimated coverages which correspond to the reference prior are appreciably closer to the nominal value 0.05 that those corresponding to the uniform prior. Notice that, although it may be shown that the reference prior *is* probability matching in the technical sense described

above, the empirical results shown in the Table do *not* follow from that fact, for probability matching is an *asymptotic* result, and one is dealing here with samples of size $n = 1$. For further details on this example, see Berger and Bernardo (1989).

3.12 Prediction and hierarchical models

Two classes of problems that are not specifically covered by the methods described above are hierarchical models and prediction problems. The difficulty with these problems is that the distributions of the quantities of interest must belong to specific families of distributions. For instance, if one wants to predict the value of y based on $\boldsymbol{x}$ when $(y, \boldsymbol{x})$ has density $p(y, \boldsymbol{x} \,|\, \boldsymbol{\theta})$, the unknown of interest is y, but its distribution is conditionally specified; thus, one needs a prior for $\boldsymbol{\theta}$, not a prior for y. Likewise, in a hierarchical model with, say, $\{\mu_1, \mu_2, \ldots, \mu_p\}$ being $\mathrm{N}(\mu_i \,|\, \theta, \lambda)$, the μ_i's may be the parameters of interest, but a prior is only needed for the hyperparameters θ and λ.

In hierarchical models, the parameters with conditionally known distributions may be integrated out (which leads to the so-called marginal overdispersion models). A reference prior for the remaining parameters based on this marginal model is then required. The difficulty that arises is how to then identify parameters of interest and nuisance parameters to construct the ordering necessary for applying the reference algorithm, the real parameters of interest having been integrated out.

A possible solution to the problems described above is to define the quantity of interest to be the conditional mean of the original parameter of interest. Thus, in the prediction problem, the quantity of interest could be defined to be $\phi(\boldsymbol{\theta}) = \mathrm{E}[y|\boldsymbol{\theta}]$, which will be either $\boldsymbol{\theta}$ or some transformation thereof, and in the hierarchical model mentioned above the quantity of interest could be defined to be $\mathrm{E}[\mu_i \,|\, \theta, \lambda] = \theta$. More sophisticated choices, in terms of appropriately chosen discrepancy functions, are currently under scrutiny.

Bayesian prediction with objective priors is a very active research area. Pointers to recent suggestions include Kuboki (1998), Eaton and Sudderth (1998, 1999) and Smith (1999). Under appropriate regularity conditions, some of these proposals lead to Jeffreys multivariate prior, $\pi(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}$. However, the use of that prior may lead to rather unappealing predictive posteriors as the following example demonstrates.

Example 23 *Normal prediction.* Let available data consist of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from $\mathrm{N}(x_j \,|\, \mu, \sigma)$, and suppose that one is interested in predicting a new, future observation x from $\mathrm{N}(x \,|\, \mu, \sigma)$. Using the argument described above, the quantity of interest could be defined to be $\phi(\mu, \sigma) = \mathrm{E}[x \,|\, \mu, \sigma] = \mu$ and hence (see Example 16) the appropriate reference prior would be $\pi^x(\mu, \sigma) = \sigma^{-1}$ $(n \geq 2)$. The corresponding joint reference posterior is $\pi(\mu, \sigma \,|\, \boldsymbol{x}) \propto \prod_{j=1}^{n} \mathrm{N}(x_j \,|\, \mu, \sigma)\, \sigma^{-1}$ and the posterior

predictive distribution is

$$\begin{aligned}\pi(x \,|\, \boldsymbol{x}) &= \int_0^\infty \int_{-\infty}^\infty \mathrm{N}(x \,|\, \mu, \sigma)\, \pi(\mu, \sigma \,|\, \boldsymbol{x})\, \mathrm{d}\mu\, \mathrm{d}\sigma \\ &\propto \{(n+1)s^2 + (\overline{x} - \mu)^2)\}^{-n/2}, \\ &\propto \mathrm{St}(x \,|\, \overline{x},\, s\{(n+1)/(n-1)\}^{1/2},\, n-1), \quad n \geq 2 \end{aligned} \tag{38}$$

where, as before, $\overline{x} = n^{-1} \sum_{j=1}^n x_j$ and $s^2 = n^{-1} \sum_{j=1}^n (x_j - \overline{x})^2$. As one would expect, the reference predictive distribution (38) is proper whenever $n \geq 2$: in the absence of prior knowledge, $n = 2$ is the minimum sample size required to identify the two unknown parameters.

It may be verified that the predictive posterior (38) has consistent coverage properties. For instance, with $n = 2$, the reference posterior predictive probability that a third observation lies within the first two is

$$\Pr[x_{(1)} < x < x_{(2)} \,|\, x_1, x_2] = \int_{x_{(1)}}^{x_{(2)}} \pi(x \,|\, x_1, x_2)\, \mathrm{d}x = \tfrac{1}{3}\,,$$

where $x_{(1)} = \min[x_1, x_2]$, and $x_{(2)} = \max[x_1, x_2]$. This is consistent with the fact that, for all μ and σ, the frequentist coverage of the corresponding region of $\mathbb{R}^3$ is precisely

$$\int\int\int_{\{(x_1,x_2,x_3);\ x_{(1)} < x_3 < x_{(2)}\}} \prod_{i=1}^3 \mathrm{N}(x_j \,|\, \mu, \sigma)\, \mathrm{d}x_1\, \mathrm{d}x_2\, \mathrm{d}x_3 = \tfrac{1}{3}\,. \tag{39}$$

In sharp contrast, if Jeffreys multivariate rule $\pi^J(\mu, \sigma) \propto |I(\mu, \sigma)|^{1/2} = \sigma^{-2}$ were used, the posterior predictive would have been a Student t centred at $\overline{x}$, with scale $s\{(n+1)/n\}^{1/2}$, and with n degrees of freedom, which is proper whenever $n \geq 1$. Thus, with $\pi^J(\mu, \sigma)$ as a prior, probabilistic predictions would be possible with only *one* observation, rather unappealing when no prior knowledge is assumed. Moreover, the probability that a third observation lies within the first two which corresponds to the prior $\pi^J(\mu, \sigma)$ is $1/2$, rather than $1/3$, a less than attractive result in view of (39).

For a recent predictive probability matching approach to objective predictive posteriors, see Datta, Mukerjee, Ghosh and Sweeting (2000).

4 Reference Inference Summaries

From a Bayesian viewpoint, the final outcome of a problem of inference about *any* unknown quantity is nothing but the posterior distribution of that quantity. Thus, given some data $\boldsymbol{x}$ and conditions C, *all* that can be said about any function $\boldsymbol{\theta}(\boldsymbol{\omega})$ of the parameters $\boldsymbol{\omega}$ which govern the model is contained in the posterior distribution $p(\boldsymbol{\theta} \,|\, \boldsymbol{x}, C)$, and *all* that can be said about some function $\boldsymbol{y}$ of future observations from the same model is contained in its posterior predictive distribution $p(\boldsymbol{y} \,|\, \boldsymbol{x}, C)$. In fact (Bernardo, 1979a),

Bayesian inference may be described as a decision problem where the space of available actions is the class of those posterior probability distributions of the quantity of interest which are compatible with accepted assumptions.

However, to make it easier for the user to assimilate the appropriate conclusions, it is often convenient to *summarize* the information contained in the posterior distribution, while retaining as much of the information as possible. This is conventionally done by (i) providing values of the quantity of interest which, in the light of the data, are likely to be "close" to its true value, and (ii) measuring the compatibility of the results with hypothetical values of the quantity of interest which might have been suggested in the context of the investigation. In this section, objective Bayesian counterparts to these traditional inference problems of *estimation* and *testing*, which are based on the joint use of intrinsic loss functions and reference analysis, are briefly considered.

4.1 Point Estimation

Let $\boldsymbol{x}$ be the available data, which are assumed to consist of one observation from $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\omega} \in \Omega\}$, and let $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \boldsymbol{\Theta}$ be the quantity of interest. Without loss of generality, the original model $\mathcal{M}$ may be written as $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, in terms of the quantity of interest $\boldsymbol{\theta}$ and a vector $\boldsymbol{\lambda}$ of nuisance parameters. A *point estimate* of $\boldsymbol{\theta}$ is some value $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ which could possibly be regarded as an appropriate proxy for the actual, unknown value of $\boldsymbol{\theta}$.

Formally, to choose a point estimate for $\boldsymbol{\theta}$ is a *decision problem*, where the action space is the class Θ of possible $\boldsymbol{\theta}$ values. From a decision-theoretic perspective, to choose a point estimate $\tilde{\boldsymbol{\theta}}$ of some quantity $\boldsymbol{\theta}$ is a *decision* to act as if $\tilde{\boldsymbol{\theta}}$ were $\boldsymbol{\theta}$, not to assert something about the value of $\boldsymbol{\theta}$ (although desire to assert something simple may well be the main reason to obtain an estimate). To solve this decision problem it is necessary to specify a *loss function* $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$ measuring the consequences of acting *as if* the true value of the quantity of interest were $\tilde{\boldsymbol{\theta}}$, when it is actually $\boldsymbol{\theta}$. The expected posterior loss if $\tilde{\boldsymbol{\theta}}$ were used is $l[\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}] = \int_{\Theta} \ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \, d\boldsymbol{\theta}$, and the *Bayes estimate* is that $\tilde{\boldsymbol{\theta}}$ value which minimizes $l[\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}]$ in $\boldsymbol{\Theta}$. The *Bayes estimator* is the function of the data $\boldsymbol{\theta}^*(\boldsymbol{x}) = \arg\min_{\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}} l[\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}]$.

For any given model and data, the Bayes estimate depends on the chosen loss function. The loss function is context specific, and should generally be chosen in terms of the anticipated uses of the estimate; however, a number of conventional loss functions have been suggested for those situations where no particular uses are envisaged. These loss functions produce estimates which may be regarded as simple descriptions of the *location* of the posterior distribution. For example, if the loss function is quadratic, so that $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^t (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$, then the Bayes estimate is the *posterior mean* $\boldsymbol{\theta}^* = E[\boldsymbol{\theta} \,|\, \boldsymbol{x}]$, assuming that the mean exists. Similarly, if the loss function is a zero-one function, so that

$\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 0$ if $\tilde{\boldsymbol{\theta}}$ belongs to a ball of radius ϵ centred in $\boldsymbol{\theta}$ and $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 1$ otherwise, then the Bayes estimate $\boldsymbol{\theta}^*$ tends to the *posterior mode* as the radius of the ball tends to zero, assuming that a unique mode exists.

If θ is univariate and the loss function is linear, so that $\ell(\tilde{\theta}, \theta) = c_1(\tilde{\theta} - \theta)$ if $\tilde{\theta} \geq \theta$, and $\ell(\tilde{\theta}, \theta) = c_2(\theta - \tilde{\theta})$ otherwise, then the Bayes estimate is the *posterior quantile* of order $c_2/(c_1 + c_2)$, so that $\Pr[\theta < \theta^*] = c_2/(c_1 + c_2)$. In particular, if $c_1 = c_2$, the Bayes estimate is the *posterior median*. The results just described for univariate linear loss functions clearly illustrate the fact that *any* possible parameter value may turn out be the Bayes estimate: it all depends on the loss function describing the consequences of the anticipated uses of the estimate.

Conventional loss functions are typically *not* invariant under reparametrization. As a consequence, the Bayes estimator ϕ^* of a one-to-one transformation $\phi = \phi(\boldsymbol{\theta})$ of the original parameter $\boldsymbol{\theta}$ is not necessarily $\phi(\boldsymbol{\theta}^*)$ (the *univariate* posterior median, which *is* coherent under reparametrization, is an interesting exception). Moreover, conventional loss functions, such as the quadratic loss, focus attention on the discrepancy between the estimate $\tilde{\boldsymbol{\theta}}$ and the true value $\boldsymbol{\theta}$, rather than on the more relevant discrepancy between the statistical *models* they label. The intrinsic discrepancy $\delta\{\mathcal{M}_{\tilde{\boldsymbol{\theta}}},\ p_{\boldsymbol{x}\,|\,\boldsymbol{\theta},\boldsymbol{\lambda}}\}$ (Definition 1) directly measures how different the probability *model* $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})$ is from its closest approximation within the family $\mathcal{M}_{\tilde{\boldsymbol{\theta}}} \equiv \{p(\boldsymbol{x}\,|\,\tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}),\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, and its value does not depend on the particular parametrization chosen to describe the problem.

Definition 8 (Intrinsic estimation) *Let available data $\boldsymbol{x}$ consist of one observation from $\mathcal{M} \equiv \{p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, let $\mathcal{M}_{\tilde{\boldsymbol{\theta}}}$ be the restricted model $\mathcal{M}_{\tilde{\boldsymbol{\theta}}} \equiv \{p(\boldsymbol{x}\,|\,\tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, and let*

$$\delta\{\tilde{\boldsymbol{\theta}},\ (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \delta\{\mathcal{M}_{\tilde{\boldsymbol{\theta}}},\ p_{\boldsymbol{x}\,|\,\boldsymbol{\theta},\boldsymbol{\lambda}}\} = \min_{\tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}}\ \delta\{p(\boldsymbol{x}\,|\,\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}),\ p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})\} \tag{40}$$

be the intrinsic discrepancy between the distribution $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})$ and the set of distributions $\mathcal{M}_{\tilde{\boldsymbol{\theta}}}$. The reference posterior expected intrinsic loss is

$$d(\tilde{\boldsymbol{\theta}}\,|\,\boldsymbol{x}) = \mathrm{E}[\delta\,|\,\boldsymbol{x}] = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \delta\{\tilde{\boldsymbol{\theta}},\ (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, \pi^{\delta}(\boldsymbol{\theta}, \boldsymbol{\lambda}\,|\,\boldsymbol{x})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}, \tag{41}$$

where $\pi^{\delta}(\boldsymbol{\theta}, \boldsymbol{\lambda}\,|\,\boldsymbol{x}) \propto p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})\, \pi^{\delta}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is the reference posterior of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ when δ is the quantity of interest. Given $\boldsymbol{x}$, the intrinsic estimate $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(\boldsymbol{x})$ *is that value $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ which minimizes the posterior reference expected intrinsic loss $d(\tilde{\boldsymbol{\theta}}\,|\,\boldsymbol{x})$. As a function of $\boldsymbol{x}$, $\boldsymbol{\theta}^*(\boldsymbol{x})$ is the* intrinsic estimator *of $\boldsymbol{\theta}$.*

The intrinsic estimate is well defined for any dimensionality, and it is coherent under transformations, in the sense that, if $\boldsymbol{\phi}(\boldsymbol{\theta})$ is a one-to-one function of $\boldsymbol{\theta}$, then the intrinsic estimate $\boldsymbol{\phi}^*$ of $\boldsymbol{\phi}(\boldsymbol{\theta})$ is simply $\boldsymbol{\phi}(\boldsymbol{\theta}^*)$. Under broad regularity conditions (Juárez, 2004), the intrinsic estimator is admissible under the in-

trinsic loss. Moreover, the reference expected intrinsic loss $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x})$ is typically a convex function of $\tilde{\boldsymbol{\theta}}$ in a neighbourhood of its minimum, in which case the intrinsic estimate $\boldsymbol{\theta}^*$ is unique, and it is easily derived by either analytical or numerical methods.

Example 24 *Intrinsic estimation of a Binomial parameter.* Consider estimation of a Binomial proportion θ from r successes given n trials; the reference prior (see Example 12) is $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, the corresponding reference posterior is $\pi(\theta \,|\, n, r) = \mathrm{Be}(\theta \,|\, r + \frac{1}{2}, n - r + \frac{1}{2})$, and the quadratic loss based estimator (the posterior mean) of θ is $\mathrm{E}[\theta \,|\, n, r] = (r + 1/2)/(n + 1)$. However, the quadratic loss based estimator of the log-odds $\phi(\theta) = \log[\theta/(1-\theta)]$, is $\mathrm{E}[\phi \,|\, n, r] = \psi(r + 1/2) - \psi(n - r + 1/2)$ (where $\psi(x) = d \log[\Gamma(x)]/dx$ is the *digamma* function), which is *not* equal to $\phi(\mathrm{E}[\theta \,|\, n, r])$.

On the other hand the intrinsic discrepancy between two Binomial distributions with parameters θ and $\tilde{\theta}$ and the same value of n, the loss to be suffered if $\tilde{\theta}$ were used as a proxy for θ, is $\delta\{\tilde{\theta}, \theta \,|\, n\} = n\, \delta_1\{\tilde{\theta}, \theta\}$, where (see Example 1)

$$\begin{aligned} \delta_1\{\theta_i, \theta_j\} &= \min\{k(\theta_i \,|\, \theta_i), k(\theta_j \,|\, \theta_i)\}, \\ k(\theta_i \,|\, \theta_j) &= \theta_j \log[\,\theta_j/\theta_i\,] + (1-\theta_j) \log[(1-\theta_j)/(1-\theta_i)]. \end{aligned}$$

The intrinsic estimator $\theta^* = \theta^*(r, n)$ is obtained by minimizing the reference expected posterior loss

$$d(\tilde{\theta} \,|\, n, r) = \int_0^1 \delta(\tilde{\theta}, \theta \,|\, n)\, \mathrm{Be}(\theta \,|\, r + \tfrac{1}{2}, n - r + \tfrac{1}{2})\, d\theta. \tag{42}$$

Since intrinsic estimation is coherent under reparametrization, the intrinsic estimator of, say, the log-odds is simply the log-odds of the intrinsic estimator of θ. The exact value of θ^* may be easily obtained by numerical methods, but a very good linear approximation, based on the reference posterior mean of the approximate location parameter (Definition 7) $\phi(\theta) = \int_0^\theta \theta^{-1/2}(1-\theta)^{-1/2}\, \mathrm{d}\theta = \frac{2}{\pi} \arcsin \sqrt{\theta}$, is

$$\theta^*(r, n) \approx \sin^2\{\tfrac{\pi}{2} E[\phi \,|\, r, n]\} \approx (r + \tfrac{1}{3})/(n + \tfrac{2}{3}). \tag{43}$$

The linear approximation (43) remains good even for small samples and extreme r values. For instance, the exact value of the intrinsic estimator with $r = 0$ and $n = 12$ (see Example 28 later in this section) is $\theta^* = 0.02631$, while the approximation yields 0.02632.

Example 25 *Intrinsic estimation of normal variance.* The intrinsic discrepancy $\delta\{p_1, p_2\}$ between two normal densities $p_1(x)$ and $p_2(x)$, with $p_i(x) = \mathrm{N}(x \,|\, \mu_i, \sigma_i)$, is $\delta\{p_1, p_2\} = \min\{k\{p_1 \,|\, p_2\},\ k\{p_2 \,|\, p_1\}\}$, where the

relevant Kullback-Leibler directed divergences are

$$\kappa\{p_i \,|\, p_j\} = \int_{\mathcal{X}} p_j(x) \log \frac{p_j(x)}{p_i(x)} \, \mathrm{d}x = \frac{1}{2}\Big\{ \log \frac{\sigma_i^2}{\sigma_j^2} + \frac{\sigma_j^2}{\sigma_i^2} - 1 + \frac{(\mu_i - \mu_j)^2}{\sigma_i^2} \Big\}.$$

The intrinsic discrepancy between the normal $\mathrm{N}(x \,|\, \mu, \sigma)$ and the set of normals with standard deviation $\tilde{\sigma}$, $\mathcal{M}_{\tilde{\sigma}} \equiv \{\mathrm{N}(x \,|\, \tilde{\mu}, \tilde{\sigma}),\ \tilde{\mu} \in \mathbb{R}\}$ is achieved when $\tilde{\mu} = \mu$, and is found to be

$$\delta\{\mathcal{M}_{\tilde{\sigma}},\ N(x \,|\, \mu, \sigma)\} = \delta(\theta) = \begin{cases} \frac{1}{2}[\log \theta^{-1} + \theta - 1], & \theta < 1 \\ \frac{1}{2}[\log \theta + \theta^{-1} - 1], & \theta \geq 1 \end{cases}$$

which only depends on the ratio $\theta = \tilde{\sigma}^2/\sigma^2$. Since, for any fixed $\tilde{\sigma}$, the intrinsic discrepancy, $\delta\{\tilde{\sigma}, (\mu, \sigma)\} = \delta(\theta)$ is a one-to-one function of σ, the reference prior when δ is the quantity of interest is $\pi^{\delta}(\mu, \sigma) = \sigma^{-1}$, the same as if the quantity of interest were σ (see Example 16). The corresponding posterior distribution of $\theta = \tilde{\sigma}^2/\sigma^2$, after a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of fixed size $n \geq 2$ has been observed, is the gamma density $\pi(\theta \,|\, \boldsymbol{x}) = \mathrm{Ga}(\theta \,|\, (n-1)/2, ns^2/\tilde{\sigma}^2)$, where $s^2 = \sum_j (x_j - \overline{x})^2/n$. The intrinsic estimate of σ is that value σ^* of $\tilde{\sigma}$ which minimizes the expected posterior loss,

$$\int_0^{\infty} \delta(\theta)\, \pi(\theta \,|\, \boldsymbol{x}) \, \mathrm{d}\theta = \int_0^{\infty} \delta(\theta)\, \mathrm{Ga}(\theta \,|\, (n-1)/2, ns^2/\tilde{\sigma}^2) \, \mathrm{d}\theta.$$

The exact value of $\sigma^*(\boldsymbol{x})$ is easily obtained by one-dimensional numerical integration. However, for $n > 2$, a very good approximation is given by

$$\sigma^* = \sqrt{\frac{\sum_j (x_j - \overline{x})^2}{n-2}} \tag{44}$$

which is larger than both the mle estimate s (which divides by n the sum of squares) and the squared root of the conventional unbiased estimate of the variance (which divides by $n-1$). A good approximation for $n = 2$ is $\sigma^* = (\sqrt{5}/2)|x_1 - x_2|$. Since intrinsic estimation is coherent under one-to-one reparametrizations, the intrinsic estimator of the variance is $(\sigma^*)^2$, and the intrinsic estimator of, say, $\log \sigma$ is simply $\log \sigma^*$.

Intrinsic estimation is a very powerful, general procedure for objective, invariant point estimation. For further discussion, see Bernardo and Juárez (2003).

4.2 Region (interval) estimation

To describe the inferential content of the posterior distribution $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ of the quantity of interest it is often convenient to quote regions $R \subset \boldsymbol{\Theta}$ of given (posterior) probability under $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$. Any subset of the parameter space

$R_q \subset \boldsymbol{\Theta}$ such that $\int_{R_q} \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta} = q$, $0 < q < 1$, so that, given data $\boldsymbol{x}$, the true value of $\boldsymbol{\theta}$ belongs to R_q with probability q, is said to be a (posterior) *q-credible region* of $\boldsymbol{\theta}$. Credible regions are coherent under reparametrization; thus, for any q-credible region R_q of $\boldsymbol{\theta}$ a one-to-one transformation $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$, $\boldsymbol{\phi}(R_q)$ is a q-credible region of $\boldsymbol{\phi}$. However, for any given q there are generally infinitely many credible regions.

Sometimes, credible regions are selected to have minimum size (length, area, volume), resulting in *highest probability density* (HPD) regions, where all points in the region have larger probability density than all points outside. However, HPD regions are *not* coherent under reparametrization: the image $\boldsymbol{\phi}(R_q)$ of an HPD q-credible region R_q will be a q-credible region for $\boldsymbol{\phi}$, but will not generally be HPD; indeed, there is no compelling reason to restrict attention to HPD credible regions. In one dimension, posterior quantiles are often used to derive credible regions. Thus, if $\theta_q = \theta_q(\boldsymbol{x})$ is the $100q\%$ posterior quantile of θ, then $R_q^l = \{\theta;\, \theta \leq \theta_q\}$ is a one-sided, typically unique q-credible region, and it is coherent under reparametrization. *Probability centred* q-credible regions of the form $R_q^c = \{\theta;\, \theta_{(1-q)/2} \leq \theta \leq \theta_{(1+q)/2}\}$ are easier to compute, and are often quoted in preference to HPD regions. However, centred credible regions are only really appealing when the posterior density has a unique interior mode, and have a crucial limitation: they are not uniquely defined in problems with more than one dimension.

For reasonable loss functions, a typically unique credible region may be selected as a *lowest posterior loss* (LPL) region, where all points in the region have smaller (posterior) expected loss than all points outside.

Definition 9 (Intrinsic credible region) *Let available data $\boldsymbol{x}$ consist of one observation from $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, let $\mathcal{M}_{\tilde{\boldsymbol{\theta}}}$ be the restricted model $\mathcal{M}_{\tilde{\boldsymbol{\theta}}} \equiv \{p(\boldsymbol{x} \,|\, \tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ and let $\delta\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ be the intrinsic discrepancy between the distribution $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ and the set $\mathcal{M}_{\tilde{\boldsymbol{\theta}},}$. An* intrinsic q-credible region $R_q^* = R_q^*(\boldsymbol{x}) \subset \boldsymbol{\Theta}$ *is a subset of the parameter space $\boldsymbol{\Theta}$ such that,*

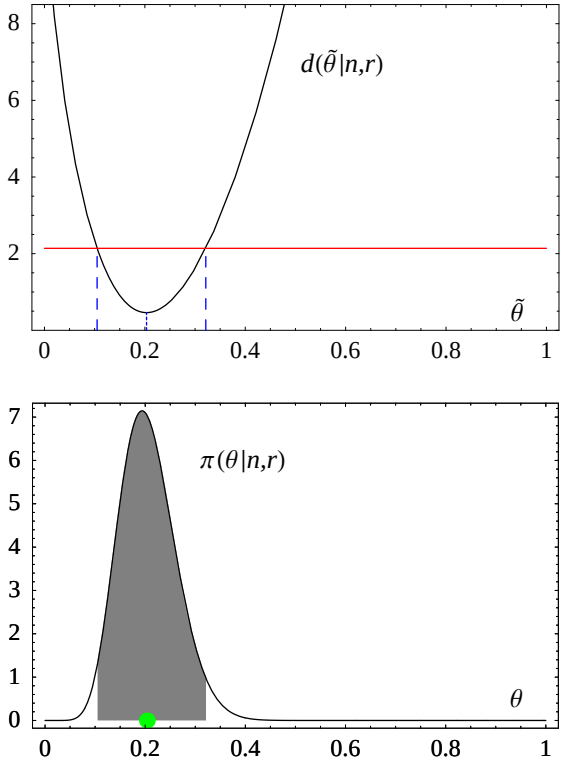
$$\int_{R_q^*(\boldsymbol{x})} \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta} = q, \quad \forall \boldsymbol{\theta}_i \in R_q^*(\boldsymbol{x}),\ \forall \boldsymbol{\theta}_j \notin R_q^*(\boldsymbol{x}),\ d(\tilde{\boldsymbol{\theta}}_i \,|\, \boldsymbol{x}) \leq d(\tilde{\boldsymbol{\theta}}_j \,|\, \boldsymbol{x}),$$

where, as before, $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}) = \mathrm{E}[\delta \,|\, \boldsymbol{x}] = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \delta\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, \pi^{\delta}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}$ is the reference posterior expected intrinsic loss.

Intrinsic credible regions are well defined for any dimensionality, and they are coherent under one-to-one transformations, in the sense that, if $\boldsymbol{\phi}\{\boldsymbol{\theta}\}$ is a one-to-one transformation of $\boldsymbol{\theta}$ and $R_q^* \subset \boldsymbol{\Theta}$ is an intrinsic q-credible region for $\boldsymbol{\theta}$, then $\boldsymbol{\phi}\{R_q^*\} \subset \boldsymbol{\Phi}$ is an intrinsic q-credible region for $\boldsymbol{\phi}$. As mentioned above, the reference expected intrinsic loss $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x})$ is often a convex function of $\tilde{\boldsymbol{\theta}}$; in that case, for each $q \in (0, 1)$ there is a unique (convex) intrinsic q-credible region.

Example 26 *Intrinsic Binomial credible regions.* Let r be the number of successes observed in n independent Bernoulli trials with parameter θ.

Figure 4 *Intrinsic* 0.95*-credible region for a Binomial parameter.*



As described in Example 24, the reference posterior expected intrinsic loss which corresponds to using $\tilde{\theta}$ instead of the actual (unknown) θ is the convex function $d\{\tilde{\theta}\,|\,n,r\}$ of Equation (42), which is represented in the upper panel of Figure 4 as a function of $\tilde{\theta}$, for $r = 10$ and $n = 50$. Using the invariance of the intrinsic loss with respect to one-to-one transformations, and a normal approximation to the posterior distribution of the approximate location parameter $\phi(\theta) = 2\arcsin\sqrt{\theta}$, it is found that

$$d\{\tilde{\theta}\,|\,n,r\} \approx \tfrac{1}{2} + 2\,n\left(\arcsin\sqrt{\tilde{\theta}} - \arcsin\sqrt{(r+\alpha_n)/(n+2\alpha_n)}\right)^2,$$

where $\alpha_n = (n+4)/(4n+10)$, rapidly concerging to $\frac{1}{4}$. A lowest posterior loss (LDL) q-credible region consists of the set of $\tilde{\theta}$ points with posterior probability q and minimum expected loss. In this problem, the intrinsic q-credible region $R^*_q(r,n)$, is therefore obtained as the interval $R^*_q(r,n) = [\theta_a(r,n),\, \theta_b(r,n)]$ defined by the solution (θ_a, θ_b) to the system

$$\Big\{d\{\theta_a\,|\,n,r\} = d\{\theta_b\,|\,n,r\}, \quad \int_{\theta_a}^{\theta_b} \pi(\theta\,|\,n,r)\,\mathrm{d}\theta = q\Big\}.$$

In particular, the intrinsic 0.95-credible region is the set of $\tilde{\theta}$ points with

posterior expected loss smaller than 2.139 (shaded region in the lower panel of Figure 4), which is $R^*_{0.95} = \{\tilde{\theta};\ 0.105 \leq \tilde{\theta} \leq 0.321]$. Notice that this is neither a HPD interval nor a centred interval. The point with minimum expected loss is the intrinsic estimator, $\theta^* = 0.2034$. Since intrinsic estimation is coherent under one-to-one reparametrizations, the intrinsic estimator and the 0.95-intrinsic credible region of the log-odds, $\psi = \psi(\theta) = \log[\theta/(1-\theta)]$ are immediately derived as $\psi(\theta^*) = -1.365$ and $\psi(R^*_{0.95}) = [-2.144,\ -0.747]$.

It may be argued that, in practice, it is reasonable for credible regions to give privilege to the most probable values of the parameters, as HPD regions do. This is obviously incompatible with an invariance requirement, but it is interesting to notice that, in one-parameter problems, intrinsic credible regions are approximately HPD in the approximate location parametrization. Thus, in Example 26, the 0.95-credible region for the approximate location parameter, $\phi(\theta) = \frac{2}{\pi} \arcsin \sqrt{\theta}$, $\phi(R^*_{0.95}) = [0.210, 0.384]$, is nearly an HPD interval for ϕ.

4.3 Hypothesis Testing

Let $\boldsymbol{x}$ be the available data, which are assumed to consist of one observation from model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, parametrized in terms of the vector of interest $\boldsymbol{\theta}$ and a vector $\boldsymbol{\lambda}$ of nuisance parameters. The posterior distribution $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ of the quantity of interest $\boldsymbol{\theta}$ conveys immediate intuitive information on the values of $\boldsymbol{\theta}$ which, given $\mathcal{M}$, might be declared to be *compatible* with the observed data $\boldsymbol{x}$, namely, those with a relatively high probability density. Sometimes, a *restriction*, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$, of the possible values of the quantity of interest (where $\boldsymbol{\Theta}_0$ may possibly consist of a single value $\boldsymbol{\theta}_0$) is suggested in the course of the investigation as deserving special consideration, either because restricting $\boldsymbol{\theta}$ to $\boldsymbol{\Theta}_0$ would greatly simplify the model, or because there are additional, context specific arguments suggesting that $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. Intuitively, the (null) *hypothesis* $H_0 \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}$ should be judged to be *compatible* with the observed data $\boldsymbol{x}$ if there are elements in $\boldsymbol{\Theta}_0$ with a relatively high posterior density. However, a more precise conclusion is typically required and this is made possible by adopting a decision-oriented approach. Formally, testing the hypothesis $H_0 \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}$ is a *decision problem* where the action space $\mathcal{A} = \{a_0, a_1\}$ only contains two elements: to accept (a_0) or to reject (a_1) the proposed restriction.

To solve this decision problem, it is necessary to specify an appropriate loss function, $\ell(a_i, \boldsymbol{\theta})$, measuring the consequences of accepting or rejecting H_0 as a function of the actual value $\boldsymbol{\theta}$ of the vector of interest. Notice that this requires the statement of an *alternative* 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 the null, $\int_{\boldsymbol{\Theta}} \ell(a_0, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}$, is larger than the expected posterior loss of

rejecting, $\int_{\boldsymbol{\Theta}} \ell(a_1, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}$, that is, if (and only if)

$$\int_{\boldsymbol{\Theta}} [\ell(a_0, \boldsymbol{\theta}) - \ell(a_1, \boldsymbol{\theta})]\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}} \Delta\ell(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta} > 0. \tag{45}$$

Therefore, only the loss difference $\Delta\ell(\boldsymbol{\Theta}_0, \boldsymbol{\theta}) = \ell(a_0, \boldsymbol{\theta}) - \ell(a_1, \boldsymbol{\theta})$, which measures the *advantage* of rejecting $H_0 \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}$ as a function of $\boldsymbol{\theta}$, has to be specified: the hypothesis H_0 should be rejected whenever the expected advantage of rejecting is positive.

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* H_0 were true, *i.e.*, as if $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, but there are at least two options for the alternative action a_1. This may either mean (i) the *negation* of H_0, that is to act as if $\boldsymbol{\theta} \notin \boldsymbol{\Theta}_0$ or, alternatively, it may rather mean (ii) to reject the simplification implied by H_0 and to keep the unrestricted model, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, which is true by assumption. Both options have been analyzed in the literature, although it may be argued that the problems of scientific data analysis, where hypothesis testing procedures are typically used, are better described by the second alternative. Indeed, an established model, identified by $H_0 \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}$, is often embedded into a more general model, $\{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}\}$, constructed to include promising departures from H_0, and it is then required to verify whether presently available data $\boldsymbol{x}$ are still compatible with $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, or whether the extension to $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is really required.

The simplest loss structure has, for all values of the nuisance parameter vector $\boldsymbol{\lambda}$, a zero-one form, with $\{\ell(a_0, \boldsymbol{\theta}) = 0, \ell(a_1, \boldsymbol{\theta}) = 1\}$ if $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, and $\{\ell(a_0, \boldsymbol{\theta}) = 1, \ell(a_1, \boldsymbol{\theta}) = 0\}$ if $\boldsymbol{\theta} \notin \boldsymbol{\Theta}_0$, so that the *advantage* $\Delta\ell\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ of rejecting H_0 is

$$\Delta\ell\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \begin{cases} 1, \text{ if } \boldsymbol{\theta} \notin \boldsymbol{\Theta}_0 \\ -1, \text{ if } \boldsymbol{\theta} \in \boldsymbol{\Theta}_0. \end{cases} \tag{46}$$

With this (rather naïve) loss function it is immediately found that the optimal action is to reject H_0 if (and only if) $\Pr(\boldsymbol{\theta} \notin \boldsymbol{\Theta}_0 \,|\, \boldsymbol{x}) > \Pr(\boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \,|\, \boldsymbol{x})$. Notice that this formulation *requires* that $\Pr(\boldsymbol{\theta} \in \boldsymbol{\Theta}_0) > 0$, that is, that the (null) hypothesis H_0 has a strictly positive prior probability. If $\boldsymbol{\theta}$ is a continuous parameter and $\boldsymbol{\Theta}_0$ has zero measure (for instance if H_0 consists of a single point $\boldsymbol{\theta}_0$), this requires the use of a non-regular "sharp" prior concentrating a positive probability mass on $\boldsymbol{\theta}_0$. With no mention to the loss structure implicit behind, this solution was early advocated by Jeffreys (1961, Ch. 5). However, this is known to lead to the difficulties associated to Lindley's paradox (Lindley, 1957; Bartlett, 1957; Bernardo, 1980; Robert, 1993; Brewer, 2002).

The intrinsic discrepancy loss may also be used to provide an attractive general alternative to Bayesian hypothesis testing, the *Bayesian reference cri-*

terion, BRC (Bernardo, 1999a; Bernardo and Rueda, 2002). This follows from assuming that the loss structure is such that

$$\Delta\ell\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \delta\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - d^*, \quad d^* > 0, \tag{47}$$

where $\delta\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, which describes as a function of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ the loss suffered by assuming that $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, is the intrinsic discrepancy between the distribution $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ and the set $\mathcal{M}_0 \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}), \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$. The function $\delta\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, which is invariant under one-to-one reparametrization, is non-negative and it is zero if, and only if, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. The constant d^* is the (strictly positive) advantage of being able to work with the null model when it is true, measured in the same units as δ; the choice of d^*, in terms of posterior expected log-likelihood ratios, is discussed below.

Definition 10 (Intrinsic hypothesis testing: BRC) *Let available data $\boldsymbol{x}$ consist of one observation from $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, let $\mathcal{M}_0$ be the restricted model $\mathcal{M}_0 \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}), \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ and let $\delta\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ be the intrinsic discrepancy between the distribution $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ and the set $\mathcal{M}_0$. The* Bayesian reference criterion (BRC) *rejects model $\mathcal{M}_0$ if the* intrinsic statistic *$d(\boldsymbol{\Theta}_0 \,|\, \boldsymbol{x})$, defined as the reference posterior expected intrinsic loss, exceeds a critical value d^*. Formally, the null hypothesis $H_0 \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}$ is rejected if, and only if,*

$$d(\boldsymbol{\Theta}_0 \,|\, \boldsymbol{x}) = \mathrm{E}[\delta \,|\, \boldsymbol{x}] = \int_{\boldsymbol{\Theta}} \delta\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, \pi^{\delta}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda} > d^*,$$

where $\pi^{\delta}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})\, \pi^{\delta}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is the reference posterior of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ when $\delta = \delta\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ is the quantity of interest. The conventional value $d^* = \log(100)$ may be used for scientific communication.

As the sample size increases, the expected value of $d(\boldsymbol{\Theta}_0 \,|\, \boldsymbol{x})$ under sampling tends to one when H_0 is true, and tends to infinity otherwise; thus $d(\boldsymbol{\Theta}_0 \,|\, \boldsymbol{x})$ may be regarded as a continuous, positive measure of the expected loss (in information units) from simplifying the model by accepting $\mathcal{M}_0$. In traditional language, $d(\boldsymbol{\Theta}_0 \,|\, \boldsymbol{x})$ is a test statistic, and the BRC criterion rejects the null if this *intrinsic test statistic* $d(\boldsymbol{\Theta}_0 \,|\, \boldsymbol{x})$ exceeds some *critical value* d^*. However, in sharp contrast to frequentist hypothesis testing, the critical value d^* is simply a utility constant which measures the number of *information units* which the decision maker is prepared to loose in order to be able to work with the null model H_0, *not* a function of sampling properties of the model.

The interpretation of the intrinsic discrepancy in terms of the minimum posterior expected likelihood ratio in favour of the true model (see Section 2) provides a direct *calibration* of the required critical value. Indeed, $d(\boldsymbol{\Theta}_0 \,|\, \boldsymbol{x})$ is the minimum posterior expected log-likelihood ratio in favour of the unrestricted model. For instance, values around $\log[10] \approx 2.3$ should be regarded as mild evidence against H_0, while values around $\log[100] \approx 4.6$ suggest strong evidence against the null, and values larger than $\log[1000] \approx 6.9$ may be safely

used to reject H_0. Notice that, in contrast to frequentist hypothesis testing, where it is hazily recommended to adjust the significance level for dimensionality and sample size, the intrinsic statistic in measured on an absolute scale which remains valid for *any* sample size and *any* dimensionality.

Example 27 *Testing the value of a normal mean.* Let data consist of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from a normal $\mathrm{N}(x \,|\, \mu, \sigma)$ distribution, and consider the "canonical" problem of testing whether or not these data are compatible with some specific sharp hypothesis $H_0 \equiv \{\mu = \mu_0\}$ on the value of the mean. The intrinsic discrepancy is easily found to be

$$\delta(\mu_0, \mu \,|\, \sigma) = \frac{n}{2}\left(\frac{\mu - \mu_0}{\sigma}\right)^2, \tag{48}$$

a simple transformation of the standardized distance between μ and μ_0, which generalizes to $\delta(\boldsymbol{\mu}_0, \boldsymbol{\mu}) = (n/2)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^t \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)$, a linear function of the Mahalanobis distance, in the multivariate normal case.

Consider first the case where σ is assumed to be known. The reference prior for μ is then uniform; this is also the reference prior when the parameter of interest is δ, since $\delta(\mu_0, \mu)$ is a piecewise invertible function of μ (see Theorem 6). The corresponding posterior distribution, is $\pi(\mu \,|\, \boldsymbol{x}) = \mathrm{N}(\mu \,|\, \overline{x}, \sigma/\sqrt{n})$, $(n \geq 1)$. The expected value of $\delta(\mu_0, \mu)$ with respect to this posterior yields the corresponding intrinsic statistic,

$$d(\mu_0 \,|\, \boldsymbol{x}) = \tfrac{1}{2}(1 + z^2), \quad z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \tag{49}$$

a simple function of the standardized distance between the sample mean $\overline{x}$ and μ_0. As prescribed by the general theory, the expected value of $d(\mu_0, \,|\, \boldsymbol{x})$ under repeated sampling is one if $\mu = \mu_0$, and increases linearly with n otherwise. In this canonical example, to reject H_0 whenever $|z| > 1.96$ (the frequentist suggestion with the conventional 0.05 significance level), corresponds to rejecting H_0 whenever $d(\mu_0 \,|\, \boldsymbol{x})$ is larger than 2.42, a rather weak evidence, since this means that the posterior expected likelihood ratio against H_0 is only about $\exp[2.42] = 11.25$. Conversely, to reject whenever posterior expected likelihood ratio against H_0 is about 100, so that $d^* = \log[100] \approx 4.6$, is to reject whenever $|z| > 2.86$, which is close to the conventional 3σ rule often used by engineers. The extreme 6σ rule, apparently popular these days, would correspond (under normality) to $d^* = 18.5 \approx \log[10^8]$.

If the scale parameter σ is also unknown, the intrinsic discrepancy is

$$\delta\{\mu_0, (\mu, \sigma)\} = \frac{n}{2} \log\left[1 + \left(\frac{\mu - \mu_0}{\sigma}\right)^2\right], \tag{50}$$

which is *not* the same as (48). The intrinsic test statistic $d(\mu_0, \boldsymbol{x})$ may then be found as the expected value of $\delta\{\mu_0, (\mu, \sigma)\}$ under the corresponding

joint reference posterior distribution $\pi^{\delta}(\mu, \sigma \,|\, \boldsymbol{x})$ when δ is the quantity of interest. After some algebra, the exact result may be expressed in terms of hypergeometric functions (Bernardo, 1999a), but is very well approximated by the simple function

$$d(\mu_0 \,|\, \boldsymbol{x}) \approx \frac{1}{2} + \frac{n}{2} \log\left(1 + \frac{t^2}{n}\right), \tag{51}$$

where t is the conventional statistic $t = \sqrt{n-1}\,(\overline{x} - \mu_0)/s$, written here in terms of the sample variance $s^2 = \sum_j (x_j - \overline{x})^2/n$. For instance, for samples sizes 5, 30 and 1000, and using the threshold $d^* = \log[100]$, the null hypothesis $H_0 \equiv \{\mu = \mu_0\}$ would be rejected whenever $|t|$ is respectively larger than 4.564, 3.073, and 2.871.

Example 28 *A lady tasting tea.* A lady claims that by tasting a cup of tea made with milk she can discriminate whether milk has been poured over the tea infusion or the other way round, and she is able to give the correct answer in n consecutive trials. Are these results compatible with the hypothesis that she is only guessing and has been lucky? The example, a variation suggested by Neyman (1950, Sec. 5.2) to a problem originally proposed by Fisher (1935, Sec. 2.5), has often been used to compare alternative approaches to hypothesis testing. See Lindley (1984) for a subjectivist Bayesian analysis.

The intrinsic objective Bayesian solution is immediate from the results in Examples 24 and 26. Indeed, using Definition 10, if data are assumed to consist of n Bernoulli observations and r successes have been observed, the intrinsic statistic to test the precise null $\theta = \theta_0$ is

$$d(\theta_0 \,|\, r, n) = \int_0^1 \delta\{\theta_0, \theta \,|\, n\}\, \mathrm{Be}(\theta \,|\, r + \tfrac{1}{2}, n - r + \tfrac{1}{2})\, \mathrm{d}\theta,$$

where $\delta\{\theta_0, \theta \,|\, n\}$ is given by (7). In this case, one has $r = n$ and $\theta_0 = \frac{1}{2}$. For the values $n = 8$, $n = 10$ and $n = 12$ traditionally discussed, the intrinsic test statistic, $d(\theta_0 \,|\, r, n)$, respectively yields the values $d(\frac{1}{2} \,|\, 8, 8) \approx 4.15$, $d(\frac{1}{2} \,|\, 10, 10) \approx 5.41$ and $d(\frac{1}{2} \,|\, 12, 12) \approx 6.70$. Since $\log[100] \approx 4.61$, the hypothesis of pure guessing would not be rejected with $n = 8$ with the conventional threshold $d^* = \log[100]$, but would be rejected with $n = 10$ successes (and *a fortiori* with $n = 12$). Actually, the value of $d(\frac{1}{2} \,|\, 8, 8)$ says that the observed data are only estimated to be about $\exp[4.15] \approx 64$ times more likely under the unrestricted model (unknown θ) that under the null model (no discrimination power, $\theta = \theta_0 = \frac{1}{2}$). However, with $n = 10$ and $n = 12$ the observed data are respectively estimated to be about 224 and 811 times more likely under the unrestricted model that under the null.

The Bayesian reference criterion may also be used with non-nested problems. Thus, given two alternative models for $\boldsymbol{x} \in \mathcal{X}$, $\mathcal{M}_1 = \{p_1(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1), \boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1\}$

and $\mathcal{M}_2 = \{p_2(\boldsymbol{x} \,|\, \boldsymbol{\theta}_2),\, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_2\}$, one may introduce the a new parameter α to define a mixture model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_1, \alpha) = p_1(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1)^{\alpha}\, p_2(\boldsymbol{x} \,|\, \boldsymbol{\theta}_2)^{1-\alpha}$ (with either a continuous $\alpha \in [0, 1]$ or, more simply, a discrete $\alpha \in \{0, 1\}$), and use BRC to verify whether $\mathcal{M}_1$, or $\mathcal{M}_2$, or both, are compatible with the data, assuming the mixture is. For further discussion on hypothesis testing and the development of the Bayesian reference criterion see Bernardo (1982, 1985a, 1999a), Bernardo and Bayarri (1985), Rueda (1992) and Bernardo and Rueda (2002).

5 Further Reading

Reference analysis already has a long history, but it still is a very active area of research. The original paper on reference analysis, (Bernardo, 1979b), is easily read and it is followed by a very lively discussion; Bernardo (1981), extends the theory to general decision problems; see also Bernardo and Smith (1994, Sec. 5.4.1) and Rabena (1998). Berger and Bernardo (1989, 1992c) contain crucial mathematical extensions. Bernardo (1997) is a non-technical analysis, in a dialogue format, of the foundational issues involved, and it is followed by a discussion. A textbook level description of reference analysis is provided in Bernardo and Smith (1994, Sec. 5.4); Bernardo and Ramón (1998) contains a simple introduction to reference distributions. BRC, the Bayesian reference criterion for hypothesis testing, was introduced by Bernardo (1999a) and further refined in Bernardo and Rueda (2002). Intrinsic estimation was introduced in Bernardo and Juárez (2003). Berger, Bernardo and Sun (2005) contains the last mathematical developments of reference theory at the moment of writing.

Papers which contain either specific derivations or new applications of reference analysis include, in chronological order of the first related paper by the same author(s), Bernardo (1977a,b, 1978, 1980, 1982, 1985a,b, 1999b), Bayarri (1981, 1985), Ferrándiz (1982, 1985), Sendra (1982), Eaves (1983a,b, 1985), Armero (1985), Bernardo and Bayarri (1985), Chang and Villegas (1986), Chang and Eaves (1990), Hills (1987), Mendoza (1987, 1988, 1990), Bernardo and Girón (1988), Lindley (1988), Berger and Bernardo (1989, 1992a,b,c), Clarke and Barron (1990, 1994), Polson and Wasserman (1990), Phillips (1991), Severini (1991, 1993, 1995, 1999), Ye and Berger (1991), Ghosh and Mukerjee (1992), Singh and Upadhyay (1992), Stephens and Smith (1992), Berger and Sun (1993), Clarke and Wasserman (1993), Dey and Peng (1993, 1995), Kuboki (1993, 1998), Liseo (1993, 2003, 2005), Ye (1993, 1994, 1995, 1998), Berger and Yang (1994), Kubokawa and Robert (1994), Sun (1994, 1997), Sun and Berger (1994, 1998), Yang and Berger (1994, 1997), Datta and J. K. Ghosh (1995a,b), Datta and M. Ghosh (1995a); Datta and M. Ghosh (1995b), Giudici (1995), Ghosh, Carlin and Srivastava (1995), du Plessis, van der Merwe and Groenewald (1995), Sun and Ye (1995, 1996, 1999), de Waal, Groenewald and Kemp (1995), Yang and Chen (1995), Bernard (1996), Clarke (1996), Ghosh and Yang (1996), Armero and Bayarri (1997), Fernández, Osiewalski and Steel

(1997), Garvan and Ghosh (1997, 1999), Ghosal and Samanta (1997), Ghosal (1997, 1999), Sugiura and Ishibayashi (1997), Berger, Philippe and Robert (1998), Bernardo and Ramón (1998), Chung and Dey (1998, 2002), Scholl (1998), Sun, Ghosh and Basu (1998), Philippe and Robert (1998), Berger, Liseo and Wolpert (1999), Burch and Harris (1999), Brewer (1999), Scricciolo (1999), Vernotte and Zalamansky (1999), Yuan and Clarke (1999), Berger, Pericchi and Varshavsky (1998), Lee (1998), Fernández and Steel (1998b, 1999a,b, 2000), Mendoza and Gutiérrez-Peña (1999), Mukerjee and Reid (1999, 2001), Aguilar and West (2000), Eno and Ye (2000, 2001), Elhor and Pensky (2000), Fernández and Steel (2000), Kim, Kang and Cho (2000), van der Linde (2000), Berger, de Oliveira and Sansó (2001), Fan (2001), Ghosh and Kim (2001), Ghosh, Rousseau and Kim (2001), Kim, Chang and Kang (1961), Kim, Kang and Lee (2001, 2002), Komaki (2001, 2004), Natarajan (2001), Rosa and Gianola (2001), Aslam (2002a,b), Daniels (2002), Datta, Ghosh and Kim (2002), Millar (2002), Philippe and Rousseau (2002), Pretorius and van der Merwe (2002), Tardella (2002), Consonni and Veronese (2003), Datta and Smith (2003), Fraser, Reid, Wong and Yi (2003), Ghosh and Heo (2003a,b), Ghosh, Yin and Kim (2003), Gutiérrez-Peña and Rueda (2003), He (2003), Leucari and Consonni (2003), Lauretto, Pereira, Stern and Zacks (2003), Madruga, Pereira and Stern (2003), Ni and Sun (2003), Sareen (2003), Consonni, Veronese, and Gutiérrez-Peña (2004), Sun and Ni (2004), Grünwald and Dawid (2004), Roverato and Consonni (2004), Stern (2004a,b), van der Merwe and Chikobvu (2004) and Liseo and Loperfido (2005).

This chapter concentrates on reference analysis. It is known, however, that ostensibly different approaches to the derivation of objective priors often produce the same result, a testimony of the robustness of many solutions to the definition of what an appropriate objective prior may be in a particular problem. Many authors have proposed alternative objective priors (often comparing the resulting inferences with those obtained within the frequentist paradigm), either as general methods or as *ad hoc* solutions to specific inferential problems, and a few are openly critical with objective Bayesian methods. Many relevant papers in this very active field of Bayesian mathematical statistics are listed in the references section below. For reviews of many of these, see Dawid (1983), Bernardo and Smith (1994, Sec. 5.6.2), Kass and Wasserman (1996) and Bernardo (1997).

6 Acknowledgements

The author is indebted to many colleagues for helpful comments on an earlier draft of this paper. Special recognition is due, however, to the very detailed comments by my *maestro*, Professor Dennis Lindley.

Research supported by grant BMF 2002-2889 of the former *Ministerio de Ciencia y Tecnología*, Madrid, Spain.

References

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*. Cambridge: University Press.

de Alba, E. and Mendoza, M. (1996). A discrete model for Bayesian forecasting with stable seasonal patterns. *Advances in Econometrics II* (R. Carter Hill, ed.) New York: JAI Press, 267–281.

Agliari, A., and Calvi-Parisetti, C. (1988). A g-reference informative prior: A note on Zellner's g-prior. *The Statistician* **37**, 271–275.

Aguilar, O. and West, M. (2000). Bayesian Dynamic factor models and portfolio allocation. *J. Business Econ. Studies* **18**, 338–357.

Akaike, H. (1977). On Entropy Maximization Principle. *Applications of Statistics* (P. R. Krishnaiah, ed.) Amsterdam: North-Holland, 27-41.

Akaike, H. (1978). A new look at the Bayes procedure. *Biometrika* **65**, 53–59.

Akaike, H. (1980). The interpretation of improper prior distributions as limits of data dependent proper prior distributions. *J. Roy. Statist. Soc. B* **45**, 46–52.

Akaike, H. (1980b). Likelihood and the Bayes procedure. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) Valencia: University Press, 144–166 and 185–203 (with discussion).

Akaike, H. (1980c). Ignorance prior distribution of a hyperparameter and Stein's estimator. *Ann. Inst. Statist. Math.* **32**, 171–178.

Akaike, H. (1983). On minimum information prior distributions. *Ann. Inst. Statist. Math.* **35**, 139–149.

Armero, C. (1985). Bayesian analysis of M/M/1/∞/Fifo Queues. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 613–618.

Armero, C. and Bayarri, M. J. (1997). A Bayesian analysis of a queuing system with unlimited service. *J. Statist. Planning and Inference* **58**, 241–261.

Aslam, M. (2002a). Bayesian analysis for paired comparison models allowing ties and not allowing ties. *Pakistan J. Statist.* **18**, 53–69.

Aslam, M. (2002b). Reference Prior for the Parameters of the Rao-Kupper Model. *Proc. Pakistan Acad. Sci.* **39**, 215–224

Atwood, C. L. (1996). Constrained noninformative priors in risk assessment. *Reliability Eng. System Safety* **53**, 37–46.

Banerjee, A. K. and Bhattacharyya G. K. (1979). Bayesian results for the inverse Gaussian distribution with an application. *Technometrics* **21**, 247–251.

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

Barnard, G. A. (1988). The future of statistics: Teaching and research. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) Oxford: University Press, 17–24.

Barron, A. R. (1999). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 27–52, (with discussion).

Barron, A., Rissanen, J. and Yu, B. (1998). The Minimum Description Length principle in coding and modelling. *IEEE Trans. Information Theory* **44**, 2743–2760.

Bartlett, M. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika* **44**, 533–534.

Bayarri, M. J. (1981). Inferencia Bayesiana sobre el coeficiente de correlación de una población normal bivariante. *Trab. Estadist.* **32**, 18–31.

Bayarri, M. J. (1985). Bayesian inference on the parameters of a Beta distribution. *Statistics and Decisions* **2**, 17–22.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. Published posthumously in *Phil. Trans. Roy. Soc. London* **53**, 370–418 and **54**, 296–325. Reprinted in *Biometrika* **45** (1958), 293–315, with a biographical note by G. A. Barnard.

Belzile, E. and Angers, J.-F. (1995). Choice of noninformative priors for the variance components of an unbalanced one-way random model. *Comm. Statist. Theory and Methods* **24** 1325–1341.

Berger, J. O. (2000). Bayesian analysis: A look at today and thoughts of tomorrow. *J. Amer. Statist. Assoc.* **95**, 1269–1276.

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

Berger, J. O. and Bernardo, J. M. (1992a). Ordered group reference priors with applications to a multinomial problem. *Biometrika* **79**, 25–37.

Berger, J. O. and Bernardo, J. M. (1992b). Reference priors in a variance components problem. *Bayesian Analysis in Statistics and Econometrics* (P. K. Goel and N. S. Iyengar, eds.) Berlin: Springer, 323–340.

Berger, J. O. and Bernardo, J. M. (1992c). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 35–60 (with discussion).

Berger, J. O., Bernardo, J. M. and Mendoza, M. (1989). On priors that maximize expected information. *Recent Developments in Statistics and their Applications* (J. P. Klein and J. C. Lee, eds.). Seoul: Freedom Academy, 1–20.

Berger, J. O., Bernardo, J. M. and Sun, D. (2005). Reference priors from first principles: A general definition. *Tech. Rep.*, SAMSI, NC, USA.

Berger, J. O., de Oliveira, V. and Sansó, B. (2001). Objective Bayesian analysis of spatially correlated data. *J. Amer. Statist. Assoc.* **96**, 1361–1374.

Berger, J. O., Liseo, B. and Wolpert, R. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.* **14**, 1–28.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91**, 109–122.

Berger, J. O., Pericchi, L. R. and Varshavsky, J. A. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā A* **60**, 307–321.

Berger, J. O., Philippe, A. and Robert, C. (1998). Estimation of quadratic functions: noninformative priors for noncentrality parameters. *Statistica Sinica* **8**, 359–376.

Berger, J. O. and Robert, C. P. (1990). Subjective hierarchical Bayes estimation of a multivariate mean: On the frequentist interface. *Ann. Statist.* **18**, 617–651.

Berger, J. O. and Strawderman, W. E. (1996). Choice of hierarchical priors: Admissibility in estimation of normal means. *Ann. Statist.* **24**, 931–951.

Berger, J. O. and Sun, D. (1993). Bayesian analysis for the poly-Weibull distribution. *J. Amer. Statist. Assoc.* **88**, 1412–1418.

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle* (2nd ed.). Hayward, CA: IMS.

Berger, J. O. and Yang, R. (1994). Noninformative priors and Bayesian testing for the AR(1) model. *Econometric Theory* **10**, 461–482.

Berger, R. L. (1979). Gamma minimax robustness of Bayes rules. *Comm. Statist. Theory and Methods* **8**, 543–560.

Bernard, J.-M. (1996). Bayesian interpretation of frequentist procedures for a Bernoulli process. *Amer. Statist.* **50**, 7–13.

Bernardo, J. M. (1977a). Inferences about the ratio of normal means: a Bayesian approach to the Fieller-Creasy problem. *Recent Developments in Statistics* (J. R. Barra, F. Brodeau, G. Romier and B. van Cutsem eds.). Amsterdam: North-Holland, 345–349.

Bernardo, J. M. (1977b). Inferencia Bayesiana sobre el coeficiente de variación: una solución a la paradoja de marginalización. *Trab. Estadist.* **28**, 23-80.

Bernardo, J. M. (1978). Unacceptable implications of the left Haar measure in a standard normal theory inference problem *Trab. Estadist.* **29**, 3–9.

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

Bernardo, J. M. (1979b). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113–147 (with discussion). Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.) Brookfield, VT: Edward Elgar, 1995, 229–263.

Bernardo, J. M. (1980). A Bayesian analysis of classical hypothesis testing. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) Valencia: University Press, 605–647 (with discussion).

Bernardo, J. M. (1981). Reference decisions. *Symposia Mathematica* **25**, 85–94.

Bernardo, J. M. (1982). Contraste de modelos probabilísticos desde una perspectiva Bayesiana. *Trab. Estadist.* **33**, 16–30.

Bernardo, J. M. (1985a). Anlisis Bayesiano de los contrastes de hipótesis paramtricos. *Trab. Estadist.* **36**, 45–54.

Bernardo, J. M. (1985b). On a famous problem of induction. *Trab. Estadist.* **36**, 24–30.

Bernardo, J. M. (1997). Non-informative priors do not exist. *J. Statist. Planning and Inference* **65**, 159–189 (with discussion).

Bernardo, J. M. (1999a). Nested hypothesis testing: The Bayesian reference criterion. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 101–130 (with discussion).

Bernardo, J. M. (1999b). Model-free objective Bayesian prediction. *Rev. Acad. Ciencias de Madrid* **93**, 295–302.

Bernardo, J. M. and Bayarri, M. J. (1985). Bayesian model criticism. *Model Choice* (J.-P. Florens, M. Mouchart, J.-P. Raoult and L. Simar, eds.) Brussels: Pub. Fac. Univ. Saint Louis, 43–59.

Bernardo, J. M. and Girón F. J. (1988). A Bayesian analysis of simple mixture problems. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) Oxford: University Press, 67–88 (with discussion).

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

Bernardo, J. M. and Ramón, J. M. (1998). An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters. *The Statistician* **47**, 1–35.

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

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

Blyth, S. and Smith, A. F. M. (1998). Bayesian meta-analysis of frequentist p-values. *Comm. Statist. Theory and Methods* **27**, 2707–2724.

Box, G. E. P. and Cox D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. B* **26**, 211–252 (with discussion).

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

Boulton, D. M. and Wallace, C. S. (1970). A program for numerical classification. *The Computer J.* 13, 63–69.

Brewer, K. R. W. (1999). Testing a precise null hypothesis using reference posterior odds. *Rev. Acad. Ciencias de Madrid* **93**, 303–310.

Brewer, K. R. W. (2002). The Lindley and Bartlett paradoxes. *Pak./ J./ Statist.* **18**, 1–13.

Brown, L. D., Cai, T. T. and DasGupta, A. (2001). Interval estimation of a Binomial proportion. *Statist. Sci.* **16**, 101-133, (with discussion).

Brown, L. D., Cai, T. T. and DasGupta, A. (2002). Confidence intervals for a Binomial proportion and asymptotic expansions. *Ann. Statist.* **30**, 160–201.

Burch, B. D. and Harris, I. R. (1999). Bayesian estimators of the intraclass correlation coefficient in the one-way random effects model. *Comm. Statist. Theory and Methods* **28**, 1247–1272.

Casella, G. (1996). Statistical inference and Monte Carlo algorithms. *Test* **5**, 249–344 (with discussion).

Casella, G. and Hwang, J. T. (1987). Employing vague prior information in the construction of confidence sets. *J. Multivariate Analysis* **21**, 79–104.

Chaloner, K. (1987). A Bayesian approach to the estimation of variance components for the unbalanced one way random model. *Technometrics* **29**, 323–337.

Chang, T. and Eaves, D. M. (1990). Reference priors for the orbit of a group model. *Ann. Statist.* **18**, 1595–1614.

Chang, T. and Villegas, C. (1986). On a theorem of Stein relating Bayesian and classical inferences in group models. *Can. J. Statist.* **14**, 289–296.

Chao, J. C. and Phillips, P. C. B. (1998). Posterior distributions in limited information analysis of the simultaneous equations model using the Jeffreys prior. *J. Econometrics* **87**, 49–86.

Chao, J. C. and Phillips, P. C. B. (2002). Jeffreys prior analysis of the simultaneous equations model in the case with $n+1$ endogenous variables. *J. Econometrics* **111**, 251–283.

Chen, M. H., Ibrahim, J. G. and Shao, Q. M. (2000). Power prior distributions for generalized linear models. *J. Statist. Planning and Inference* **84**, 121-137.

Cho, J. S. and Baek, S. U. (2002). Development of matching priors for the ratio of the failure rate in the Burr model. *Far East J. Theor. Stat.* **8**, 79–87.

Chung, Y. and Dey, D. K. (1998). Bayesian approach to estimation in intraclass correlation using reference priors. *Comm. Statist. Theory and Methods* **27**, 2241–2255.

Chung, Y. and Dey, D. K. (2002). Model determination for the variance component model using reference priors. *J. Statist. Planning and Inference* **100**, 49–65.

Cifarelli, D. M. and Regazzini, E. (1987). Priors for exponential families which maximize the association between past and future observations. *Probability and Bayesian Statistics* (R. Viertl, ed.) London: Plenum, 83–95.

Cover, M. and Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley

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

Crowder, M. (1992). Bayesian priors based on a parameter transformation using the distribution function.*Ann. Inst. Statist. Math.* **44**, 405–416.

Crowder, M. and Sweeting, T. (1989). Bayesian inference for a bivariate Binomial distribution. *Biometrika* **76**, 599–603.

Csiszár, I. (1985). An extended maximum entropy principle and a Bayesian justification. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 83–98, (with discussion).

Csiszár, I. (1991). Why least squared and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Statist.* **19**, 2032–2066.

Clarke, B. (1996). Implications of reference priors for prior information and for sample size. *J. Amer. Statist. Assoc.* **91**, 173–184.

Clarke, B. and Barron, A. R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Information Theory* **36**, 453–471.

Clarke, B. and Barron, A. R. (1994). Jeffreys' prior is asymptotically least favourable under entropy risk. *J. Statist. Planning and Inference* **41**, 37–60.

Clarke, B. and Sun, D. (1997). Reference priors under the chi-squared distance. *Sankhyā A* **59**, 215–231.

Clarke, B., and Wasserman, L. (1993). Noninformative priors and nuisance parameters, *J. Amer. Statist. Assoc.* **88**, 1427–1432.

Consonni, G. and Veronese, P. (1989a). Some remarks on the use of improper priors for the analysis of exponential regression problems. *Biometrika* **76**, 101–106.

Consonni, G. and Veronese, P. (1989b). A note on coherent invariant distributions as non-informative priors for exponential and location-scale families. *Comm. Statist. Theory and Methods* **18**, 2883–2907.

Consonni, G. and Veronese, P. (1992). Bayes factors for linear models and improper priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 587–594.

Consonni, G. and Veronese, P. (1993). Unbiased Bayes estimates and improper priors. *Ann. Inst. Statist. Math.* **45**, 303–315.

Consonni, G. and Veronese, P. (2003). Enriched conjugate and reference priors for the Wishart family on symmetric cones. *Ann. Statist.* **31**, 1491–1516.

Consonni, G., Veronese, P. and Gutiérrez-Peña, E. (2004). Reference priors for exponential families with simple quadratic variance function. *J. Multivariate Analysis* **88**, 335–364.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. B* **49**, 1–39 (with discussion).

Cornfield, J. (1969). The Bayesian outlook and its application. *Biometrics* **25**, 617–657, (with discussion).

Daniels, M. J. (1999). A prior for the variance in hierarchical models. *Can. J. Statist.* **27**, 567–578.

Daniels, M. J. (2005). A class of shrinkage priors for the dependence structure in longitudinal data. *J. Statist. Planning and Inference* **127**, 119-130.

Daniels, M. J. and Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* **89**, 553–566.

DasGupta, A. (1985). Bayes minimax estimation in multiparameter families when the parameter space is restricted to a bounded convex set. *Sankhyā A* **47**, 326–332.

Datta, G. S. (1996). On priors providing frequentist validity for Bayesian inference of multiple parametric functions. *Biometrika* **83**, 287–298.

Datta, G. S. and Ghosh, J. K. (1995a). On priors providing a frequentist validity for Bayesian inference. *Biometrika* **82**, 37–45.

Datta, G. S. and Ghosh, J. K. (1995b). Noninformative priors for maximal invariant parameter in group models. *Test* **4**, 95–114.

Datta, G. S. and Ghosh, M. (1995a). Some remarks on noninformative priors. *J. Amer. Statist. Assoc.* **90**, 1357–1363.

Datta, G. S. and Ghosh, M. (1995b). Hierarchical Bayes estimators of the error variance in one-way ANOVA models. *J. Statist. Planning and Inference* **45**, 399–411.

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

Datta, G. S., Ghosh, M. and Kim, Y.-H. (2002). Probability matching priors for one-way unbalanced random effect models. *Statistics and Decisions* **20**, 29–51.

Datta, G. S., Ghosh, M. and Mukerjee, R. (2000). Some new results on probability matching priors. *Calcutta Statist. Assoc. Bull.* **50**, 179–192. Corr: **51**, 125.

Datta, G. S., and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*. Berlin: Springer,

Datta, G. S., Mukerjee, R., Ghosh, M. and Sweeting, T. J. (2000). Bayesian prediction with approximate frequentist validity . *Ann. Statist.* **28**, 1414–1426.

Datta, G. S. and Smith, D. D. (2003). On property of posterior distributions of variance components in small area estimation. *J. Statist. Planning and Inference* **112**, 175–183.

Datta, G. S. and Sweeting, T. (2005). Probability matching priors. *In this volume.*

Dawid, A. P. (1980). A Bayesian look at nuisance parameters. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) Valencia: University Press, 167-203, (with discussion).

Dawid, A. P. (1983). Invariant prior distributions. *Encyclopedia of Statistical Sciences* **4** (S. Kotz, N. L. Johnson and C. B. Read, eds.) New York: Wiley, 228–236.

Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference. *J. Roy. Statist. Soc. B* **53**, 79—109 (with discussion).

Dawid, A. P. and Stone, M. (1972). Expectation consistency of inverse probability distributions. *Biometrika* **59**, 486–489.

Dawid, A. P. and Stone, M. (1973). Expectation consistency and generalized Bayes inference. *Ann. Statist.* **1**, 478–485.

Dawid, A. P., Stone, M. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. B* **35**, 189–233 (with discussion).

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

Dey, D. K. and Peng, F. (1993). On the choice of prior for the Bayes estimation in accelerated life testing.*J. Statist. Computation and Simulation* **48**, 207–217.

Dey, D. K. and Peng, F. (1995). Elimination of nuisance parameters using information measure. *Parisankhyan Samikkha* **2**, 9–29.

Delampady, M., DasGupta, A., Casella, G. Rubin, H. and Strawderman, W. E. (2001). A new approach to default priors and robust Bayesian methodology. *Can. J. Statist.* **29**, 437–450.

Diaconis P. and Freedman, D. A. (1998). Consistency of Bayes estimates for nonparametric regression: Normal theory. *Bernoulli* **4**, 411–444.

DiCiccio, T. J. and Martin, M. A. (1991). Approximations of marginal tail probabilities for a class of smooth functions with applications to bayesian and conditional inference. *Biometrika* **78**, 891–902.

DiCiccio, T. J. and Stern, S. E. (1994). Frequentist and Bayesian Bartlett correction of test statistics based on adjusted profile likelihood. *J. Roy. Statist. Soc. B* **56**, 397–408.

Eaton, M. L. (1982). A method for evaluating improper prior distributions. *Statistical Decision Theory and Related Topics III* **1** (S. S. Gupta and J. O. Berger, eds.) New York: Academic Press,

Eaton, M. L. (1992). A statistical diptych: admissible inferences, recurrence of symmetric Markov chains. *Ann. Statist.* **20**, 1147–1179.

Eaton, M. L. and Freedman, D. A. (2004). Dutch book against some 'objective' priors. *Bernoulli* **10**, 861–872.

Eaton, M. L., Sudderth, W. D. (1998). A new predictive distribution for normal multivariate linear models. *Sankhyā A* **60**, 363–382.

Eaton, M. L., Sudderth, W. D. (1999). Consistency and strong inconsistency of group-invariant predictive inferences. *Bernoulli* **5**, 833–854.

Eaton, M. L., Sudderth, W. D. (2002). Group invariant inference and right Haar measure. *J. Statist. Planning and Inference* **103**, 87–99.

Eaton, M. L., Sudderth, W. D. (2004). Properties of right Haar predictive inference. *Sankhyā A* **66**, 487–512.

Eaves, D. M. (1983a). On Bayesian nonlinear regression with an enzyme example. *Biometrika* **70**, 373-379.

Eaves, D. M. (1983b). Minimally informative prior analysis of a non-linear model. *The Statistician* **32**, 117.

Eaves, D. M. (1985). On maximizing the missing information about a hypothesis. *J. Roy. Statist. Soc. B* **47**, 263-266.

Eaves, D. and Chang, T.(1992). Priors for ordered conditional variance and vector partial correlation *J. Multivariate Analysis* **41**, 43–55.

Efron, B. (1986). Why isn't everyone a Bayesian? *Amer. Statist.* **40**, 1–11, (with discussion).

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80**, 3–26.

Elhor, A. and Pensky, M. (2000). Bayesian estimation of location of lightning events. *Sankhyā B* **62**, 202–206.

Eno, D. R. and Ye, K. (2000). Bayesian reference prior analysis for polynomial calibration models. *Test* **9**, 191–202.

Eno, D. R. and Ye, K. (2001). Probability matching priors for an extended statistical calibration problem. *Can. J. Statist.* **29**, 19–35.

Erickson, T. (1989). Proper posteriors from improper priors for an unidentified error-in-variables model. *Econometrica* **57**, 1299-1316.

Evans, I. G. and Nigm A. M. (1980). Bayesian prediction for two parameter Weibull lifetime models. *Comm. Statist. Theory and Methods* **9**, 649–658.

Everson, P. J. and Bradlow, E. T. (2002). Bayesian inference for the Beta-Binomial distribution via polynomial expansions. *J. Comp. Graphical Statist.* **11**, 202–207.

Fan, T.-H. (2001). Noninformative Bayesian estimation for the optimum in a single factor quadratic response model. *Test* **10**, 225–240.

Fan, T.-H. and Berger, J. O. (1992). Behaviour of the posterior distribution and inferences for a normal mean with t prior distributions.*Statistics and Decisions* **10**, 99–120.

Fatti, L. P. (1982). Predictive discrimination under the random effect model. *South African Statist. J.* **16**, 55–77.

Fernández, C., Osiewalski, J. and Steel, M. (1997). On the use of panel data in stochastic frontier models with improper priors.*J. Econometrics* **79**, 169–193.

Fernández, C. and Steel, M. (1998a). Reference priors for the non-normal two-sample problems. *Test* **7**, 179–205.

Fernández, C. and Steel, M. (1998b) On Bayesian modelling of fat tails and skewness. *J. Amer. Statist. Assoc.* **93**, 359–371.

Fernández, C. and Steel, M. (1999b). On the dangers of modelling through continuous distributions: A bayesian perspective. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 213–238, /diss

Fernández, C. and Steel, M. (1999b). Reference priors for the general location-scale model. *Statistics and Probability Letters* **43**, 377–384.

Fernández, C. and Steel, M. (2000). Bayesian regression analysis with scale mixtures of normals. *J. Economic Theory* **16**, 80–101.

Ferrándiz, J. R. (1982). Una solución Bayesiana a la paradoja de Stein. *Trab. Estadist.* **33**, 31–46.

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

de Finetti, B. (1970). *Teoria delle Probabilit* **1**. Turin: Einaudi. English translation as *Theory of Probability* **1** in 1974, Chichester: Wiley.

Fisher, R. A. (1935). *The Design of Experiments.* Edinburgh: Oliver and Boyd.

Florens, J.-P. (1978). Mesures à priori et invariance dans une expérience Bayésienne. *Pub. Inst. Statist. Univ. Paris* **23**, 29–55.

Florens, J.-P. (1982). Expériences Bayésiennes invariantes. *Ann. Inst. M. Poincaré* **18**, 309–317.

Fraser, D. A. S., McDunnough, P. and Taback, N. (1997). Improper priors, posterior asymptotic normality, and conditional inference. *Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz* (N. L. Johnson and N. Balakrishnan, eds.) New York: Wiley, 563–569.

Fraser, D. A. S., Monette, G., and Ng, K. W. (1985). Marginalization, likelihood and structural models, *Multivariate Analysis* **6** (P. R. Krishnaiah, ed.) Amsterdam: North-Holland, 209–217.

Fraser, D. A. S., Reid, N. (1996). Bayes posteriors for scalar interest parameters. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 581–585.

Fraser, D. A. S. and Reid, N. (2002). Strong matching of frequentist and Bayesian parametric inference. *J. Statist. Planning and Inference* **103**, 263–285.

Fraser, D. A. S., Reid, N., Wong, A. and Yi, G. Y. (2003). Direct Bayes for interest parameters . *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press, 529–534.

Fraser, D. A. S., Reid, N. and Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249–264.

Fraser, D. A. S. and Yi, G. Y. (2002). Location reparametrization and default priors for statistical analysis. *J. Iranian Statist. Soc.* **1**, 55–78.

Freedman, D. A. (1966). A note on mutual singularity of priors. *Ann. Math. Statist.* **37**, 375–381.

Freedman, D. A. (1995). Some issues in the foundation of statistics. *Topics in the Foundation of Statistics* (B. C. van Fraassen, ed.) Dordrecht: Kluwer19–83. (with discussion).

Garvan, C. W. and Ghosh, M. (1997). Noninformative priors for dispersion models. *Biometrika* **84**, 976–982.

Garvan, C. W. and Ghosh, M. (1999). On the property of posteriors for dispersion models. *J. Statist. Planning and Inference* **78**, 229–241.

Gatsonis, C. A. (1984). Deriving posterior distributions for a location parameter: A decision theoretic approach. *Ann. Statist.* **12**, 958–970.

Geisser, S. (1965). Bayesian estimation in multivariate analysis. *Ann. Math. Statist.* **36**, 150–159.

Geisser, S. (1979). In discussion of Bernardo (1979b). *J. Roy. Statist. Soc. B* **41**, 136–137.

Geisser, S. (1980). A predictivist primer. *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.) Amsterdam: North-Holland, 363–381.

Geisser, S. (1984). On prior distributions for binary trials. *J. Amer. Statist. Assoc.* **38**, 244–251 (with discussion).

Geisser, S. (1993). *Predictive inference: An introduction.* London: Chapman and Hall

Geisser, S. and Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *J. Roy. Statist. Soc. B* **25**, 368–376.

George, E. I. and McCulloch, R. (1993). On obtaining invariant prior distributions. *J. Statist. Planning and Inference* **37**, 169–179.

Ghosal, S. (1997). Reference priors in multiparameter nonregular cases. *Test* **6**, 159–186.

Ghosal, S. (1999). Probability matching priors for non-regular cases. *Biometrika* **86**, 956–964.

Ghosal, S. and Samanta, T. (1997). Expansion of Bayes risk for entropy loss and reference prior in nonregular cases. *Statistics and Decisions* **15**, 129–140.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1997). Non-informative priors via sieves and packing numbers. *Advances in Decision Theory and Applications* (S. Panchpakesan and N. Balakrisnan, eds.) Boston: Birkhauser, 119–132.

Ghosh, J. K. and Mukerjee, R. (1991). Characterization of priors under which Bayesian and frequentist Bartlett corrections are equivalent in the multivariate case. *J. Multivariate Analysis* **38**, 385–393.

Ghosh, J. K. and Mukerjee, R. (1992). Non-informative priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 195–210 (with discussion).

Ghosh, J. K. and Mukerjee, R. (1993a). Frequentist validity of highest posterior density regions in the multiparameter case. *Ann. Math. Statist.* **45**, 293–302; corr: 602.

Ghosh, J. K. and Mukerjee, R. (1993b). On priors that match posterior and frequentist distribution functions. *Can. J. Statist.* **21**, 89–96.

Ghosh, J. K. and Mukerjee, R. (1995a). Frequentist validity of highest posterior density regions in the presence of nuisance parameters. *Statistics and Decisions* **13**, 131–139.

Ghosh, J. K. and Mukerjee, R. (1995b). On perturbed ellipsoidal and highest posterior density regions with approximate frequentist validity. *J. Roy. Statist. Soc. B* **57**, 761–769.

Ghosh, J. K. and Samanta, T. (2002). Nonsubjective Bayes testing – An overview. *J. Statist. Planning and Inference* **103**, 205–223.

Ghosh, M., Carlin, B. P. and Srivastava, M. S. (1995). Probability matching priors for linear calibration. *Test* **4**, 333–357.

Ghosh, M., Chen, M.-H., Ghosh, A. and Agresti, A. (2000a). Hierarchical Bayesian analysis of binary matched pairs data. *Statistica Sinica* **10**, 647–657.

Ghosh, M., Ghosh, A., Chen, M.-H. and Agresti, A. (2000b). Noninformative priors for one-parameter item response models. *J. Statist. Planning and Inference* **88**, 99–115.

Ghosh M. and Heo J. (2003). Default Bayesian priors for regression models with first-order autoregressive residuals. *J. Time Series Analysis* **24**, 269–282.

Ghosh, M., and Heo J. (2003). Noninformative priors, credible sets and bayesian hypothesis testing for the intraclass model. *J. Statist. Planning and Inference* **112**, 133–146.

Ghosh, M., and Meeden G. (1997). *Bayesian Methods for Finite Population Sampling* London: Chapman and Hall.

Ghosh, M. and Mukerjee, R. (1998). Recent developments on probability matching priors. *Applied Statistical Science III* (S. E. Ahmed, M. Ashanullah and B. K. Sinha eds.) New York: Science Publishers, 227–252.

Ghosh, M., Rousseau, J. and Kim, D. H. (2001). Noninformative priors for the bivariate Fieller-Creasy problem. *Statistics and Decisions* **19**, 277–288.

Ghosh, M. and Yang, M.-Ch. (1996). Non-informative priors for the two sample normal problem. *Test* **5**, 145–157.

Ghosh, M. and Kim, Y.-H. (2001). The Behrens-Fisher problem revisited: a Bayes-frequentist synthesis. *Can. J. Statist.* **29**, 5–17.

Ghosh, M., Yin, M. and Kim, Y.-H. (2003). Objective Bayesian inference for ratios of regression coefficients in linear models. *Statistica Sinica* **13**, 409–422.

Giudici, P. (1995). Bayes factors for zero partial covariances. *J. Statist. Planning and Inference* **46**, 161–174.

Gokhale, D. V. and Press, S. J. (1982). Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution. *J. Roy. Statist. Soc. A* **145**, 237–249.

Good, I. J. (1952). Rational decisions. *J. Roy. Statist. Soc. B* **14**, 107–114.

Good, I. J. (1969). What is the use of a distribution? *Multivariate Analysis* **2** (P. R. Krishnaiah, ed.) New York: Academic Press, 183–203.

Good, I. J. (1981). Flexible priors for estimating a normal distribution. *J. Statist. Computation and Simulation* **13**, 151–153.

Good, I. J. (1986). Invariant hyperpriors. *J. Statist. Computation and Simulation* **24**, 323-324.

Goudie, I. B. J. and Goldie, C. M. (1981). Initial size estimation for the pure death process. *Biometrika* **68**. 543–550.

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Ann. Statist.* **32**, 1367-1433.

Gutiérrez-Peña, E. (1992). Expected logarithmic divergence for exponential families. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 669–674.

Gutiérrez-Peña, E. and Muliere, P. (2004). Conjugate priors represent strong pre-experimental assumptions. *Scandinavian J. Statist.* **31**, 235–246.

Gutiérrez-Peña, E. and Rueda, R. (2003). Reference priors for exponential families. *J. Statist. Planning and Inference* **110**, 35–54.

Hadjicostas, P. (1998). Improper and proper posteriors with improper priors in a hierarchical model with a beta-Binomial likelihood. *Comm. Statist. Theory and Methods* **27**, 1905–1914.

Hadjicostas, P. and Berry, S. M. (1999). Improper and proper posteriors with improper priors in a Poisson-gamma hierarchical model. *Test* **8**, 147–166.

Haldane, J. B. S. (1948). The precision of observed values of small frequencies. *Biometrika* **35**, 297–303.

Hartigan, J. A. (1964). Invariant prior distributions. *Ann. Math. Statist.* **35**, 836–845.

Hartigan, J. A. (1965). The asymptotically unbiased prior distribution. *Ann. Math. Statist.* **36**, 1137–1152.

Hartigan, J. A. (1966). Note on the confidence prior of Welch and Peers. *J. Roy. Statist. Soc. B* **28**, 55-56.

Hartigan, J. A. (1971). Similarity and probability. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) Toronto: Holt, Rinehart and Winston, 305–313 (with discussion).

Hartigan, J. A. (1983). *Bayes Theory.* Berlin: Springer.

Hartigan, J. A. (1996). Locally uniform prior distributions. *Ann. Statist.* **24**, 160–173.

Hartigan, J. A. (1998). The maximum likelihood prior. *Ann. Statist.* **26**, 2083–2103.

Hartigan, J. A. (2004). Uniform priors on convex sets improve risk. *Statistics and Probability Letters* **67**, 285–288.

Hartigan, J. A. and Murphy, T. B. (2002). Inferred probabilities. *J. Statist. Planning and Inference* **105**, 23–34.

He, C. Z. (2003). Bayesian modelling of age-specific survival in bird nesting studies under irregular visits. *Biometrics* **59**, 962–973.

Heath, D. and Sudderth, W. (1978). On finitely additive priors, coherence, and extended admissibility. *Ann. Statist.* **6**, 333–345.

Heath, D. and Sudderth, W. (1989). Coherent inference from improper priors and from finitely additive priors. *Ann. Statist.* **17**, 907–919.

Hill, B. M. (1965). Inference about variance components in the one-way model. *J. Amer. Statist. Assoc.* **60**, 806–825.

Hill, S. D. and Spall, J. C. (1988). Shannon information-theoretic priors for state-space model parameters. *Bayesian Analysis of Time series and Dynamic Models* (J. C. Spall, ed.) New York: Marcel Dekker, 509–524.

Hill, S. D. and Spall, J. C. (1994). Sensitivity of a Bayesian analysis to the prior distribution. *IEEE Trans. Systems, Man and Cybernetics* **24**, 216–221.

Hills, S. E. (1987). Reference priors and identifiability problems in non-linear models. *The Statistician* **36**, 235–240.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Amer. Statist. Assoc.* **91**, 1461–1473.

Hobert, J. P. and Casella, G. (1996). Functional compatibility, markov chains and Gibbs sampling with improper posteriors. *J. Comp. Graphical Statist.* **7**, 42-60.

Hobert, J. P., Marchev, D. and Schweinsberg, J. (2004). Stability of the tail markov chain and the evaluation of improper priors for an exponential rate parameter. *Bernoulli* **10**, 549–564.

Howlader, H. A. and Weiss, G. (1988). Bayesian reliability estimation of a two parameter Cauchy distribution. *Biom. J.* **30**, 329–337.

Ibrahim, J. G. (1997). On properties of predictive priors in linear models. *Amer. Statist.* **51**, 333–337.

Ibrahim, J. G. and Chen, M.-H. (1998). Prior distributions and Bayesian computation for proportional hazards models. *Sankhy$\overline{a}$ B* **60**, 48–64.

Ibrahim, J. G. and Laud, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffreys' prior. *J. Amer. Statist. Assoc.* **86**, 981–986.

Ibragimov, I. A. and Khasminskii, R. Z. (1973). On the information in a sample about a parameter. *Proc. 2nd Internat. Symp. Information Theory.* (B. N. Petrov and F. Csaki, eds.), Budapest: Akademiaikiadó, 295–309.

Inaba, T. (1984). Non-informative prior distribution in a simultaneous equations model. *J. Japan Statist. Soc.* **14**, 93–103.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Trans. Systems, Science and Cybernetics* **4**, 227–291.

Jaynes, E. T. (1976). Confidence intervals vs. Bayesian intervals. *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science* **2** (W. L. Harper and C. A. Hooker eds.). Dordrecht: Reidel, 175–257 (with discussion).

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proc. of the IEEE* **70**, 939–952.

Jaynes, E. T. (1985). Some random observations. *Synthèse* **3**, 115–138.

Jaynes, E. T. (1989). *Papers on Probability, Statistics and Statistical Physics*, 2nd ed. Dordrecht: Kluwer.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A* **186**, 453–461.

Jeffreys, H. (1955). The present position in probability theory. *Brit. J. Philos. Sci.* **5**, 275–289.

Jeffreys, H. (1961). *Theory of Probability* (Third edition). Oxford: University Press.

Joshi, S. and Shah, M. (1991). Estimating the mean of an inverse Gaussian distribution with known coefficient of variation. *Comm. Statist. Theory and Methods* **20**, 2907–2912.

Juárez, M. (2004). *Métodos Bayesianos Objetivos de Estimación y Contraste de Hipótesis*. Ph.D. Thesis, Universitat de València, Spain.

Kashyap, R. L. (1971). Prior probability and uncertainty. *IEEE Trans. Information Theory* **14**, 641–650.

Kappenman, R. F., Geisser, S. and Antle, C. F. (1970). Bayesian and fiducial solutions to the Fieller Creasy problem. *Sankhyā B* **25**, 331–330.

Kass, R. E. (1989). The geometry of asymptotic inference. *Statist. Sci.* **4**, 188–235, (with discussion).

Kass, R. E. (1990). Data-translated likelihood and Jeffreys' rule. *Biometrika* **77**, 107–114.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91**, 1343–1370. Corr: 1998 **93**, 412.

Keyes, T. K. and Levy, M. S. (1996). Goodness of prediction fit for multivariate linear models. *J. Amer. Statist. Assoc.* **91**, 191–197.

Kim, B. H., Chang, I. H. and Kang, C. K. (2001). Bayesian estimation for the reliability in Weibull stress-strength systems using noninformative priors. *Far East J. Theor. Stat.* **5**, 299–315.

Kim, D. H., Kang, S. G. and Cho, J. S. (2000). Noninformative Priors for Stress-Strength System in the Burr-Type X Model *J. Korean Statist. Soc.* **29**, 17–28.

Kim, D. H., Kang, S. G. and Lee, W. D. (2001). Noninformative priors for intraclass correlation coefficient in familial data. *Far East J. Theor. Stat.* **5**, 51–65.

Kim, D. H., Kang, S. G. and Lee, W. D. (2002). Noninformative priors for the power law process. *J. Korean Statist. Soc.* **31**, 17–31.

Kim, D. H., Lee, W. D. and Kang, S. G. (2000). Bayesian model selection for life time data under type II censoring. *Comm. Statist. Theory and Methods* **29**, 2865–2878.

Kim, S. W. and Ibrahim, J. G. (2000). On Bayesian inference for proportional hazards models using noninformative priors. *Lifetime Data Analysis* **6**, 331–341.

Komaki, F. (2001). A shrinkage predictive distribution for multivariate Normal observables. *Biometrika* **88**, 859–864.

Komaki, F. (2004). Simultaneous prediction of independent Poisson observables *Ann. Statist.* **32**, 1744-1769

Kubokawa, T. and Robert, C. P. (1994). New perspectives in linear calibration. *J. Multivariate Analysis* **51**, 178–200.

Kuboki, H. (1993). Inferential distributions for non-Bayesian predictive fit. *Ann. Inst. Statist. Math.* **45**, 567–578.

Kuboki, H. (1998). Reference priors for prediction. *J. Statist. Planning and Inference* **69**, 295–317.

Kullback, S. (1968). *Information Theory and Statistics* (2nd ed.). New York: Dover. Reprinted in 1978, Gloucester, MA: Peter Smith.

Kullback, S. (1983). Kullback information. *Encyclopedia of Statistical Sciences* **4** (S. Kotz, N. L. Johnson and C. B. Read, eds.) New York: Wiley, 421–425.

Kullback, S. (1983). The Kullback-Leibler distance. *Amer. Statist.* **41**, 340–341.

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

Lafferty, J. and Wasserman, L. (2001). Iterative Markov chain Monte Carlo computation of reference priors and minimax risk. *Uncertainty in Artificial Intelligence* (UAI, Seattle).

Lane, D. and Sudderth, W. D. (1984). Coherent predictive inference. *Sankhy$\overline{a}$ A* **46**, 166–185.

Laplace, P. S. (1812). *Théorie Analytique des Probabilités*. Paris: Courcier. Reprinted as *Oeuvres Complètes de Laplace* **7**, 1878–1912. Paris: Gauthier-Villars.

Laplace, P. S. (1825). *Essai Philosophique sur les Probabilitiés*. Paris: Courcier (5th ed). English translation in 1952 as *Philosophical Essay on Probabilities*. New York: Dover.

Lauretto, M., Pereira, C. A. B. Stern, J. M. and Zacks, S. (2003). Comparing parameters of two bivariate normal bistributions using the invariant FBST, Full Bayesian Significance Test. *Brazilian J. Probab. Statist.* **17**, 147–168.

Lee, H. K. H. (2003). A noninformative prior for neural networks. *Machine Learning* **50**, 197–212.

Lee, G. (1998). Development of matching priors for $P(X < Y)$ in exponential distributions. *J. Korean Statist. Soc.* **27**, 421–433.

Lee, J. C. and Chang, C. H. (2000). Bayesian analysis of a growth curve model with a general autoregressive covariance structure. *Scandinavian J. Statist.* **27**, 703–713.

Lee, J. C. and Hwang, R. C. (2000). On estimation and prediction for temporally correlated longitudinal data. *J. Statist. Planning and Inference* **87**, 87–104.

Lee, J. J. and Shin, W. S. (1990). Prior distributions using the entropy principles. *Korean J. Appl. Statist.* **3**, 91–104.

Leucari, V. and Consonni, G. (2003). Compatible priors for causal Bayesian networks. *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press, 597-606.

van der Linde, A. (2000). Reference priors for shrinkage and smoothing parameters. *J. Statist. Planning and Inference* **90**, 245–274.

Lindley, D. V. (1956). On a measure of information provided by an experiment. *Ann. Math. Statist.* **27**, 986–1005.

Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.

Lindley, D. V. (1958). Fiducial distribution and Bayes' Theorem. *J. Roy. Statist. Soc. B* **20**, 102–107.

Lindley, D. V. (1961). The use of prior probability distributions in statistical inference and decision. *Proc. Fourth Berkeley Symp.* **1** (J. Neyman and E. L. Scott, eds.) Berkeley: Univ. California Press, 453–468.

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

Lindley, D. V. (1984). A Bayesian lady tasting tea. *Statistics: An Appraisal*. (H. A.David and H. T.David, eds.) Ames, IA: Iowa State Press455–479.

Lindley, D. V. (1988). Statistical inference concerning the Hardy-Weinberg equilibrium. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) Oxford: University Press, 307–326 (with discussion).

Liseo, B. (1993). Elimination of nuisance parameters with reference priors. *Biometrika* **80**, 295–304.

Liseo, B. (2003). Bayesian and conditional frequentist analyses of the Fieller's problem. A critical review. *Metron* **61**, 133–152.

Liseo, B. and Loperfido, N. (2005). A note on reference priors for the scalar skew-normal distribution. *J. Statist. Planning and Inference* (to appear).

Liseo, B. (2005). The problem of the elimination of nuisance parameters in a Bayesian framework. *In this volume*.

Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statist. Computing* **10**, 325–337.

Madruga, M. R., Pereira, C. A. B. and Stern, J. M. (2003). Bayesian evidence test for precise hypothesis. *J. Statist. Planning and Inference* **117**, 185–198.

Maryak, J. L. and Spall, J. C. (1987). Conditions for the insensitivity of the Bayesian posterior distribution to the choice of prior distribution. *Statistics and Probability Letters* **5**, 401–407.

Marinucci, D. and Petrella, L. (1999). A Bayesian proposal for the analysis of stationary and nonstationary AR(1) time series. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 821–828.

McCulloch, R. E., Polson, N. G. and Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *J. Econometrics* **99**, 173–193.

Meeden, G. and Vardeman, S. (1991). A non-informative Bayesian approach to interval estimation in finite population sampling. *J. Amer. Statist. Assoc.* **86**, 972–986.

Mendoza, M. (1987). A Bayesian analysis of a generalized slope ratio bioassay. *Probability and Bayesian Statistics* (R. Viertl, ed.) London: Plenum, 357–364.

Mendoza, M. (1988). Inferences about the ratio of linear combinations of the coefficients in a multiple regression problem. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) Oxford: University Press, 705–711.

Mendoza, M. (1990). A Bayesian analysis of the slope ratio bioassay. *Biometrika* **46**, 1059–1069.

Mendoza, M. (1994). Asymptotic normality under transformations. A result with Bayesian applications. *Test* **3**, 173–180.

Mendoza, M. and Gutiérrez-Peña, E. (1999). Bayesian inference for the ratio of the means of two normal populations with unequal variances. *Biom. J.* **41**, 133–147.

Mendoza, M. and Gutiérrez-Peña, E. (2000). Bayesian conjugate analysis of the Galton-Walton process. *Test* **9**, 149–171.

Meng, X.-L. and Zaslavsky, A. M. (2002). Single observation unbiased priors. *Ann. Statist.* **30**, 1345–1375.

Mengersen, K. L. and Robert, C. P. (1996). Testing for mixtures: A Bayesian entropic approach. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 255-276 (with discussion).

Van der Merwe, A. J. and Chikobvu, D. (2004). Bayesian analysis of the process capability Index C_{pk}. *South African Statist. J.* **38**, 97–117.

Millar, R. B. (2002). Reference priors for Bayesian fisheries models. *Can. J. Fish. Aquat. Sci.* **59**, 1492–1502.

Miller, R. B. (1980). Bayesian analysis of the two-parameter gamma distribution. *Technometrics* **22**, 65-69.

Molitor, J. and Sun, D. (2002). Bayesian analysis under ordered functions of parameters. *Environ. Ecological Statist.* **9**, 179–193.

Moreno, E. and Girón, F. J. (1997). Estimating with incomplete count data: a Bayesian approach. *J. Statist. Planning and Inference* **66**, 147–159.

Mukerjee R. and Chen Z. (2003). On Expected Lengths of Predictive Intervals *Scandinavian J. Statist.* **30**, 757–766

Mukerjee, R. and Dey, D. K. (1993). Frequentist validity of posterior quantiles in the presence of a nuisance parameter: Higher order asymptotics. *Biometrika* **80**, 499–505.

Mukerjee, R. and Ghosh, M. (1997). Second-order probability matching priors. *Biometrika* **84**, 970–975.

Mukerjee, R. and Reid, N. (1999). On a property of probability matching priors: Matching the alternative coverage probabilities. *Biometrika* **86**, 333–340.

Mukerjee, R. and Reid, N. (2001). Second-order probability matching priors for a parametric function with application to Bayesian tolerance limits. *Biometrika* **88**, 587–592.

Mukhopadhyay, S. and DasGupta, A. (1997). Uniform approximation of Bayes solutions and posteriors: Frequentist valid Bayes inference. *Statistics and Decisions* **15**, 51–73.

Mukhopadhyay, S. and Ghosh, M. (1995). On the uniform approximation of Laplace's prior by t-priors in location problems. *J. Multivariate Analysis* **54**, 284–294.

Natarajan, R. (2001). On the propriety of a modified Jeffreys's prior for variance components in binary random effects models. *Statistics and Probability Letters* **51**, 409–414.

Natarajan, R. and Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. *J. Amer. Statist. Assoc.* **95**, 227–237.

Natarajan, R. and McCulloch, C. E. (1998). Gibbs sampling with diffuse proper priors: A valid approach to data driven inference?. *J. Comp. Graphical Statist.* **7**, 267–277.

Neyman, J. (1950). *First Course in probability and Statistics*. New York: Holt.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

Ni, S. X, and Sun, D. (2003). Noninformative priors and frequentist risks of Bayesian estimators of vector-autoregressive models. *J. Econom.* **115**, 159–197.

Nicolau, A. (1993). Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. *J. Roy. Statist. Soc. B* **55**, 377–390.

Novick, M. R. (1969). Multiparameter Bayesian indifference procedures. *J. Roy. Statist. Soc. B* **31**, 29–64.

Novick, M. R. and Hall, W. K. (1965). A Bayesian indifference procedure. *J. Amer. Statist. Assoc.* **60**, 1104–1117.

Ogata, Y. (1990). A Monte Carlo method for an objective Bayesian procedure. *Ann. Inst. Statist. Math.* **42**, 403–433.

Oh, M.-S. and Kim, Y. (2000). Jeffreys noninformative prior in Bayesian conjoint analysis. *J. Korean Statist. Soc.* **29**, 137–153.

Palmer, J. L. and Pettit, L. I. (1996). Risks of using improper priors with Gibbs sampling and autocorrelated errors. *J. Comp. Graphical Statist.* **5** 245 - 249.

Paris, J. B. (1994) *The Uncertain Reasoner's Companion: A Mathematical Perspective*, Cambridge: University Press.

Pauler, D. K., Wakefield, J. C. and Kass, R. E. (1999). Bayes factors and approximations for variance component models. *J. Amer. Statist. Assoc.* **94**, 1242–1253.

Peers, H. W. (1965). On confidence points and Bayesian probability points in the case of several parameters. *J. Roy. Statist. Soc. B* **27**, 9–16.

Peers, H. W. (1968). Confidence properties of Bayesian interval estimates. *J. Roy. Statist. Soc. B* **30**, 535–544.

Pericchi, L. R. (1981). A Bayesian approach to transformation to normality. *Biometrika* **68**, 35–43.

Pericchi, L. R. (2005). Model selection and hypothesis testing based on probabilities. *In this volume.*

Pericchi L.R. and Sansó B. (1995) A note on bounded influence in Bayesian analysis. *Biometrika* **82**, 223-225.

Pericchi, L. R. and Walley, P. (1991). Robust Bayesian credible intervals and prior ignorance. *Internat. Statist. Rev.* **59**, 1–23.

Perks, W. (1947). Some observations on inverse probability, including a new indifference rule. *J. Inst. Actuaries* **73**, 285–334 (with discussion).

Philippe, A. and Robert, C. P. (1998). A note on the confidence properties of reference priors for the calibration model. *Test* **7**, 147–160.

Philippe A. and Rousseau, J.. (2002). Non informative priors in the case of Gaussian long-memory processes. *Bernoulli* **8**, 451–473.

Phillips, P. C. B. (1991). To criticize the critics: an objective Bayesian analysis of stochastic trends. *J. Applied Econometrics* **6**, 333-473, (with discussion).

Piccinato, L. (1973). Un metodo per determinare distribuzioni iniziali relativamente non-informative. *Metron* **31**, 124–156.

Piccinato, L. (1977). Predictive distributions and non-informative priors. *Trans. 7th. Prague Conf. Information Theory* (M. Uldrich, ed.). Prague: Czech. Acad. Sciences, 399–407.

du Plessis, J. L., van der Merwe, A. J. and Groenewald, P. C. N. (1995). Reference priors for the multivariate calibration problem. *South African Statist. J.* **29**, 155–168.

Poirier, D. J. (1985). Bayesian hypothesis testing in linear models with continuously induced conjugate priors across hypotheses. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 711–722.

Poirier, D. J. (1994). Jeffreys prior for logit models. *J. Econometrics* **63**, 327–339.

Pole, A. and West, M. (1989). Reference analysis of the dynamic linear model. *J. Time Series Analysis* **10**, 13–147.

Polson, N. G. (1992). In discussion of Ghosh and Mukerjee (1992). *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 203–205.

Polson, N. G. and Wasserman, L. (1990). Prior distributions for the bivariate Binomial. *Biometrika* **76**, 604–608.

Press, S. J. (1972). *Applied Multivariate Analysis: using Bayesian and Frequentist Methods of Inference*. Second edition in 1982, Melbourne, FL: Krieger.

Pretorius, A. L. and van der Merwe A. J. (2002). Reference and probability-matching priors in Bayesian analysis of mixed linear models. *J. Anim. Breed. Genet.* **119**, 311–324.

Price, R. M., Bonett, D. G. (2000). Estimating the ratio of two Poisson rates. *Comput. Statist. Data Anal.* **34**, 345–356.

Rabena, M. T. (1998). Deriving reference decisions. *Test* **7**, 161–177.

Rai, G. (1976). Estimates of mean life and reliability function when failure rate is randomly distributed. *Trab. Estadist.* **27**: 247–251.

Raftery, A. E. (1988) Inference for the Binomial N parameter: A hierarchical Bayes approach *Biometrika* **75**. 223–228.

Raftery, A. E. (1988) Analysis of a simple debugging model. *Appl. Statist.* **37**, 12–22.

Raftery, A. E. and Akman, V. E. (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika* **73**, 85–89.

Reid, N. (1996). Likelihood and Bayesian approximation methods. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 349–366 (with discussion).

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11**, 416–431.

Rissanen, J. (1986). Stochastic complexity and modelling *Ann. Statist.* **14**, 1080–1100.

Rissanen, J. (1987). Stochastic complexity. *J. Roy. Statist. Soc. B* **49**, 223-239 and 252-265 (with discussion).

Rissanen, J. (1988). Minimum description length principle. *Encyclopedia of Statistical Sciences* **5** (S. Kotz, N. L. Johnson and C. B. Read, eds.) New York: Wiley, 523–527.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Enquiry.* Singapore: World Scientific Pub..

Roberts, G. O. and Rosenthal, J. S. (2001). Infinite hierarchies and prior distributions. *Bernoulli* **7**, 453–471.

Robert, C. P. (1993). A note on Jeffreys-Lindley paradox. *Statistica Sinica* **3**, 601–608.

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

Rodríguez, C. C. (1991). From Euclid to entropy. *Maximum Entropy and Bayesian Methods* (W. T. Grandy and L. H. Schick eds.). Dordrecht: Kluwer, 343–348.

Rousseau, J. (2000). Coverage properties of one-sided intervals in the discrete case and application to matching priors. *Ann. Inst. Statist. Math.* **52**, 28–42.

Rosa, G. J. M. and Gianola, D. (2001). Inferences about the coefficient of correlation in the standard bivariate normal distribution. *South African Statist. J.* **35**, 69–93.

Roverato, A. and Consonni, G. (2004). Compatible Prior Distributions for DAG models *J. Roy. Statist. Soc. B* **66**, 47–61

Rueda, R. (1992). A Bayesian alternative to parametric hypothesis testing. *Test* **1**, 61–67.

Sansó, B. and Pericchi, L. R. (1992). Near ignorance classes of log-concave priors for the location model. *Test* **1**, 39–46.

Sansó, B. and Pericchi, L. R. (1994). On near ignorance classes. *Rev. Brasilera Prob. Estatist.* **8**, 119–126.

Sareen, S. (2003). Reference Bayesian inference in nonregular models. *J. Econom.* **113**, 265–288.

Savage, L. J. (1954). *The Foundations of Statistics.* New York: Wiley. Second edition in 1972, New York: Dover.

Scholl, H. R. (1998). Shannon optimal priors on independent identically distributed statistical experiments converge weakly to Jeffreys' prior. *Test* **7**, 75–94.

Scricciolo, C. (1999). Probability matching priors: A review. *J. It. Statist. Soc.* **8**, 83–100.

Sendra, M. (1982). Distribución final de referencia para el problema de Fieller-Creasy. *Trab. Estadist.* **33**, 55–72.

Severini, T. A. (1991). On the relationship between Bayesian and non-Bayesian interval estimates. *J. Roy. Statist. Soc. B* **53**, 611–618.

Severini, T. A. (1993). Bayesian interval estimates which are also confidence intervals. *J. Roy. Statist. Soc. B* **53**, 611-618.

Severini, T. A. (1995). Information and conditional inference. *J. Amer. Statist. Assoc.* **90**, 1341-1346.

Severini, T. A. (1999). On the relationship between Bayesian and non-Bayesian elimination of nuisance parameters. *Statistica Sinica* **9**, 713–724.

Severini, T. A., Mukerjee, R. and Ghosh, M. (2002). On an exact probability matching property of right-invariant priors. *Biometrika* **89**, 953–957.

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

Shieh, G. and Lee, J. C. (2002). Bayesian prediction analysis for growth curve model using noninformative priors. *Ann. Inst. Statist. Math.* **54**, 324–337.

Singh, U. and Upadhyay, S. K. (1992). Bayes estimators in the presence of a guess value. *Comm. Statist. Simul. and Comput.* **21**, 1181–1198.

Singh, U., Gupta, P. K. and Upadhyay, S. K. (2002). Estimation of exponentiated Weibull shape parameters under linex loss functions. *Comm. Statist. Simul. and Comput.* **31**, 523–537.

Singpurwalla, N. D. and Wilson, P. (2004) When can finite testing ensure infinite trustworthiness? *J. Iranian Statist. Soc.* **3**, 1–37 (with discussion).

Smith, R. L. (1999). Bayesian and frequentist approaches to parametric predictive inference. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 589–612, (with discussion).

Smith, R. L. and Naylor, J. C. (1987). A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. *Appl. Statist.* **36**, 358–369.

Sono, S. (1983). On a noninformative prior distribution for Bayesian inference of multinomial distribution's parameters. *Ann. Inst. Statist. Math.* **35**, 167–174.

Spall, J. C. and Hill, S. D. (1990). Least informative Bayesian prior distributions for finite samples based on information theory. *IEEE Trans. Automatic Control* **35**, 580–583.

Spiegelhalter, D. J. (1985). Exact Bayesian inference on the parameters of a Cauchy distribution with vague prior information. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 743–749.

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

Stein, C. (1962). Confidence sets for the mean of a multivariate normal distribution. *J. Roy. Statist. Soc. B* **24**, 265–296 (with discussion).

Stein, C. (1986). On the coverage probability of confidence sets based on a prior distribution. *Sequential Methods in Statistics, 3rd ed* (G. B. Wetherbil and K. D. Glazebrook, eds.) London: Chapman and Hall, 485—514.

Stephens, D. A. and Smith, A. F. M. (1992). Sampling-resampling techniques for the computation of posterior densities in normal means problems. *Test* **1**, 1–18.

Stern, J. M. (2004a). Paraconsistent sensitivity analysis for Bayesian significance tests. *Lecture Notes in Artificial Intelligence* **3171**, 134–143.

Stern, J. M. (2004b). Uninformative reference sensitivity in possibilistic sharp hypotheses tests. *Amer. Inst. Physics Proc.* 735, 581–588.

Stewart, W. E. (1987). Multiresponse parameter estimation with a new and noninformative prior. *Biometrika* **74**, 557–562.

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

Stone, M. (1963). The posterior t distribution. *Ann. Math. Statist.* **34**, 568–573.

Stone, M. (1965). Right Haar measures for convergence in probability to invariant posterior distributions. *Ann. Math. Statist.* **36**, 440–453.

Stone, M. (1970). Necessary and sufficient conditions for convergence in probability to invariant posterior distributions. *Ann. Math. Statist.* **41**, 1939–1953.

Stone, M. (1976). Strong inconsistency from uniform priors. *J. Amer. Statist. Assoc.* **71**, 114–125 (with discussion).

Stone, M. and Dawid, A. P. (1972). Un-Bayesian implications of improper Bayesian inference in routine statistical problems. *Biometrika* **59**, 369–375.

Stone, M. and Springer, B. G. F. (1965). A paradox involving quasi prior distributions. *Biometrika* **52**, 623–627.

Strachan, R. W. and van Dijk, H. K. (2003). Bayesian model selection with an uninformative prior. *Oxford Bul. Econ. Statist.* **65**, 863–76.

Strawderman, W. E. (2000). Minimaxity. *J. Amer. Statist. Assoc.* **95**, 1364-1368.

Sugiura, N. and Ishibayashi, H. (1997). Reference prior Bayes estimator for bivariate normal covariance matrix with risk comparison. *Comm. Statist. Theory and Methods* **26**, 2203–2221.

Sun, D. (1994). Integrable expansions for posterior distributions for a two-parameter exponential family.*Ann. Statist.* **22**, 1808–1830.

Sun, D. (1997). A note on noninformative priors for Weibull distributions. *J. Statist. Planning and Inference* **61**, 319–338.

Sun, D. and Berger, J. O. (1994). Bayesian sequential reliability for Weibull and related distributions. *Ann. Inst. Statist. Math.* **46**, 221–249.

Sun, D. and Berger, J. O. (1998). Reference priors with partial information. *Biometrika* **85**, 55–71.

Sun, D., Ghosh, M. and Basu, A. P. (1998). Bayesian analysis for a stress-strength system under noninformative priors. *Can. J. Statist.* **26**, 323–332.

Sun, D. and Ni, S. (2004). Bayesian analysis of vector-autoregressive models with noninformative priors. *J. Statist. Planning and Inference* **121**, 291–309.

Sun, D., Tsutakawa, R. K. and He, Z. (2001). Propriety of posteriors with improper priors in hierarchical linear mixed models. *Statistica Sinica* **11**, 77–95.

Sun, D. and Ye, K. (1995). Reference prior Bayesian analysis for normal mean products. *J. Amer. Statist. Assoc.* **90**, 589–597.

Sun, D. and Ye, K. (1996). Frequentist validity of posterior quantiles for a two-parameter exponential family. *Biometrika* **83**, 55-65.

Sun, D. and Ye, K. (1999). Reference priors for a product of normal means when variances are unknown. *Can. J. Statist.* **27**, 97–103.

Sweeting, T. J. (1985). Consistent prior distributions for transformed models. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 755–762.

Sweeting, T. J. (1984). On the choice of prior distribution for the Box-Cox transformed linear model. *Biometrika* **71**, 127–134.

Sweeting, T. J. (1995a). A framework for Bayesian and likelihood approximations in statistics. *Biometrika* **82**, 1–24.

Sweeting, T. J. (1995b). A Bayesian approach to approximate conditional inference. *Biometrika* **82**, 25–36.

Sweeting, T. J. (1996). Approximate Bayesian computation based on signed roots of log-density rations. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 427–444, (with discussion).

Sweeting, T. J. (2001). Coverage probability bias, objective Bayes and the likelihood principle. *Biometrika* **88**, 657–675.

Tardella, L. (2002). A new Bayesian method for nonparametric capture-recapture models in presence of heterogeneity. *Biometrika* **89**, 807–817.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76**, 604–608.

Tiwari, R. C., Chib, S. and Jammalamadaka, S. R. (1989). Bayes estimation of the multiple correlation coefficient. *Comm. Statist. Theory and Methods* **18**, 1401–1413.

Torgersen, E. N. (1981). Measures of information based on comparison with total information and with total ignorance. *Ann. Statist.* **9**, 638–657.

Tsutakawa, R. K. (1992). Prior distribution for item response curves. *British J. Math. Statist. Philosophy* **45**, 51–74.

Upadhyay, S. K. and Pandey, M. (1989). Prediction limits for an exponential distribution: a Bayes predictive distribution approach. *IEEE Trans. Reliability* **38**, 599–602.

Upadhyay, S. K. and Peshwani, M. (2003). Choice between Weibull and lognormal models: A simulation-based bayesian study. *Comm. Statist. Theory and Methods* **32**, 381–405.

Upadhyay, S. K. and Smith, A. F. M. (1994). Modelling complexities in reliability, and the role of simulation in Bayesian computation. *Internat. J. Continuing Eng. Education* **4**, 93–104.

Upadhyay, S. K., Agrawal, R. and Smith, A. F. M. (1996). Bayesian analysis of inverse Gaussian non-linear regression by simulation. *Sankhyā B* **58**, 363–378.

Upadhyay, S. K., Vasishta, N. and Smith, A. F. M. (2001). Bayes inference in life testing and reliability via Markov chain Montecarlo simulation. *Sankhyā A* **63**, 15–40.

Vaurio, J. K. (1992). Objective prior distributions and Bayesian updating. *Reliability Eng. System Safety* **35**, 55–59.

Vernotte, F. and Zalamansky, G. (1999). A Bayesian method for oscillator stability analysis. *IEEE Trans. Ultrasonics, Ferroelec. and Freq. Controls* **46**, 1545–1550.

Villegas, C. (1969). On the a priori distribution of the covariance matrix. *Ann. Math. Statist.* **40**, 1098–1099.

Villegas, C. (1971). On Haar priors. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) Toronto: Holt, Rinehart and Winston, 409–414 (with discussion).

Villegas, C. (1977a). Inner statistical inference. *J. Amer. Statist. Assoc.* **72**, 453–458.

Villegas, C. (1977b). On the representation of ignorance. *J. Amer. Statist. Assoc.* **72**, 651–654.

Villegas, C. (1981). Inner statistical inference II. *Ann. Statist.* **9**, 768–776.

Villegas, C. (1982). Maximum likelihood and least squares estimation in linear and affine functional models. *Ann. Statist.* **10**, 256–265.

Villegas, C. (1990). Bayesian inference in models with Euclidean structures. *J. Amer. Statist. Assoc.* **85**, 1159–1164.

de Waal, D. J. and Nel, D. G. (1988). A procedure to select a ML-II prior in a multivariate normal case. *Comm. Statist. Simul. and Comput.* **17**, 1021–1035.

de Waal, D. J., and Groenewald, P. C. N. (1989). On measuring the amount of information from the data in a Bayesian analysis. *South African Statist. J.* **23**, 23–62, (with discussion).

de Waal, D. J., Groenewald, P. C. N. and Kemp, C. J. (1995). Perturbation of the Normal Model in Linear Regression. *South African Statist. J.* **29**, 105–130.

Wallace, C. S. and Dowe, D. L. (1999). Minimum message length and Kolmogorov complexity. *The Computer J.* 42, 270–283.

Wallace, C. S. and Freeman, P. R. (1987). Estimation and inference by compact coding. *J. Roy. Statist. Soc. B* **49**, 240–265.

Walker, S. and Muliere, P, (1999). A characterization of a neutral to the right prior via an extension of Johnson's sufficientness postulate. *Ann. Statist.* **27**, 589–599.

Walker, S. G. and Gutiérrez-Peña, E. (1999). Robustifying Bayesian procedures. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 685–710, (with discussion).

Wasserman, L. (1996). The conflict between improper priors and robustness. *J. Statist. Planning and Inference* **52**, 1–15.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *J. Roy. Statist. Soc. B* **62**, 159–180.

Wasserman, L. and Clarke, B. (1995). Information tradeoff. *Test* **4**, 19–38.

Welch, B. L. (1965). On comparisons between confidence point procedures in the case of a single parameter. *J. Roy. Statist. Soc. B* **27**, 1–8.

Welch, B. L. and Peers, H. W. (1963). On formulae for confidence points based on intervals of weighted likelihoods. *J. Roy. Statist. Soc. B* **25**, 318–329.

Wolfinger, R. D., Kass, R. E. (2000). Nonconjugate Bayesian analysis of variance component models. *Biometrics* **56**, 768–774.

Yang, R. (1995). Invariance of the reference prior under reparametrization. *Test* **4**, 83–94.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22**, 1195–1211.

Yang, R. and Berger, J. O. (1997). A catalogue of noninformative priors. *Tech. Rep.*, Duke University, ISDS 97-42.

Yang, R and Chen, M.-H. (1995). Bayesian analysis of random coefficient regression models using noninformative priors. *J. Multivariate Analysis* **55**, 283–311.

Ye, K. (1993). Reference priors when the stopping rule depends on the parameter of interest. *J. Amer. Statist. Assoc.* **88**, 360–363.

Ye, K. (1994). Bayesian reference prior analysis on the ratio of variances for the balanced one-way random effect model. *J. Statist. Planning and Inference* **41**, 267–280.

Ye, K. (1995). Selection of the reference priors for a balanced random effects model. *Test* **4**, 1–17.

Ye, K. (1998). Estimating a ratio of the variances in a balanced one-way random effects model. *Statistics and Decisions* **16**, 163–180.

Ye, K. and Berger, J. O. (1991). Non-informative priors for inferences in exponential regression models. *Biometrika* **78**, 645–656.

Yuan, A. and Clarke, B. S. (1999). A minimally informative likelihood for decision analysis: Illustration and robustness, *Can. J. Statist.* **27**, 649–665.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley. Reprinted in 1987, Melbourne, FL: Krieger.

Zellner, A. (1977). Maximal data information prior distributions. *New Developments in the Applications of Bayesian Methods* (A. Ayka and C. Brumat, eds.) Amsterdam: North-Holland, 211–232.

Zellner, A. (1983). Applications of Bayesian analysis in econometrics. *The Statistician* **32**, 23–34.

Zellner, A. (1986a). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.) Amsterdam: North-Holland, 233–243.

Zellner, A. (1986b). Bayesian estimation and prediction using asymmetric loss functions. *J. Amer. Statist. Assoc.* **81**, 446–451.

Zellner, A. (1988). Optimal information processing and Bayes' theorem. *Amer. Statist.* **42**, 278–284 (with discussion).

Zellner, A. (1991). Bayesian methods and entropy in economics and econometrics. *Maximum Entropy and Bayesian Methods* (W. T. Grandy and L. H. Schick eds.). Dordrecht: Kluwer, 17–31.

Zellner, A. (1996). Models, prior information, and Bayesian analysis. *J. Econometrics* **75**, 51–68.

Zellner, A. (1997). *Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers*. Brookfield, VT: Edward Elgar.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypothesis. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) Valencia: University Press, 585–603 and 618–647 (with discussion).

Zidek, J. V.(1969). A representation of Bayes invariant procedures in terms of Haar measure. *Ann. Inst. Statist. Math.* **21**, 291–308.