Chapter 5

Inference

Summary

The role of Bayes' theorem in the updating of beliefs about observables in the light of new information is identified and related to conventional mechanisms of predictive and parametric inference. The roles of sufficiency, ancillarity and stopping rules in such inference processes are also examined. Forms of common statistical decisions and inference summaries are introduced and the problems of implementing Bayesian procedures are discussed at length. In particular, conjugate, asymptotic and reference forms of analysis and numerical approximation approaches are detailed.

5.1 THE BAYESIAN PARADIGM

5.1.1 Observables, Beliefs and Models

Our development has focused on the foundational issues which arise when we aspire to formal quantitative coherence in the context of decision making in situations of uncertainty. This development, in combination with an operational approach to the basic concepts, has led us to view the problem of statistical modelling as that of identifying or selecting particular forms of representation of beliefs about observables.

For example, in the case of a sequence $x_1, x_2, \ldots,$ of 0 – 1 random quantities for which beliefs correspond to a judgement of infinite exchangeability, Proposition 4.1, (de Finetti's theorem) identifies the representation of the joint mass function for $x_1, \ldots, x_n$ as having the form

$$p(x_1, \ldots, x_n) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}\, dQ(\theta),$$

for some choice of distribution Q over the interval $[0, 1]$.

More generally, for sequences of real-valued or integer-valued random quantities, $x_1, x_2, \ldots,$ we have seen, in Sections 4.3 – 4.5, that beliefs which combine judgements of exchangeability with some form of further structure (either in terms of invariance or sufficient statistics), often lead us to work with representations of the form

$$p(x_1, \ldots, x_n) = \int_{\Re^k} \prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta})\, dQ(\boldsymbol{\theta}),$$

where $p(x \,|\, \boldsymbol{\theta})$ denotes a specified form of labelled family of probability distributions and Q is some choice of distribution over $\Re^k$.

Such representations, and the more complicated forms considered in Section 4.6, exhibit the various ways in which the element of primary significance from the subjectivist, operationalist standpoint, namely the *predictive model* of beliefs about observables, can be thought of *as if* constructed from a *parametric model* together with a *prior distribution* for the labelling parameter.

Our primary concern in this chapter will be with the way in which the updating of beliefs in the light of new information takes place within the framework of such representations.

5.1.2 The Role of Bayes' Theorem

In its simplest form, within the formal framework of predictive model belief distributions derived from quantitative coherence considerations, the problem corresponds to identifying the joint conditional density of

$$p(x_{n+1}, \ldots, x_{n+m} \,|\, x_1, \ldots, x_n)$$

for any $m \geq 1$, given, for any $n \geq 1$, the form of representation of the joint density $p(x_1, \ldots, x_n)$.

In general, of course, this simply reduces to calculating

$$p(x_{n+1}, \ldots, x_{n+m} \,|\, x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_{n+m})}{p(x_1, \ldots, x_n)}$$

and, in the absence of further structure, there is little more that can be said. However, when the predictive model admits a representation in terms of parametric models and prior distributions, the learning process can be essentially identified, in conventional terminology, with the standard parametric form of Bayes' theorem.

Thus, for example, if we consider the general parametric form of representation for an exchangeable sequence, with $dQ(\boldsymbol{\theta})$ having density representation, $p(\boldsymbol{\theta})d\boldsymbol{\theta}$, we have

$$p(x_1, \ldots, x_n) = \int \prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

from which it follows that

$$\begin{aligned} p(x_{n+1}, \ldots, x_{n+m} \,|\, x_1, \ldots, x_n) &= \frac{\int \prod_{i=1}^{n+m} p(x_i \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int \prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \\ &= \int \prod_{i=n+1}^{n+m} p(x_i \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, x_1, \ldots, x_n)\, d\boldsymbol{\theta}, \end{aligned}$$

where

$$p(\boldsymbol{\theta} \,|\, x_1, \ldots, x_n) = \frac{\prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int \prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}\,.$$

This latter relationship is just *Bayes' theorem*, expressing the *posterior density* for $\boldsymbol{\theta}$, given $x_1, \ldots, x_n$, in terms of the *parametric model* for $x_1, \ldots, x_n$ given $\boldsymbol{\theta}$, and the *prior density* for $\boldsymbol{\theta}$. The (conditional, or posterior) predictive model for $x_{n+1}, \ldots, x_{n+m}$, given $x_1, \ldots, x_n$ is seen to have precisely the same general form of representation as the initial predictive model, except that the corresponding parametric model component is now integrated with respect to the posterior distribution of the parameter, rather than with respect to the prior distribution.

We recall from Chapter 4 that, considered as a function of $\boldsymbol{\theta}$,

$$\text{lik}(\boldsymbol{\theta} \,|\, x_1, \ldots, x_n) = p(x_1, \ldots, x_n \,|\, \boldsymbol{\theta})$$

is usually referred to as the *likelihood function*. A formal definition of such a concept is, however, problematic; for details, see Bayarri *et al.* (1988) and Bayarri and DeGroot (1992b).

5.1.3 Predictive and Parametric Inference

Given our operationalist concern with modelling and reporting uncertainty in terms of *observables*, it is not surprising that Bayes' theorem, in its role as the key to a coherent learning process for *parameters*, simply appears as a step within the predictive process of passing from

$$p(x_1, \ldots, x_n) = \int p(x_1, \ldots, x_n \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

to

$$p(x_{n+1},\ldots,x_{n+m}\,|\,x_1,\ldots,x_n) = \int p(x_{n+1},\ldots,x_{n+m}\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta}\,|\,x_1,\ldots,x_n)\,d\boldsymbol{\theta},$$

by means of

$$p(\boldsymbol{\theta}\,|\,x_1,\ldots,x_n) = \frac{p(x_1,\ldots,x_n\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(x_1,\ldots,x_n\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta}}\,.$$

Writing $\boldsymbol{y} = \{y_1,\ldots,y_m\} = \{x_{n+1},\ldots,x_{n+m}\}$ to denote future (or, as yet unobserved) quantities and $\boldsymbol{x} = \{x_1,\ldots,x_n\}$ to denote the already observed quantities, these relations may be re-expressed more simply as

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta},$$
$$p(\boldsymbol{y}\,|\,\boldsymbol{x}) = \int p(\boldsymbol{y}\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta}\,|\,\boldsymbol{x})\,d\boldsymbol{\theta}$$

and

$$p(\boldsymbol{\theta}\,|\,\boldsymbol{x}) = p(\boldsymbol{x}\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\boldsymbol{x}).$$

However, as we noted on many occasions in Chapter 4, if we proceed purely formally, from an operationalist standpoint it is not at all clear, at first sight, how we should interpret "beliefs about parameters", as represented by $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}\,|\,\boldsymbol{x})$, or even whether such "beliefs" have any intrinsic interest. We also answered these questions on many occasions in Chapter 4, by noting that, in all the forms of predictive model representations we considered, the parameters had interpretations as strong law limits of (appropriate functions of) observables. Thus, for example, in the case of the infinitely exchangeable 0 – 1 sequence (Section 4.3.1) beliefs about θ correspond to beliefs about what the long-run frequency of 1's would be in a future sample; in the context of a real-valued exchangeable sequence with centred spherical symmetry (Section 4.4.1), beliefs about μ and σ^2, respectively, correspond to beliefs about what the large sample mean, and the large sample mean sum of squares about the sample mean would be, in a future sample.

Inference about parameters is thus seen to be a limiting form of predictive inference about observables. This means that, although the predictive form is primary, and the role of parametric inference is typically that of an intermediate structural step, parametric inference will often itself be the legitimate end-product of a statistical analysis in situations where interest focuses on quantities which could be viewed as large-sample functions of observables. Either way, parametric inference is of considerable importance for statistical analysis in the context of the models we are mainly concerned with in this volume.

When a parametric form is involved simply as an intermediate step in the predictive process, we have seen that $p(\boldsymbol{\theta} \mid x_1, \ldots, x_n)$, the full joint posterior density for the parameter vector $\boldsymbol{\theta}$ is all that is required. However, if we are concerned with parametric inference *per se*, we may be interested in only some subset, $\boldsymbol{\phi}$, of the components of $\boldsymbol{\theta}$, or in some transformed subvector of parameters, $\boldsymbol{g}(\boldsymbol{\theta})$. For example, in the case of a real-valued sequence we may only be interested in the large-sample mean and not in the variance; or in the case of two $0-1$ sequences we may only be interested in the difference in the long-run frequencies.

In the case of interest in a subvector of $\boldsymbol{\theta}$, let us suppose that the full parameter vector can be partitioned into $\boldsymbol{\theta} = \{\boldsymbol{\phi}, \boldsymbol{\lambda}\}$, where $\boldsymbol{\phi}$ is the subvector of interest, and $\boldsymbol{\lambda}$ is the complementary subvector of $\boldsymbol{\theta}$, often referred to, in this context, as the vector of *nuisance parameters*. Since

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{x})} \ ,$$

the (marginal) posterior density for $\boldsymbol{\phi}$ is given by

$$p(\boldsymbol{\phi} \mid \boldsymbol{x}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{x}) \, d\boldsymbol{\lambda} = \int p(\boldsymbol{\phi}, \boldsymbol{\lambda} \mid \boldsymbol{x}) \, d\boldsymbol{\lambda},$$

where

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, d\boldsymbol{\theta} = \int p(\boldsymbol{x} \mid \boldsymbol{\phi}, \boldsymbol{\lambda}) p(\boldsymbol{\phi}, \boldsymbol{\lambda}) d\boldsymbol{\phi} \, d\boldsymbol{\lambda},$$

with all integrals taken over the full range of possible values of the relevant quantities.

Expressed in terms of the notation introduced in Section 3.2.4, we have

$$p(\boldsymbol{x} \mid \boldsymbol{\phi}, \boldsymbol{\lambda}) \otimes p(\boldsymbol{\phi}, \boldsymbol{\lambda}) \equiv p(\boldsymbol{\phi}, \boldsymbol{\lambda} \mid \boldsymbol{x}),$$

$$p(\boldsymbol{\phi}, \boldsymbol{\lambda} \mid \boldsymbol{x}) \underset{\phi}{\longrightarrow} p(\boldsymbol{\phi} \mid \boldsymbol{x}).$$

In some situations, the prior specification $p(\boldsymbol{\phi}, \boldsymbol{\lambda})$ may be most easily arrived at through the specification of $p(\boldsymbol{\lambda} \mid \boldsymbol{\phi}) p(\boldsymbol{\phi})$. In such cases, we note that we could first calculate the *integrated likelihood* for $\boldsymbol{\phi}$,

$$p(\boldsymbol{x} \mid \boldsymbol{\phi}) = \int p(\boldsymbol{x} \mid \boldsymbol{\phi}, \boldsymbol{\lambda}) p(\boldsymbol{\lambda} \mid \boldsymbol{\phi}) \, d\boldsymbol{\lambda},$$

and subsequently proceed without any further need to consider the nuisance parameters, since

$$p(\boldsymbol{\phi} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \boldsymbol{\phi}) p(\boldsymbol{\phi})}{p(\boldsymbol{x})} \ .$$

In the case where interest is focused on a transformed parameter vector, $\boldsymbol{g}(\boldsymbol{\theta})$, we proceed using standard change-of-variable probability techniques as described in Section 3.2.4. Suppose first that $\boldsymbol{\psi} = \boldsymbol{g}(\boldsymbol{\theta})$ is a one-to-one differentiable transformation of $\boldsymbol{\theta}$. It then follows that

$$p_{\psi}(\boldsymbol{\psi} \,|\, \boldsymbol{x}) = p_{\theta}(\boldsymbol{g}^{-1}(\boldsymbol{\psi}) \,|\, \boldsymbol{x}) \,|\, \boldsymbol{J}_{g^{-1}}(\boldsymbol{\psi}) \,|\,,$$

where

$$\boldsymbol{J}_{g^{-1}}(\boldsymbol{\psi}) = \frac{\partial \boldsymbol{g}^{-1}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}$$

is the Jacobian of the inverse transformation $\boldsymbol{\theta} = \boldsymbol{g}^{-1}(\boldsymbol{\psi})$. Alternatively, by substituting $\boldsymbol{\theta} = \boldsymbol{g}^{-1}(\boldsymbol{\psi})$, we could write $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$ as $p(\boldsymbol{x} \,|\, \boldsymbol{\psi})$, and replace $p(\boldsymbol{\theta})$ by $p_{\theta}(\boldsymbol{g}^{-1}(\boldsymbol{\psi})) \,|\, \boldsymbol{J}_{g^{-1}}(\boldsymbol{\psi}) \,|$, to obtain $p(\boldsymbol{\psi} \,|\, \boldsymbol{x}) = p(\boldsymbol{x} \,|\, \boldsymbol{\psi}) p(\boldsymbol{\psi}) / p(\boldsymbol{x})$ directly.

If $\boldsymbol{\psi} = g(\boldsymbol{\theta})$ has dimension less than $\boldsymbol{\theta}$, we can typically define $\boldsymbol{\gamma} = (\boldsymbol{\psi}, \boldsymbol{\omega}) = \boldsymbol{h}(\boldsymbol{\theta})$, for some $\boldsymbol{\omega}$ such that $\boldsymbol{\gamma} = \boldsymbol{h}(\boldsymbol{\theta})$ is a one-to-one differentiable transformation, and then proceed in two steps. We first obtain

$$p(\boldsymbol{\psi}, \boldsymbol{\omega} \,|\, \boldsymbol{x}) = p_{\theta}(\boldsymbol{h}^{-1}(\boldsymbol{\gamma}) \,|\, \boldsymbol{x}) \,|\, \boldsymbol{J}_{h^{-1}}(\boldsymbol{\gamma}) \,|\,,$$

where

$$\boldsymbol{J}_{h^{-1}}(\boldsymbol{\gamma}) = \frac{\partial \boldsymbol{h}^{-1}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\,,$$

and then marginalise to

$$p(\boldsymbol{\psi} \,|\, \boldsymbol{x}) = \int p(\boldsymbol{\psi}, \boldsymbol{\omega} \,|\, \boldsymbol{x})\, d\boldsymbol{\omega}.$$

These techniques will be used extensively in later sections of this chapter.

In order to keep the presentation of these basic manipulative techniques as simple as possible, we have avoided introducing additional notation for the ranges of possible values of the various parameters. In particular, all integrals have been assumed to be over the full ranges of the possible parameter values.

In general, this notational economy will cause no confusion and the parameter ranges will be clear from the context. However, there are situations where specific constraints on parameters are introduced and need to be made explicit in the analysis. In such cases, notation for ranges of parameter values will typically also need to be made explicit.

Consider, for example, a parametric model, $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, together with a prior specification $p(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, for which the posterior density, suppressing explicit use of Θ, is given by

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}\,.$$

Now suppose that it is required to specify the posterior subject to the constraint $\boldsymbol{\theta} \in \Theta_0 \subset \Theta$, where $\int_{\Theta_0} p(\boldsymbol{\theta}) d\boldsymbol{\theta} > 0$.

Defining the constrained prior density by

$$p_0(\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta})}{\int_{\Theta_0} p(\boldsymbol{\theta}) d(\boldsymbol{\theta})}, \quad \boldsymbol{\theta} \in \Theta_0,$$

we obtain, using Bayes' theorem,

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{\theta} \in \Theta_0) = \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}) p_0(\boldsymbol{\theta})}{\int_{\Theta_0} p(\boldsymbol{x} \mid \boldsymbol{\theta}) p_0(\boldsymbol{\theta}) d\boldsymbol{\theta}}, \quad \boldsymbol{\theta} \in \Theta_0.$$

From this, substituting for $p_0(\boldsymbol{\theta})$ in terms of $p(\boldsymbol{\theta})$ and dividing both numerator and denominator by

$$p(\boldsymbol{x}) = \int_{\Theta} p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

we obtain

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{\theta} \in \Theta_0) = \frac{p(\boldsymbol{\theta} \mid \boldsymbol{x})}{\int_{\Theta_0} p(\boldsymbol{\theta} \mid \boldsymbol{x}) \, d\boldsymbol{\theta}}, \quad \boldsymbol{\theta} \in \Theta_0,$$

expressing the constraint in terms of the unconstrained posterior (a result which could, of course, have been obtained by direct, straightforward conditioning).

Numerical methods are often necessary to analyze models with constrained parameters; see Gelfand *et al.* (1992) for the use of Gibbs sampling in this context.

5.1.4 Sufficiency, Ancillarity and Stopping Rules

The concepts of predictive and parametric sufficient statistics were introduced in Section 4.5.2, and shown to be equivalent, within the framework of the kinds of models we are considering in this volume. In particular, it was established that a (minimal) sufficient statistic, $\boldsymbol{t}(\boldsymbol{x})$, for $\boldsymbol{\theta}$, in the context of a parametric model $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, can be characterised by either of the conditions

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}) = p\big(\boldsymbol{\theta} \mid \boldsymbol{t}(\boldsymbol{x})\big), \qquad \text{for all } p(\boldsymbol{\theta}),$$

or

$$p\big(\boldsymbol{x} \mid \boldsymbol{t}(\boldsymbol{x}), \boldsymbol{\theta}\big) = p\big(\boldsymbol{x} \mid \boldsymbol{t}(\boldsymbol{x})\big).$$

The important implication of the concept is that $\boldsymbol{t}(\boldsymbol{x})$ serves as a sufficient summary of the complete data $\boldsymbol{x}$ in forming any required revision of beliefs. The resulting data reduction often implies considerable simplification in modelling and analysis. In

many cases, the sufficient statistic $\boldsymbol{t}(\boldsymbol{x})$ can itself be partitioned into two component statistics, $\boldsymbol{t}(\boldsymbol{x}) = [\boldsymbol{a}(\boldsymbol{x}), \boldsymbol{s}(\boldsymbol{x})]$ such that, for all $\boldsymbol{\theta}$,

$$\begin{aligned} p\big(\boldsymbol{t}(\boldsymbol{x}) \,|\, \boldsymbol{\theta}\big) &= p\big(\boldsymbol{s}(\boldsymbol{x}) \,|\, \boldsymbol{a}(\boldsymbol{x}), \boldsymbol{\theta}\big) p\big(\boldsymbol{a}(\boldsymbol{x}) \,|\, \boldsymbol{\theta}\big) \\ &= p\big(\boldsymbol{s}(\boldsymbol{x}) \,|\, \boldsymbol{a}(\boldsymbol{x}), \boldsymbol{\theta}\big) p\big(\boldsymbol{a}(\boldsymbol{x})\big). \end{aligned}$$

It then follows that, for any choice of $p(\boldsymbol{\theta})$,

$$\begin{aligned} p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = p\big(\boldsymbol{\theta} \,|\, \boldsymbol{t}(\boldsymbol{x})\big) &\propto p\big(\boldsymbol{t}(\boldsymbol{x}) \,|\, \boldsymbol{\theta}\big) p(\boldsymbol{\theta}) \\ &\propto p\big(\boldsymbol{s}(\boldsymbol{x}) \,|\, \boldsymbol{a}(\boldsymbol{x}), \boldsymbol{\theta}\big) p(\boldsymbol{\theta}), \end{aligned}$$

so that, in the prior to posterior inference process defined by Bayes' theorem, it suffices to use $p(\boldsymbol{s}(\boldsymbol{x}) \,|\, \boldsymbol{a}(\boldsymbol{x}), \boldsymbol{\theta})$, rather than $p(\boldsymbol{t}(\boldsymbol{x}) \,|\, \boldsymbol{\theta})$ as the likelihood function. This further simplification motivates the following definition.

Definition 5.1. (***Ancillary statistic***). *A statistic, $\boldsymbol{a}(\boldsymbol{x})$, is said to be ancillary, with respect to $\boldsymbol{\theta}$ in a parametric model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, if $p(\boldsymbol{a}(\boldsymbol{x}) \,|\, \boldsymbol{\theta}) = p(\boldsymbol{a}(\boldsymbol{x}))$ for all values of $\boldsymbol{\theta}$.*

Example 5.1. ***(Bernoulli model)***. In Example 4.5, we saw that for the Bernoulli parametric model

$$p(x_1, \ldots, x_n \,|\, \theta) = \prod_{i=1}^{n} p(x_i \,|\, \theta) = \theta^{r_n} (1 - \theta)^{n - r_n},$$

which only depends on n and $r_n = x_1 + \cdots + x_n$. Thus, $\boldsymbol{t}_n = [n, r_n]$ provides a minimal sufficient statistic, and one may work in terms of the joint probability function $p(n, r_n \,|\, \theta)$.

If we now write

$$p(n, r_n \,|\, \theta) = p(r_n \,|\, n, \theta) p(n \,|\, \theta),$$

and make the assumption that, for all $n \geq 1$, the mechanism by which the sample size, n, is arrived at does not depend on θ, so that $p(n \,|\, \theta) = p(n)$, $n \geq 1$, we see that n *is ancillary for* θ, in the sense of Definition 5.1. It follows that prior to posterior inference for θ can therefore proceed on the basis of

$$p(\theta \,|\, \boldsymbol{x}) = p(\theta \,|\, n, r_n) \propto p(r_n \,|\, n, \theta) p(\theta),$$

for any choice of $p(\theta)$, $0 \leq \theta \leq 1$. From Corollary 4.1, we see that

$$\begin{aligned} p(r_n \,|\, n, \theta) &= \binom{n}{r_n} \theta^{r_n} (1 - \theta)^{n - r_n}, \qquad 0 \leq r_n \leq n, \\ &= \mathrm{Bi}(r_n \,|\, \theta, n), \end{aligned}$$

so that inferences in this case can be made as if we had adopted a *binomial parametric model*. However, if we write

$$p(n, r_n \mid \theta) = p(n \mid r_n, \theta)p(r_n \mid \theta)$$

and make the assumption that, for all $r_n \geq 1$, termination of sampling is governed by a mechanism for selecting r_n, which does not depend on θ, so that $p(r_n \mid \theta) = p(r_n), r_n \geq 1$, we see that r_n is ancillary for θ, in the sense of Definition 5.1. It follows that prior to posterior inference for θ can therefore proceed on the basis of

$$p(\theta \mid \boldsymbol{x}) = p(\theta \mid n, r_n) \propto p(n \mid r_n, \theta)p(\theta),$$

for any choice of $p(\theta), 0 < \theta \leq 1$. It is easily verified that

$$\begin{aligned} p(n \mid r_n, \theta) &= \binom{n-1}{r_n - 1}\theta^{r_n}(1-\theta)^{n-r_n}, \qquad n \geq r_n, \\ &= \text{Nb}(n \mid \theta, r_n) \end{aligned}$$

(see Section 3.2.2), so that inferences in this case can be made as if we had adopted a *negative-binomial parametric model*.

We note, incidentally, that whereas in the binomial case it makes sense to consider $p(\theta)$ as specified over $0 \leq \theta \leq 1$, in the negative-binomial case it may only make sense to think of $p(\theta)$ as specified over $0 < \theta \leq 1$, since $p(r_n \mid \theta = 0) = 0$, for all $r_n \geq 1$.

So far as prior to posterior inference for θ is concerned, we note that, for any specified $p(\theta)$, and assuming that either $p(n \mid \theta) = p(n)$ or $p(r_n \mid \theta) = p(r_n)$, we obtain

$$p(\theta \mid x_1, \ldots, x_n) = p(\theta \mid n, r_n) \propto \theta^{r_n}(1-\theta)^{n-r_n}p(\theta)$$

since, considered as functions of θ,

$$p(r_n \mid n, \theta) \propto p(n \mid r_n, \theta) \propto \theta^{r_n}(1-\theta)^{n-r_n}.$$

The last part of the above example illustrates a general fact about the mechanism of parametric Bayesian inference which is trivially obvious; namely, *for any specified $p(\boldsymbol{\theta})$, if the likelihood functions $p_1(\boldsymbol{x}_1 \mid \boldsymbol{\theta}), p_2(\boldsymbol{x}_2 \mid \boldsymbol{\theta})$ are proportional as functions of $\boldsymbol{\theta}$, the resulting posterior densities for $\boldsymbol{\theta}$ are identical*. It turns out, as we shall see in Appendix B, that many non-Bayesian inference procedures do not lead to identical inferences when applied to such proportional likelihoods. The assertion that they *should*, the so-called *Likelihood Principle*, is therefore a controversial issue among statisticians . In contrast, in the Bayesian inference context described above, this is a straightforward consequence of Bayes' theorem, rather than an imposed "principle". Note, however, that the above remarks are predicated on a specified $p(\boldsymbol{\theta})$. It may be, of course, that knowledge of the particular sampling mechanism employed has implications for the specification of $p(\boldsymbol{\theta})$, as illustrated, for example, by the comment above concerning negative-binomial sampling and the restriction to $0 < \theta \leq 1$.

Although the likelihood principle is implicit in Bayesian statistics, it was developed as a separate principle by Barnard (1949), and became a focus of interest when Birnbaum (1962) showed that it followed from the widely accepted sufficiency and conditionality principles. Berger and Wolpert (1984/1988) provide an extensive discussion of the likelihood principle and related issues. Other relevant references are Barnard *et al.* (1962), Fraser (1963), Pratt (1965), Barnard (1967), Hartigan (1967), Birnbaum (1968, 1978), Durbin (1970), Basu (1975), Dawid (1983a), Joshi (1983), Berger (1985b), Hill (1987) and Bayarri *et al.* (1988).

Example 5.1 illustrates the way in which ancillary statistics often arise naturally as a consequence of the way in which data are collected. In general, it is very often the case that the sample size, n, is fixed in advance and that inferences are automatically made conditional on n, without further reflection. It is, however, perhaps not obvious that inferences can be made conditional on n if the latter has arisen as a result of such familiar imperatives as "stop collecting data when you feel tired", or "when the research budget runs out". The kind of analysis given above makes it intuitively clear that such conditioning is, in fact, valid, provided that the mechanism which has led to n "does not depend on $\boldsymbol{\theta}$". This latter condition may, however, not always be immediately obviously transparent, and the following definition provides one version of a more formal framework for considering sampling mechanisms and their dependence on model parameters.

Definition 5.2. (***Stopping rule***). *A **stopping rule**, $\boldsymbol{h}$, for (sequential) sampling from a sequence of observables $x_1 \in X_1, x_2 \in X_2, \ldots$, is a sequence of functions $h_n : X_1 \times \cdots \times X_n \to [0, 1]$, such that, if $\boldsymbol{x}_{(n)} = (x_1, \ldots, x_n)$ is observed, then sampling is terminated with probability $h_n(\boldsymbol{x}_{(n)})$; otherwise, the $(n + 1)$th observation is made. A stopping rule is **proper** if the induced probability distribution $p_h(n), n = 1, 2, \ldots$, for final sample size guarantees that the latter is finite. The rule is **deterministic** if $h_n(\boldsymbol{x}_{(n)}) \in \{0, 1\}$ for all $(n, \boldsymbol{x}_{(n)})$; otherwise, it is a **randomised** stopping rule.*

In general, we must regard the data resulting from a sampling mechanism defined by a stopping rule $\boldsymbol{h}$ as consisting of $(n, \boldsymbol{x}_{(n)})$, the sample size, together with the observed quantities $x_1, \ldots, x_n$. A parametric model for these data thus involves a probability density of the form $p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{h}, \boldsymbol{\theta})$, conditioning both on the stopping rule (i.e., sampling mechanism) and on an underlying labelling parameter $\boldsymbol{\theta}$. But, either through unawareness or misapprehension, this is typically ignored and, instead, we act as if the actual observed sample size n had been fixed in advance, in effect assuming that

$$p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{h}, \boldsymbol{\theta}) = p(\boldsymbol{x}_{(n)} \,|\, n, \boldsymbol{\theta}) = p(\boldsymbol{x}_{(n)} \,|\, \boldsymbol{\theta}),$$

using the standard notation we have hitherto adopted for fixed n. The important question that now arises is the following: under what circumstances, if any, can

we proceed to make inferences about $\boldsymbol{\theta}$ on the basis of this (generally erroneous!) assumption, without considering explicit conditioning on the actual form of $\boldsymbol{h}$? Let us first consider a simple example.

Example 5.2. ***("Biased" stopping rule for a Bernoulli sequence)***. Suppose, given θ, that $x_1, x_2, \ldots$ may be regarded as a sequence of independent Bernoulli random quantities with $p(x_i \,|\, \theta) = \text{Bi}(x_i \,|\, \theta, 1)$, $x_i = 0, 1$, and that a sequential sample is to be obtained using the deterministic stopping rule $\boldsymbol{h}$, defined by: $h_1(1) = 1$, $h_1(0) = 0$, $h_2(x_1, x_2) = 1$ for all x_1, x_2. In other words, if there is a success on the first trial, sampling is terminated (resulting in $n = 1$, $x_1 = 1$); otherwise, two observations are obtained (resulting in either $n = 2$, $x_1 = 0$, $x_2 = 0$ or $n = 2$, $x_1 = 0$, $x_2 = 1$).

At first sight, it might appear essential to take explicit account of $\boldsymbol{h}$ in making inferences about θ, since the sampling procedure seems designed to bias us towards believing in large values of θ. Consider, however, the following detailed analysis:

$$\begin{aligned} p(n = 1, x_1 = 1 \,|\, \boldsymbol{h}, \theta) &= p(x_1 = 1 \,|\, n = 1, \boldsymbol{h}, \theta) p(n = 1 \,|\, \boldsymbol{h}, \theta) \\ &= 1 \cdot p(x_1 = 1 \,|\, \theta) = p(x_1 = 1 \,|\, \theta) \end{aligned}$$

and, for $x = 0, 1$,

$$\begin{aligned} p(n = 2, x_1 = 0, x_2 = x \,|\, \boldsymbol{h}, \theta) &= p(x_1 = 0, x_2 = x \,|\, n = 2, \boldsymbol{h}, \theta) p(n = 2 \,|\, \boldsymbol{h}, \theta) \\ &= p(x_1 = 0 | n = 2, \boldsymbol{h}, \theta) p(x_2 = x \,|\, x_1 = 0, n = 2, \boldsymbol{h}, \theta) p(n = 2 \,|\, \boldsymbol{h}, \theta) \\ &= 1 \cdot p(x_2 = x \,|\, x_1 = 0, \theta) p(x_1 = 0 \,|\, \theta) \\ &= p(x_2 = x, x_1 = 0 \,|\, \theta). \end{aligned}$$

Thus, for all $(n, \boldsymbol{x}_{(n)})$ having non-zero probability, we obtain in this case

$$p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{h}, \theta) = p(\boldsymbol{x}_{(n)} \,|\, \theta),$$

the latter considered pointwise as functions of θ (i.e., likelihoods). It then follows trivially from Bayes' theorem that, *for any specified* $p(\theta)$, inferences for θ based on assuming n to have been fixed at its observed value will be identical to those based on a likelihood derived from explicit consideration of $\boldsymbol{h}$.

Consider now a randomised version of this stopping rule which is defined by $h_1(1) = \pi$, $h_1(0) = 0$, $h_2(x_1, x_2) = 1$ for all x_1, x_2. In this case, we have

$$\begin{aligned} p(n = 1, x_1 = 1 \,|\, \boldsymbol{h}, \theta) &= p(x_1 = 1 \,|\, n = 1, \boldsymbol{h}, \theta) p(n = 1 \,|\, \boldsymbol{h}, \theta) \\ &= 1 \cdot \pi \cdot p(x_1 = 1 \,|\, \theta), \end{aligned}$$

with, for $x = 0, 1$,

$$\begin{aligned} p(n =& 2, x_1 = 0, x_2 = x \,|\, \boldsymbol{h}, \theta) \\ &= p(n = 2 \,|\, x_1 = 0, \boldsymbol{h}, \theta) \\ &\quad \times p(x_1 = 0 \,|\, \boldsymbol{h}, \theta) p(x_2 = x \,|\, x_1 = 0, n = 2, \boldsymbol{h}, \theta) \\ &= 1 \cdot p(x_1 = 0 \,|\, \theta) p(x_2 = x \,|\, \theta) \end{aligned}$$

and

$$\begin{aligned} p(n=2, x_1=1, x_2=x \,|\, \boldsymbol{h}, \theta) &= p(n=2 \,|\, x_1=1, \boldsymbol{h}, \theta) p(x_1=1 \,|\, \boldsymbol{h}, \theta) \\ &\quad \times p(x_2=x \,|\, x_1=1, n=2, \boldsymbol{h}, \theta) \\ &= (1-\pi) p(x_1=1 \,|\, \theta) p(x_2=x \,|\, \theta). \end{aligned}$$

Thus, for all $(n, \boldsymbol{x}_{(n)})$ having non-zero probability, we again find that

$$p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{h}, \theta) \propto p(\boldsymbol{x}_{(n)} \,|\, \theta)$$

as functions of θ, so that the proportionality of the likelihoods once more implies identical inferences from Bayes' theorem, for any given $p(\theta)$.

The analysis of the preceding example showed, perhaps contrary to intuition, that, although seemingly biasing the analysis towards beliefs in larger values of θ, the stopping rule does not in fact lead to a different likelihood from that of the a priori fixed sample size. The following, rather trivial, proposition makes clear that this is true for all stopping rules as defined in Definition 5.2, which we might therefore describe as "likelihood non-informative stopping rules".

Proposition 5.1. **(*Stopping rules are likelihood non-informative*)**.
For any stopping rule $\boldsymbol{h}$, for (sequential) sampling from a sequence of observables $x_1, x_2, \ldots$, having fixed sample size parametric model $p(\boldsymbol{x}_{(n)} \,|\, n, \boldsymbol{\theta}) = p(\boldsymbol{x}_{(n)} \,|\, \boldsymbol{\theta})$,

$$p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{h}, \boldsymbol{\theta}) \propto p(\boldsymbol{x}_{(n)} \,|\, \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta,$$

for all $(n, \boldsymbol{x}_{(n)})$ such that $p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{h}, \boldsymbol{\theta}) \neq 0$.

Proof. This follows straightforwardly on noting that

$$p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{h}, \boldsymbol{\theta}) = \Big[\boldsymbol{h}_n(\boldsymbol{x}_{(n)}) \prod_{i=1}^{n-1} \big(1 - \boldsymbol{h}_i(\boldsymbol{x}_{(i)})\big) \Big] p(\boldsymbol{x}_{(n)} \,|\, \boldsymbol{\theta}),$$

and that the term in square brackets does not depend on $\boldsymbol{\theta}$. ◁

Again, it is a trivial consequence of Bayes' theorem that, *for any specified prior density*, prior to posterior inference for $\boldsymbol{\theta}$ given data $(n, \boldsymbol{x}_{(n)})$ obtained using a likelihood non-informative stopping rule $\boldsymbol{h}$ can proceed by acting as if $\boldsymbol{x}_{(n)}$ were obtained using a fixed sample size n. However, a notationally precise rendering of Bayes' theorem,

$$\begin{aligned} p(\boldsymbol{\theta} \,|\, n, \boldsymbol{x}_{(n)}, \boldsymbol{h}) &\propto p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{h}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, \boldsymbol{h}) \\ &\propto p(\boldsymbol{x}_{(n)} \,|\, \theta) p(\boldsymbol{\theta} \,|\, \boldsymbol{h}), \end{aligned}$$

reveals that *knowledge of $\boldsymbol{h}$ might well affect the specification of the prior density*! It is for this reason that we use the term "likelihood non-informative" rather than just "non-informative" stopping rules. It cannot be emphasised too often that, although it is often convenient for expository reasons to focus at a given juncture on one or other of the "likelihood" and "prior" components of the model, our discussion in Chapter 4 makes clear their basic inseparability in coherent modelling and analysis of beliefs. This issue is highlighted in the following example.

Example 5.3. ***("Biased" stopping rule for a normal mean)***. Suppose, given θ, that $x_1, x_2, \ldots$, may be regarded as a sequence of independent normal random quantities with $p(x_i \,|\, \theta) = \mathrm{N}(x_i \,|\, \theta, 1)$, $x_i \in \Re$. Suppose further that an investigator has a particular concern with the parameter value $\theta = 0$ and wants to stop sampling if $\overline{x}_n = \sum_i x_i/n$ ever takes on a value that is "unlikely", assuming $\theta = 0$ to be true.

For any fixed sample size n, if "unlikely" is interpreted as "an event having probability less than or equal to α", for small α, a possible stopping rule, using the fact that $p(\overline{x}_n \,|\, n, \theta) = \mathrm{N}(\overline{x}_n \,|\, \theta, n)$, might be

$$h_n(\boldsymbol{x}_{(n)}) = \begin{cases} 1, & \text{if } |\,\overline{x}_n\,| > k(\alpha)/\sqrt{n} \\ 0, & \text{if } |\,\overline{x}_n\,| \leq k(\alpha)/\sqrt{n} \end{cases}$$

for suitable $k(\alpha)$ (for example, $k = 1.96$ for $\alpha = 0.05$, $k = 2.57$ for $\alpha = 0.01$, or $k = 3.31$ for $\alpha = 0.001$). It can be shown, using the law of the iterated logarithm (see, for example, Section 3.2.3), that this is a proper stopping rule, so that termination will certainly occur for some finite n, yielding data $(n, \boldsymbol{x}_{(n)})$. Moreover, defining

$$S_n = \left\{ \boldsymbol{x}_{(n)};\ |\bar{x}_1| \leq k(\alpha),\ |\bar{x}_2| \leq \frac{k(\alpha)}{\sqrt{2}}, \ldots, \right. \\ \left. |\bar{x}_{n-1}| \leq \frac{k(\alpha)}{\sqrt{n-1}},\ |\bar{x}_n| > \frac{k(\alpha)}{\sqrt{n}} \right\},$$

we have

$$\begin{aligned} p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{h}, \theta) &= p(\boldsymbol{x}_{(n)} \,|\, n, \boldsymbol{h}, \theta) p(n \,|\, \boldsymbol{h}, \theta) \\ &= p(\boldsymbol{x}_{(n)} \,|\, S_n, \theta) p(S_n \,|\, \theta) \\ &= p(\boldsymbol{x}_{(n)} \,|\, \theta), \end{aligned}$$

as a function of θ, for all $(n, \boldsymbol{x}_{(n)})$ for which the left-hand side is non-zero. It follows that $\boldsymbol{h}$ is a likelihood non-informative stopping rule.

Now consider prior to posterior inference for θ, where, for illustration, we assume the prior specification $p(\theta) = \mathrm{N}(\theta \,|\, \mu, \lambda)$, with precision $\lambda \simeq 0$, to be interpreted as indicating extremely vague prior beliefs about θ, which take no explicit account of the stopping rule $\boldsymbol{h}$. Since the latter is likelihood non-informative, we have

$$\begin{aligned} p(\theta \,|\, \boldsymbol{x}_{(n)}, n) &\propto p(\boldsymbol{x}_{(n)} \,|\, n, \theta) p(\theta) \\ &\propto p(\overline{x}_n \,|\, n, \theta) p(\theta) \\ &\propto \mathrm{N}(\overline{x}_n \,|\, \theta, n) \mathrm{N}(\theta \,|\, \mu, \lambda) \end{aligned}$$

by virtue of the sufficiency of $(n, \overline{x}_n)$ for the normal parametric model. The right-hand side is easily seen to be proportional to $\exp\{-\frac{1}{2}Q(\theta)\}$, where

$$Q(\theta) = (n+h)\left[\theta - \frac{n\bar{x}_n + \lambda\mu}{n+\lambda}\right]^2,$$

which implies that

$$\begin{aligned} p(\theta \,|\, \boldsymbol{x}_{(n)}, n) &= \mathrm{N}\left(\theta \middle| \frac{n\bar{x}_n + \lambda\mu}{n+\lambda}, (n+\lambda)\right) \\ &\simeq \mathrm{N}(\theta \,|\, \overline{x}_n, n) \end{aligned}$$

for $\lambda \simeq 0$.

One consequence of this vague prior specification is that, having observed $(n, \boldsymbol{x}_{(n)})$, we are led to the posterior probability statement

$$P\left[\theta \in \left(\overline{x}_n \pm \frac{k(\alpha)}{\sqrt{n}}\right) \middle| n, \overline{x}_n\right] = 1 - \alpha.$$

But the stopping rule $\boldsymbol{h}$ ensures that $|\,\overline{x}_n\,| > k(\alpha)/\sqrt{n}$. This means that the value $\theta = 0$ certainly does not lie in the posterior interval to which someone with initially very vague beliefs would attach a high probability. An investigator *knowing* $\theta = 0$ to be the true value can therefore, by using this stopping rule, mislead someone who, unaware of the stopping rule, acts as if initially very vague.

However, let us now consider an analysis which takes into account the stopping rule. The nature of $\boldsymbol{h}$ might suggest a prior specification $p(\theta \,|\, \boldsymbol{h})$ that recognises $\theta = 0$ as a possibly "special" parameter value, which should be assigned non-zero prior probability (rather than the zero probability resulting from any continuous prior density specification). As an illustration, suppose that we specify

$$p(\theta \,|\, \boldsymbol{h}) = \pi\ 1_{(\theta=0)}(\theta) + (1-\pi)1_{(\theta\neq 0)}(\theta)\mathrm{N}(\theta \,|\, 0, \lambda_0),$$

which assigns a "spike" of probability, π, to the special value, $\theta = 0$, and assigns $1 - \pi$ times a $\mathrm{N}(\theta \,|\, 0, \lambda_0)$ density to the range $\theta \neq 0$.

Since $\boldsymbol{h}$ is a likelihood non-informative stopping rule and $(n, \overline{x}_n)$ are sufficient statistics for the normal parametric model, we have

$$p(\theta \,|\, n, \boldsymbol{x}_{(n)}, \boldsymbol{h}) \propto \mathrm{N}(\overline{x}_n \,|\, \theta, n)p(\theta \,|\, \boldsymbol{h}).$$

The complete posterior $p(\theta \,|\, n, \boldsymbol{x}_{(n)}, \boldsymbol{h})$ is thus given by

$$\begin{aligned} &\frac{\pi\ 1_{(\theta=0)}(\theta)\mathrm{N}(\overline{x}_n \,|\, 0, n) + (1-\pi)1_{(\theta\neq 0)}(\theta)\mathrm{N}(\overline{x}_n \,|\, \theta, n)\mathrm{N}(\theta \,|\, 0, \lambda_0)}{\pi\ \mathrm{N}(\overline{x}_n \,|\, 0, n) + (1-\pi)\int_{-\infty}^{\infty} \mathrm{N}(\overline{x}_n \,|\, \theta, n)\mathrm{N}(\theta \,|\, 0, \lambda_0)d\theta} \\ &\qquad = \pi^* 1_{(\theta=0)}(\theta) + (1-\pi^*)1_{(\theta\neq 0)}\mathrm{N}\left(\theta \middle| \frac{n\bar{x}_n}{n+\lambda_0}, n+\lambda_0\right), \end{aligned}$$

where, since

$$\int_{-\infty}^{\infty} \mathrm{N}(\overline{x}_n \,|\, \theta, n)\mathrm{N}(\theta \,|\, 0, \lambda_0)d\theta = \mathrm{N}\left(\overline{x}_n \,|\, 0, n\frac{\lambda_0}{n+\lambda_0}\right),$$

it is easily verified that

$$\begin{aligned}\pi^* &= \left\{1 + \frac{1-\pi}{\pi} \cdot \frac{\mathrm{N}(\overline{x}_n \,|\, 0, n\lambda_0(n+\lambda_0)^{-1})}{\mathrm{N}(\overline{x}_n \,|\, 0, n)}\right\}^{-1} \\ &= \left\{1 + \frac{1-\pi}{\pi}\left(1+\frac{n}{\lambda_0}\right)^{-1/2} \exp\left[\tfrac{1}{2}(\sqrt{n}\overline{x}_n)^2\left(1+\frac{\lambda_0}{n}\right)^{-1}\right]\right\}^{-1}.\end{aligned}$$

The posterior distribution thus assigns a "spike" π^* to $\theta = 0$ and assigns $1 - \pi^*$ times a $\mathrm{N}(\theta \,|\, (n+\lambda_0)^{-1}n\overline{x}_n,\, n+\lambda_0)$ density to the range $\theta \neq 0$.

The behaviour of this posterior density, derived from a prior taking account of $\boldsymbol{h}$, is clearly very different from that of the posterior density based on a vague prior taking no account of the stopping rule. For qualitative insight, consider the case where actually $\theta = 0$ and α has been chosen to be very small, so that $k(\alpha)$ is quite large. In such a case, n is likely to be very large and at the stopping point we shall have $\overline{x}_n \simeq k(\alpha)/\sqrt{n}$. This means that

$$\pi^* \simeq \left[1 + \frac{1-\pi}{\pi}\left(1+\frac{n}{\lambda_0}\right)^{-1/2} \exp\left(\tfrac{1}{2}k^2(\alpha)\right)\right]^{-1} \simeq 1,$$

for large n, so that knowing the stopping rule and then observing that it results in a large sample size leads to an increasing conviction that $\theta = 0$. On the other hand, if θ is appreciably different from 0, the resulting n, and hence π^*, will tend to be small and the posterior will be dominated by the $\mathrm{N}(\theta \,|\, (n+\lambda_0)^{-1}n\overline{x}_n, n+\lambda_0)$ component.

5.1.5 Decisions and Inference Summaries

In Chapter 2, we made clear that our central concern is the representation and revision of beliefs as the basis for decisions. Either beliefs are to be used directly in the choice of an action, or are to be recorded or reported in some selected form, with the possibility or intention of subsequently guiding the choice of a future action.

With slightly revised notation and terminology, we recall from Chapters 2 and 3 the elements and procedures required for coherent, quantitative decision-making. The elements of a decision problem in the inference context are:

(i) $\boldsymbol{a} \in \mathcal{A}$, available "answers" to the inference problem;

(ii) $\boldsymbol{\omega} \in \Omega$, unknown states of the world;

(iii) $u : \mathcal{A} \times \Omega \to \Re$, a function attaching utilities to each consequence $(\boldsymbol{a}, \boldsymbol{\omega})$ of a decision to summarise inference in the form of an "answer", $\boldsymbol{a}$, and an ensuing state of the world, $\boldsymbol{\omega}$;

(iv) $p(\boldsymbol{\omega})$, a specification, in the form of a probability distribution, of current beliefs about the possible states of the world.

The optimal choice of answer to an inference problem is an $\boldsymbol{a} \in \mathcal{A}$ which *maximises the expected utility*,

$$\int_{\Omega} u(\boldsymbol{a}, \boldsymbol{\omega}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega}.$$

Alternatively, if instead of working with $u(\boldsymbol{a}, \boldsymbol{\omega})$ we work with a so-called *loss function*,

$$l(\boldsymbol{a}, \boldsymbol{\omega}) = f(\boldsymbol{\omega}) - u(\boldsymbol{a}, \boldsymbol{\omega}),$$

where f is an arbitrary, fixed function, the optimal choice of answer is an $\boldsymbol{a} \in \mathcal{A}$ which *minimises the expected loss*,

$$\int_{\Omega} l(\boldsymbol{a}, \boldsymbol{\omega}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega}.$$

It is clear from the forms of the expected utilities or losses which have to be calculated in order to choose an optimal answer, that, if beliefs about unknown states of the world are to provide an appropriate basis for future decision making, where, as yet, $\mathcal{A}$ and u (or l) may be unspecified, we need to report the complete belief distribution $p(\boldsymbol{\omega})$.

However, if an immediate application to a particular decision problem, with specified $\mathcal{A}$ and u (or l), is all that is required, the optimal answer—maximising the expected utility or minimising the expected loss—may turn out to involve only limited, specific features of the belief distribution, so that these "summaries" of the full distribution suffice for decision-making purposes.

In the following headed subsections, we shall illustrate and discuss some of these commonly used forms of summary. Throughout, we shall have in mind the context of parametric and predictive inference, where the unknown states of the world are parameters or future data values (observables), and current beliefs, $p(\boldsymbol{\omega})$, typically reduce to one or other of the familiar forms:

$p(\boldsymbol{\theta})$	initial beliefs about a parameter vector, $\boldsymbol{\theta}$;
$p(\boldsymbol{\theta} \mid \boldsymbol{x})$	beliefs about $\boldsymbol{\theta}$, given data $\boldsymbol{x}$;
$p(\boldsymbol{\psi} \mid \boldsymbol{x})$	beliefs about $\boldsymbol{\psi} = \boldsymbol{g}(\boldsymbol{\theta})$, given data $\boldsymbol{x}$;
$p(\boldsymbol{y} \mid \boldsymbol{x})$	beliefs about future data $\boldsymbol{y}$, given data $\boldsymbol{x}$.

Point Estimates

In cases where $\boldsymbol{\omega} \in \Omega$ corresponds to an unknown quantity, so that Ω is $\Re$, or $\Re^k$, or $\Re^+$, or $\Re \times \Re^+$, etc., and the required answer, $\boldsymbol{a} \in \mathcal{A}$, is an estimate of the true value of $\boldsymbol{\omega}$ (so that $\mathcal{A} = \Omega$), the corresponding decision problem is typically referred to as one of *point estimation*.

If $\boldsymbol{\omega} = \boldsymbol{\theta}$ or $\boldsymbol{\omega} = \boldsymbol{\psi}$, we refer to *parametric* point estimation; if $\boldsymbol{\omega} = \boldsymbol{y}$, we refer to *predictive* point estimation. Moreover, since one is almost certain not to get the answer exactly right in an estimation problem, statisticians typically work directly with the loss function concept, rather than with the utility function. A point estimation problem is thus completely defined once $\mathcal{A} = \Omega$ and $l(\boldsymbol{a}, \boldsymbol{\omega})$ are specified. Direct intuition suggests that in the one-dimensional case, distributional summaries such as the mean, median or mode of $p(\boldsymbol{\omega})$ could be reasonable point estimates of a random quantity $\boldsymbol{\omega}$. Clearly, however, these could differ considerably, and more formal guidance may be required as to when and why particular functionals of the belief distribution are justified as point estimates. This is provided by the following definition and result.

Definition 5.3. (***Bayes estimate***). *A Bayes estimate of $\boldsymbol{\omega}$ with respect to the loss function $l(\boldsymbol{a}, \boldsymbol{\omega})$ and the belief distribution $p(\boldsymbol{\omega})$ is an $\boldsymbol{a} \in \mathcal{A} = \Omega$ which minimises $\int_\Omega l(\boldsymbol{a}, \boldsymbol{\omega}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega}$.*

Proposition 5.2. (***Forms of Bayes estimates***).

(i) *If $\mathcal{A} = \Omega = \Re^k$, $l(\boldsymbol{a}, \boldsymbol{\omega}) = (\boldsymbol{a} - \boldsymbol{\omega})^t \boldsymbol{H} (\boldsymbol{a} - \boldsymbol{\omega})$, and $\boldsymbol{H}$ is symmetric definite positive, the Bayes estimate satisfies*

$$\boldsymbol{H}\boldsymbol{a} = \boldsymbol{H} E(\boldsymbol{\omega}).$$

*If $\boldsymbol{H}^{-1}$ exists, $\boldsymbol{a} = E(\boldsymbol{\omega})$, and so **the Bayes estimate with respect to quadratic form loss is the mean** of $p(\boldsymbol{\omega})$, assuming the mean to exist.*

(ii) *If $\mathcal{A} = \Omega = \Re$ and $l(\boldsymbol{a}, \boldsymbol{\omega}) = c_1(a - \omega) 1_{(\omega \le a)}(a) + c_2(\omega - a) 1_{(\omega > a)}(a)$, the **Bayes estimate with respect to linear loss is the quantile** such that*

$$P(\omega \le a) = c_2/(c_1 + c_2).$$

*If $c_1 = c_2$, the right-hand side equals $1/2$ and so **the Bayes estimate with respect to absolute value loss is a median** of $p(\boldsymbol{\omega})$.*

(iii) *If $\mathcal{A} = \Omega \subseteq \Re^k$ and $l(\boldsymbol{a}, \boldsymbol{\omega}) = 1 - 1_{(B_\varepsilon(\boldsymbol{a}))}(\boldsymbol{\omega})$, where $B_\varepsilon(\boldsymbol{a})$ is a ball of radius ε in Ω centred at $\boldsymbol{a}$, the Bayes estimate maximises*

$$\int_{B_\varepsilon(\boldsymbol{a})} p(\boldsymbol{\omega})\, d\boldsymbol{\omega}.$$

*As $\varepsilon \to 0$, the function to be maximised tends to $p(\boldsymbol{a})$ and so **the Bayes estimate with respect to zero-one loss is a mode** of $p(\boldsymbol{\omega})$, assuming a mode to exist.*

Proof. Differentiating $\int (\boldsymbol{a}-\boldsymbol{\omega})^t \boldsymbol{H}(\boldsymbol{a}-\boldsymbol{\omega})p(\boldsymbol{\omega})\,d\boldsymbol{\omega}$ with respect to $\boldsymbol{a}$ and equating to zero yields

$$2\boldsymbol{H}\int (\boldsymbol{a}-\boldsymbol{\omega})p(\boldsymbol{\omega})\,d\boldsymbol{\omega} = 0.$$

This establishes (i). Since

$$\int l(\boldsymbol{a},\boldsymbol{\omega})p(\boldsymbol{\omega})\,d\boldsymbol{\omega} = c_1\int_{\{\boldsymbol{\omega}\leq\boldsymbol{a}\}}(\boldsymbol{a}-\boldsymbol{\omega})p(\boldsymbol{\omega})\,d\boldsymbol{\omega} + c_2\int_{\{\boldsymbol{\omega}>\boldsymbol{a}\}}(\boldsymbol{\omega}-\boldsymbol{a})p(\boldsymbol{\omega})\,d\boldsymbol{\omega},$$

differentiating with respect to a and equating to zero yields

$$c_1\int_{\{\boldsymbol{\omega}\leq\boldsymbol{a}\}} p(\boldsymbol{\omega})\,d\boldsymbol{\omega} = c_2\int_{\{\boldsymbol{\omega}>\boldsymbol{a}\}} p(\boldsymbol{\omega})\,d\boldsymbol{\omega},$$

whence, adding $c_2\int_{\omega\leq a} p(\boldsymbol{\omega})\,d\boldsymbol{\omega}$ to each side, we obtain (ii). Finally, since

$$\int l(\boldsymbol{a},\boldsymbol{\omega})p(\boldsymbol{\omega})\,d\boldsymbol{\omega} = 1 - \int 1_{B_\varepsilon(a)}(\boldsymbol{\omega})p(\boldsymbol{\omega})\,d\boldsymbol{\omega},$$

and this is minimised when $\int_{B_\varepsilon(a)} p(\boldsymbol{\omega})\,d\boldsymbol{\omega}$ is maximised, we have (iii). ◁

Further insight into the nature of case (iii) can be obtained by thinking of a unimodal, continuous $p(\omega)$ in one dimension. It is then immediate by a continuity argument that a should be chosen such that

$$p(a-\varepsilon) = p(a+\varepsilon).$$

In the case of a unimodal, symmetric belief distribution, $p(\omega)$, for a single random quantity ω, the mean, median and mode coincide. In general, for unimodal, positively skewed, densities we have the relation

$$\text{mean} > \text{median} > \text{mode}$$

and the difference can be substantial if $p(\omega)$ is markedly skew. Unless, therefore, there is a very clear need for a point estimate, and a strong rationale for a specific one of the loss functions considered in Proposition 5.2, the provision of a single number to summarise $p(\omega)$ may be extremely misleading as a summary of the information available about ω. Of course, such a comment acquires even greater force if $p(\omega)$ is multimodal or otherwise "irregular".

For further discussion of Bayes estimators, see, for example, DeGroot and Rao (1963, 1966), Sacks (1963), Farrell (1964), Brown (1973), Tiao and Box (1974), Berger and Srinivasan (1978), Berger (1979, 1986), Hwang (1985, 1988), de la Horra (1987, 1988, 1992), Ghosh (1992a, 1992b), Irony (1992) and Spall and Maryak (1992).

Credible regions

We have emphasised that, from a theoretical perspective, uncertainty about an unknown quantity of interest, $\boldsymbol{\omega}$, needs to be communicated in the form of the full (prior, posterior or predictive) density, $p(\boldsymbol{\omega})$, if formal calculation of expected loss or utility is to be possible for any arbitrary future decision problem. In practice, however, $p(\boldsymbol{\omega})$ may be a somewhat complicated entity and it may be both more convenient, and also sufficient for general orientation regarding the uncertainty about $\boldsymbol{\omega}$, simply to describe regions $C \subseteq \Omega$ of given probability under $p(\boldsymbol{\omega})$. Thus, for example, in the case where $\Omega \subseteq \Re$, the identification of intervals containing 50%, 90%, 95% or 99% of the probability under $p(\omega)$ might suffice to give a good idea of the general quantitative messages implicit in $p(\omega)$. This is the intuitive basis of popular graphical representations of univariate distributions such as *box plots*.

Definition 5.4. (***Credible Region***). *A region $C \subseteq \Omega$ such that*

$$\int_C p(\boldsymbol{\omega})\, d\boldsymbol{\omega} = 1 - \alpha$$

is said to be a $100(1-\alpha)$% credible region for $\boldsymbol{\omega}$, with respect to $p(\boldsymbol{\omega})$.

*If $\Omega \subseteq \Re$, connected credible regions will be referred to as **credible intervals**.*

If $p(\boldsymbol{\omega})$ is a (prior-posterior-predictive) density, we refer to (prior-posterior-predictive) credible regions.

Clearly, for any given α there is not a unique credible region—even if we restrict attention to connected regions, as we should normally wish to do for obvious ease of interpretation (at least in cases where $p(\boldsymbol{\omega})$ is unimodal). For given Ω, $p(\boldsymbol{\omega})$ and fixed α, the problem of choosing among the subsets $C \subseteq \Omega$ such that $\int_C p(\boldsymbol{\omega})\, d\boldsymbol{\omega} = 1 - \alpha$ could be viewed as a decision problem, provided that we are willing to specify a loss function, $l(C, \boldsymbol{\omega})$, reflecting the possible consequences of quoting the $100(1-\alpha)$% credible region C. We now describe the resulting form of credible region when a loss function is used which encapsulates the intuitive idea that, for given α, we would prefer to report a credible region C whose size $||C||$ (volume, area, length) is minimised.

Proposition 5.3. (***Minimal size credible regions***). *Let $p(\boldsymbol{\omega})$ be a probability density for $\boldsymbol{\omega} \in \Omega$ almost everywhere continuous; given α, $0 < \alpha < 1$, if $\mathcal{A} = \{C;\ P(\boldsymbol{\omega} \in C) = 1 - \alpha\} \neq \emptyset$ and*

$$l(C, \boldsymbol{\omega}) = k||C|| - 1_C(\boldsymbol{\omega}), \quad C \in \mathcal{A}, \quad \boldsymbol{\omega} \in \Omega, \quad k > 0,$$

then C is optimal if and only if it has the property that $p(\boldsymbol{\omega}_1) \geq p(\boldsymbol{\omega}_2)$ for all $\boldsymbol{\omega}_1 \in C$, $\boldsymbol{\omega}_2 \notin C$ (except possibly for a subset of Ω of zero probability).

Proof. It follows straightforwardly that, for any $C \in \mathcal{A}$,

$$\int_{\Omega} l(C, \boldsymbol{\omega}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega} = k||C|| + 1 - \alpha,$$

so that an optimal C must have minimal size.

If C has the stated property and D is any other region belonging to $\mathcal{A}$, then since $C = (C \cap D) \cup (C \cap D^c)$, $D = (C \cap D) \cup (C^c \cap D)$ and $P(\boldsymbol{\omega} \in C) = P(\boldsymbol{\omega} \in D)$, we have

$$\begin{aligned} \inf_{\boldsymbol{\omega} \in C \cap D^c} p(\boldsymbol{\omega}) ||C \cap D^c|| &\leq \int_{C \cap D^c} p(\boldsymbol{\omega})\, d\boldsymbol{\omega} \\ &= \int_{C^c \cap D} p(\boldsymbol{\omega})\, d\boldsymbol{\omega} \leq \sup_{\boldsymbol{\omega} \in C^c \cap D} p(\boldsymbol{\omega}) ||C^c \cap D|| \end{aligned}$$

with

$$\sup_{\boldsymbol{\omega} \in C^c \cap D} p(\boldsymbol{\omega}) \leq \inf_{\boldsymbol{\omega} \in C \cap D^c} p(\boldsymbol{\omega})$$

so that $||C \cap D^c|| \leq ||C^c \cap D||$, and hence $||C|| \leq ||D||$.

If C does not have the stated property, there exists $A \subseteq C$ such that for all $\boldsymbol{\omega}_1 \in A$, there exists $\boldsymbol{\omega}_2 \notin C$ such that $p(\boldsymbol{\omega}_2) > p(\boldsymbol{\omega}_1)$. Let $B \subseteq C^c$ be such that $P(\boldsymbol{\omega} \in A) = P(\boldsymbol{\omega} \in B)$ and $p(\boldsymbol{\omega}_2) > p(\boldsymbol{\omega}_1)$ for all $\boldsymbol{\omega}_2 \in B$ and $\boldsymbol{\omega}_1 \in A$. Define $D = (C \cap A^c) \cup B$. Then $D \in \mathcal{A}$ and by a similar argument to that given above the result follows by showing that $||D|| < ||C||$. ◁

The property of Proposition 5.3 is worth emphasising in the form of a definition (Box and Tiao, 1965).

Definition 5.5. (***Highest probability density (HPD) regions***).
A region $C \subseteq \Omega$ is said to be a $100(1-\alpha)$% highest probability density region for $\boldsymbol{\omega}$ with respect to $p(\boldsymbol{\omega})$ if

(i) $P(\boldsymbol{\omega} \in C) = 1 - \alpha$

(ii) $p(\boldsymbol{\omega}_1) \geq p(\boldsymbol{\omega}_2)$ *for all $\boldsymbol{\omega}_1 \in C$ and $\boldsymbol{\omega}_2 \notin C$, except possibly for a subset of Ω having probability zero.*

If $p(\boldsymbol{\omega})$ is a (prior-posterior-predictive) density, we refer to highest (prior-posterior-predictive) density regions.

Clearly, the credible region approach to summarising $p(\boldsymbol{\omega})$ is not particularly useful in the case of discrete Ω, since such regions will only exist for limited choices of α. The above development should therefore be understood as intended for the case of continuous Ω.

For a number of commonly occurring univariate forms of $p(\boldsymbol{\omega})$, there exist tables which facilitate the identification of HPD intervals for a range of values of α

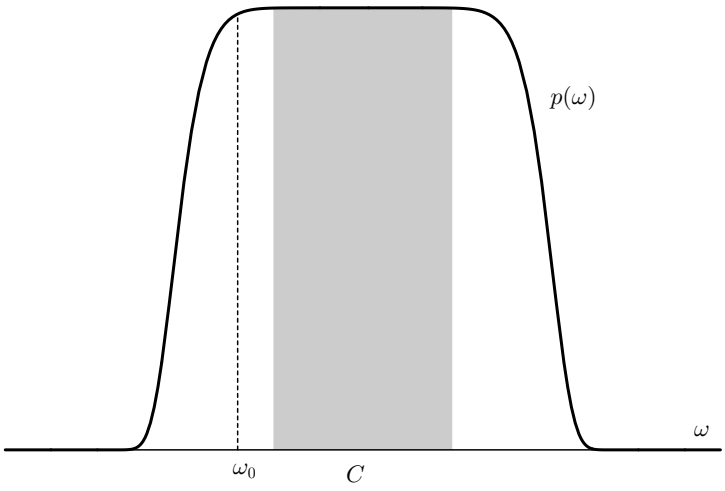

Figure 5.1a *ω_0 almost as "plausible" as all $\omega \in C$*

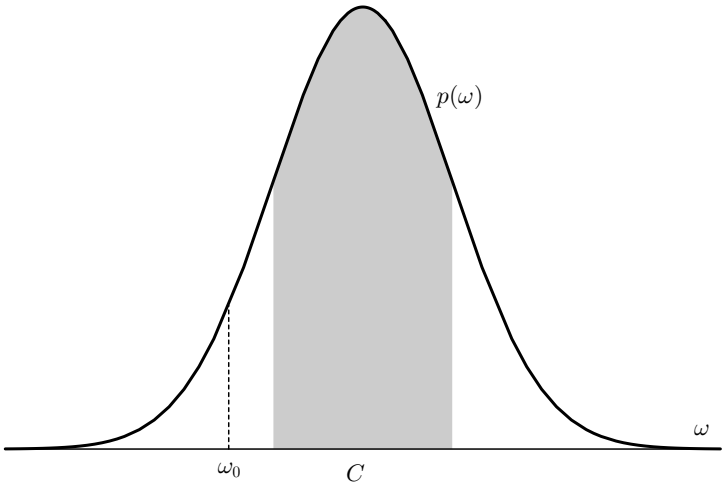

Figure 5.1b *ω_0 much less "plausible" than most $\omega \in C$*

(see, for example, Isaacs *et al.*, 1974, Ferrándiz and Sendra,1982, and Lindley and Scott, 1985).

In general, however, the derivation of an HPD region requires numerical calculation and, particularly if $p(\boldsymbol{\omega})$ does not exhibit markedly skewed behaviour, it may be satisfactory in practice to quote some more simply calculated credible re-

gion. For example, in the univariate case, conventional statistical tables facilitate the identification of intervals which exclude equi-probable tails of $p(\boldsymbol{\omega})$ for many standard distributions.

Although an appropriately chosen selection of credible regions can serve to give a useful summary of $p(\boldsymbol{\omega})$ when we focus just on the quantity $\boldsymbol{\omega}$, there is a fundamental difficulty which prevents such regions serving, in general, as a proxy for the actual density $p(\boldsymbol{\omega})$. The problem is that of lack of invariance under parameter transformation. Even if $\boldsymbol{v} = g(\boldsymbol{\omega})$ is a one-to-one transformation, it is easy to see that there is no general relation between HPD regions for $\boldsymbol{\omega}$ and $\boldsymbol{v}$. In addition, there is no way of identifying a marginal HPD region for a (possibly transformed) subset of components of $\boldsymbol{\omega}$ from knowledge of the joint HPD region.

In cases where an HPD credible region C is pragmatically acceptable as a crude summary of the density $p(\boldsymbol{\omega})$, then, particularly for small values of α (for example, 0.05, 0.01), a specific value $\boldsymbol{\omega}_0 \in \Omega$ will tend to be regarded as somewhat "implausible" if $\boldsymbol{\omega}_0 \notin C$. This, of course, provides no justification for actions such as "rejecting the hypothesis that $\boldsymbol{\omega} = \boldsymbol{\omega}_0$". If we wish to consider such actions, we must formulate a proper decision problem, specifying alternative actions and the losses consequent on correct and incorrect actions. Inferences about a specific hypothesised value $\boldsymbol{\omega}_0$ of a random quantity $\boldsymbol{\omega}$ in the absence of alternative hypothesised values are often considered in the general statistical literature under the heading of "significance testing". We shall discuss this further in Chapter 6.

For the present, it will suffice to note—as illustrated in Figure 5.1—that even the intuitive notion of "implausibility if $\boldsymbol{\omega}_0 \notin C$" depends much more on the complete characterisation of $p(\boldsymbol{\omega})$ than on an either-or assessment based on an HPD region.

For further discussion of credible regions see, for example, Pratt (1961), Aitchison (1964, 1966), Wright (1986) and DasGupta (1991).

Hypothesis Testing

The basic hypothesis testing problem usually considered by statisticians may be described as a decision problem with elements

$$\Omega = \{\omega_0 = [H_0 : \boldsymbol{\theta} \in \Theta_0], \quad \omega_1 = [H_1 : \boldsymbol{\theta} \in \Theta_1]\},$$

together with $p(\boldsymbol{\omega})$, where $\boldsymbol{\theta} \in \Theta = \Theta_0 \cup \Theta_1$, is the parameter labelling a parametric model, $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, $\mathcal{A} = \{a_0, a_1\}$, with $a_1(a_0)$ corresponding to rejecting hypothesis $H_0(H_1)$, and loss function $l(a_i, \omega_j) = l_{ij}$, $i, j \in \{0, 1\}$, with the l_{ij} reflecting the relative seriousness of the four possible consequences and, typically, $l_{00} = l_{11} = 0$.

Clearly, the main motivation and the principal use of the hypothesis testing framework is in model choice and comparison, an activity which has a somewhat different flavour from decision-making and inference within the context of an accepted model. For this reason, we shall postpone a detailed consideration of the

topic until Chapter 6, where we shall provide a much more general perspective on model choice and criticism.

General discussions of Bayesian hypothesis testing are included in Jeffreys (1939/1961), Good (1950, 1965, 1983), Lindley (1957, 1961b, 1965, 1977), Edwards *et al.* (1963), Pratt (1965), Smith (1965), Farrell (1968), Dickey (1971, 1974, 1977), Lempers (1971), Rubin (1971), Zellner (1971), DeGroot (1973), Leamer (1978), Box (1980), Shafer (1982b), Gilio and Scozzafava (1985), Smith, (1986), Berger and Delampady (1987), Berger and Sellke (1987) and Hodges (1990, 1992).

5.1.6 Implementation Issues

Given a likelihood $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ and prior density $p(\boldsymbol{\theta})$, the starting point for any form of parametric inference summary or decision about $\boldsymbol{\theta}$ is the joint posterior density

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} ,$$

and the starting point for any predictive inference summary or decision about future observables $\boldsymbol{y}$ is the predictive density

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \int p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \boldsymbol{x})\, d\boldsymbol{\theta}.$$

It is clear that to form these posterior and predictive densities there is a technical requirement to perform integrations over the range of $\boldsymbol{\theta}$. Moreover, further summarisation, in order to obtain marginal densities, or marginal moments, or expected utilities or losses in explicitly defined decision problems, will necessitate further integrations with respect to components of $\boldsymbol{\theta}$ or $\boldsymbol{y}$, or transformations thereof.

The key problem in implementing the formal Bayes solution to inference reporting or decision problems is therefore seen to be that of evaluating the required integrals. In cases where the likelihood just involves a single parameter, implementation just involves integration in one dimension and is essentially trivial. However, in problems involving a multiparameter likelihood the task of implementation is anything but trivial, since, if $\boldsymbol{\theta}$ has k components, two k-dimensional integrals are required just to form $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ and $p(\boldsymbol{y} \mid \boldsymbol{x})$. Moreover, in the case of $p(\boldsymbol{\theta} \mid \boldsymbol{x})$, for example, k $(k-1)$-dimensional integrals are required to obtain univariate marginal density values or moments, $\binom{k}{2}$ $(k-2)$-dimensional integrals are required to obtain bivariate marginal densities, and so on. Clearly, if k is at all large, the problem of implementation will, in general, lead to challenging technical problems, requiring simultaneous analytic or numerical approximation of a number of multidimensional integrals.

The above discussion has assumed a given specification of a likelihood and prior density function. However, as we have seen in Chapter 4, although a specific mathematical form for the likelihood in a given context is very often implied

or suggested by consideration of symmetry, sufficiency or experience, the mathematical specification of prior densities is typically more problematic. Some of the problems involved—such as the pragmatic strategies to be adopted in translating actual beliefs into mathematical form—relate more to practical methodology than to conceptual and theoretical issues and will be not be discussed in detail in this volume. However, many of the other problems of specifying prior densities are closely related to the general problems of implementation described above, as exemplified by the following questions:

(i) given that, for any specific beliefs, there is some arbitrariness in the precise choice of the mathematical representation of a prior density, are there choices which enable the integrations required to be carried out straightforwardly and hence permit the tractable implementation of a range of analyses, thus facilitating the kind of interpersonal analysis and scientific reporting referred to in Section 4.8.2 and again later in 6.3.3?

(ii) if the information to be provided by the data is known to be far greater than that implicit in an individual's prior beliefs, is there any necessity for a precise mathematical representation of the latter, or can a Bayesian implementation proceed purely on the basis of this qualitative understanding?

(iii) either in the context of interpersonal analysis, or as a special form of actual individual analysis, is there a formal way of representing the beliefs of an individual whose prior information is to be regarded as minimal, relative to the information provided by the data?

(iv) for general forms of likelihood and prior density, are there analytic/numerical techniques available for approximating the integrals required for implementing Bayesian methods?

Question (i) will be answered in Section 5.2, where the concept of a *conjugate* prior density will be introduced.

Question (ii) will be answered in part at the end of Section 5.2 and in more detail in Section 5.3, where an approximate "large sample" Bayesian theory involving *asymptotic posterior normality* will be presented.

Question (iii) will be answered in Section 5.4, where the information-based concept of a *reference* prior density will be introduced. An extended historical discussion of this celebrated philosophical problem of how to represent "ignorance" will be given in Section 5.6.2.

Question (iv) will be answered in Section 5.5, where classical applied analysis techniques such as *Laplace's approximation* for integrals will be briefly reviewed in the context of implementing Bayesian inference and decision summaries, together with classical numerical analytical techniques such as *Gauss-Hermite quadrature* and stochastic simulation techniques such as *importance sampling*, *sampling-importance-resampling* and *Markov chain Monte Carlo*.

5.2 CONJUGATE ANALYSIS

5.2.1 Conjugate Families

The first issue raised at the end of Section 5.1.6 is that of tractability. Given a likelihood function $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, for what choices of $p(\boldsymbol{\theta})$ are integrals such as

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \qquad \text{and} \qquad p(\boldsymbol{y} \mid \boldsymbol{x}) = \int p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{\theta}$$

easily evaluated analytically? However, since any particular mathematical form of $p(\boldsymbol{\theta})$ is acting as a representation of beliefs—either of an actual individual, or as part of a stylised sensitivity study involving a range of prior to posterior analyses—we require, in addition to tractability, that the class of mathematical functions from which $p(\boldsymbol{\theta})$ is to be chosen be both rich in the forms of beliefs it can represent and also facilitate the matching of beliefs to particular members of the class. Tractability can be achieved by noting that, since Bayes' theorem may be expressed in the form

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}),$$

both $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ and $p(\boldsymbol{\theta})$ can be guaranteed to belong to the same general family of mathematical functions by choosing $p(\boldsymbol{\theta})$ to have the same "structure" as $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, when the latter is viewed as a function of $\boldsymbol{\theta}$. However, as stated, this is a rather vacuous idea, since $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ and $p(\boldsymbol{\theta})$ would always belong to the same "general family" of functions if the latter were suitably defined. To achieve a more meaningful version of the underlying idea, let us first recall (from Section 4.5) that if $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$ is a sufficient statistic we have

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}) = p(\boldsymbol{\theta} \mid \boldsymbol{t}) \propto p(\boldsymbol{t} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}),$$

so that we can restate our requirement for tractability in terms of $p(\boldsymbol{\theta})$ having the same structure as $p(\boldsymbol{t} \mid \boldsymbol{\theta})$, when the latter is viewed as a function of $\boldsymbol{\theta}$. Again, however, without further constraint on the nature of the sequence of sufficient statistics the class of possible functions $p(\boldsymbol{\theta})$ is too large to permit easily interpreted matching of beliefs to particular members of the class. This suggests that it is only in the case of likelihoods admitting sufficient statistics of fixed dimension that we shall be able to identify a family of prior densities which ensures both tractability and ease of interpretation. This motivates the following definition.

Definition 5.6. (***Conjugate prior family***). *The conjugate family of prior densities for $\boldsymbol{\theta} \in \Theta$, with respect to a likelihood $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ with sufficient statistic $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x}) = \{n, \boldsymbol{s}(\boldsymbol{x})\}$ of a fixed dimension k independent of that of $\boldsymbol{x}$, is*

$$\{p(\boldsymbol{\theta} \mid \boldsymbol{\tau}), \boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_k) \in \mathcal{T}\},$$

where

$$\mathcal{T} = \left\{ \boldsymbol{\tau}; \int_{\Theta} p(\boldsymbol{s} = (\tau_1, \ldots, \tau_k) \,|\, \boldsymbol{\theta}, n = \tau_0) d\boldsymbol{\theta} < \infty \right\}$$

and

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{\tau}) = \frac{p(\boldsymbol{s} = (\tau_1, \ldots, \tau_k) \,|\, \boldsymbol{\theta}, n = \tau_0)}{\int_{\Theta} p(\boldsymbol{s} = (\tau_1, \ldots, \tau_k) \,|\, \boldsymbol{\theta}, n = \tau_0) d\boldsymbol{\theta}} \,.$$

From Section 4.5 and Definition 5.6, it follows that the likelihoods for which conjugate prior families exist are those corresponding to general exponential family parametric models (Definitions 4.10 and 4.11), for which, given f, $\boldsymbol{h}$, $\boldsymbol{\phi}$ and $\boldsymbol{c}$,

$$p(x \,|\, \boldsymbol{\theta}) = f(x) g(\boldsymbol{\theta}) \exp \left\{ \sum_{i=1}^{k} c_i \phi_i(\boldsymbol{\theta}) h_i(x) \right\}, \quad x \in X,$$

$$\big(g(\boldsymbol{\theta})\big)^{-1} = \int_X f(x) \exp \left\{ \sum_{i=1}^{k} c_i \phi_i(\boldsymbol{\theta}) h_i(x) \right\} dx.$$

The exponential family model is referred to as regular or non-regular, respectively, according as X does not or does depend on $\boldsymbol{\theta}$.

Proposition 5.4. (***Conjugate families for regular exponential families***). *If $\boldsymbol{x} = (x_1, \ldots, x_n)$ is a random sample from a regular exponential family distribution such that*

$$p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) = \prod_{j=1}^{n} f(x_j) \, [g(\boldsymbol{\theta})]^n \exp \left\{ \sum_{i=1}^{k} c_i \phi_i(\boldsymbol{\theta}) \left(\sum_{j=1}^{n} h_i(x_j) \right) \right\},$$

then the conjugate family for $\boldsymbol{\theta}$ has the form

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{\tau}) = [K(\boldsymbol{\tau})]^{-1} [g(\boldsymbol{\theta})]^{\tau_0} \exp \left\{ \sum_{i=1}^{k} c_i \phi_i(\boldsymbol{\theta}) \tau_i \right\}, \quad \boldsymbol{\theta} \in \Theta,$$

where $\boldsymbol{\tau}$ is such that $K(\boldsymbol{\tau}) = \int_{\Theta} [g(\boldsymbol{\theta})]^{\tau_0} \exp \left\{ \sum_{i=1}^{k} c_i \phi_i(\boldsymbol{\theta}) \tau_i \right\} d\boldsymbol{\theta} < \infty$.

Proof. By Proposition 4.10 (the Neyman factorisation criterion), the sufficient statistics for $\boldsymbol{\phi}$ have the form

$$\boldsymbol{t}_n(x_1, \ldots, x_n) = \left[n, \sum_{j=1}^{n} h_1(x_j), \ldots, \sum_{j=1}^{n} h_k(x_j) \right] = [n, \boldsymbol{s}(\boldsymbol{x})],$$

so that, for any $\boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_n)$ such that $\int_\Theta p(\boldsymbol{\theta} \mid \boldsymbol{\tau}) d\boldsymbol{\theta} < \infty$, a conjugate prior density has the form

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{\tau}) &\propto p(s_1(\boldsymbol{x}) = \tau_1, \ldots, s_k(\boldsymbol{x}) = \tau_k \mid \boldsymbol{\theta}, n = \tau_0) \\ &\propto [g(\boldsymbol{\theta})]^{\tau_0} \exp\left\{ \sum_{i=1}^{k} c_i \phi_i(\boldsymbol{\theta}) \tau_i \right\} \end{aligned}$$

by Proposition 4.2. ◁

Example 5.4. ***(Bernoulli likelihood; beta prior)***. The Bernoulli likelihood has the form

$$\begin{aligned} p(x_1, \ldots x_n \mid \theta) &= \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \qquad (0 \le \theta \le 1) \\ &= (1-\theta)^n \exp\left\{ \log\left(\frac{\theta}{1-\theta} \right) \sum_{i=1}^{n} x_i \right\}, \end{aligned}$$

so that, by Proposition 5.4, the conjugate prior density for θ is given by

$$\begin{aligned} p(\theta \mid \tau_0, \tau_1) &\propto (1-\theta)^{\tau_0} \exp\left\{ \log\left(\frac{\theta}{1-\theta} \right) \tau_1 \right\} \\ &= \frac{1}{K(\tau_0, \tau_1)} \theta^{\tau_1} (1-\theta)^{\tau_0 - \tau_1}, \end{aligned}$$

assuming the existence of

$$K(\tau_0, \tau_1) = \int_0^1 \theta^{\tau_1} (1-\theta)^{\tau_0 - \tau_1} \, d\theta.$$

Writing $\alpha = \tau_1 + 1$, and $\beta = \tau_0 - \tau_1 + 1$, we have $p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$, and hence, comparing with the definition of a beta density,

$$p(\theta \mid \tau_0, \tau_1) = p(\theta \mid \alpha, \beta) = \text{Be}(\theta \mid \alpha, \beta), \qquad \alpha > 0, \quad \beta > 0.$$

Example 5.5. ***(Poisson likelihood; gamma prior)***. The Poisson likelihood has the form

$$\begin{aligned} p(x_1, \ldots, x_n \mid \theta) &= \prod_{i=1}^{n} \frac{\theta^{x_i} \exp(-\theta)}{x_i!} \qquad (\theta > 0) \\ &= \left(\prod_{i=1}^{n} x_i! \right)^{-1} \exp(-n\theta) \exp\left(\log\theta \sum_{i=1}^{n} x_i \right), \end{aligned}$$

so that, by Proposition 5.4, the conjugate prior density for θ is given by

$$\begin{aligned} p(\theta \mid \tau_0, \tau_1) &\propto \exp(-\tau_0\theta) \exp(\tau_1 \log\theta) \\ &= \frac{1}{K(\tau_0, \tau_1)} \theta^{\tau_1} \exp(-\tau_0\theta), \end{aligned}$$

assuming the existence of

$$K(\tau_0, \tau_1) = \int_0^\infty \theta^{\tau_1} \exp(-\tau_0 \theta)\, d\theta.$$

Writing $\alpha = \tau_1 + 1$ and $\beta = \tau_0$ we have $p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1} \exp(-\beta\theta)$ and hence, comparing with the definition of a gamma density,

$$p(\theta \mid \tau_0, \tau_1) = p(\theta \mid \alpha, \beta) = \text{Ga}(\theta \mid \alpha, \beta), \qquad \alpha > 0, \quad \beta > 0.$$

Example 5.6. ***(Normal likelihood; normal-gamma prior)***. The normal likelihood, with unknown mean and precision, has the form

$$\begin{aligned} p(x_1, \ldots, x_n \mid \mu, \lambda) &= \prod_{i=1}^n \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda}{2}(x_i - \mu)^2\right\} \\ &= (2\pi)^{-n/2} \left[\lambda^{1/2} \exp\left(-\frac{\lambda}{2}\mu^2\right)\right]^n \exp\left\{\mu\lambda \sum_{i=1}^n x_i - \frac{\lambda}{2}\sum_{i=1}^n x_i^2\right\}, \end{aligned}$$

so that, by Proposition 5.4, the conjugate prior density for $\theta = (\mu, \lambda)$ is given by

$$\begin{aligned} p(\mu, \lambda \mid \tau_0, \tau_1, \tau_2) &\propto \left[\lambda^{1/2} \exp\left(-\frac{1}{2}\lambda\mu^2\right)\right]^{\tau_0} \exp\left\{\mu\lambda\tau_1 - \frac{1}{2}\lambda\tau_2\right\} \\ &= \frac{1}{K(\tau_0, \tau_1, \tau_2)} \lambda^{(\tau_0 - 1)/2} \exp\left(-\frac{\lambda}{2}(\tau_2 - \frac{\tau_1^2}{\tau_0})\right) \lambda^{\frac{1}{2}} \exp\left\{-\frac{\lambda\tau_0}{2}\left(\mu - \frac{\tau_1}{\tau_0}\right)^2\right\}, \end{aligned}$$

assuming the existence of $K(\tau_0, \tau_1, \tau_2)$, given by

$$\int_0^\infty \lambda^{\frac{\tau_0 - 1}{2}} \exp\left(-\frac{\lambda}{2}(\tau_2 - \frac{\tau_1^2}{\tau_0})\right) \left\{\int_{-\infty}^\infty \lambda^{\frac{1}{2}} \exp\left[-\frac{\lambda\tau_0}{2}\left(\mu - \frac{\tau_1}{\tau_0}\right)^2\right] d\mu\right\} d\lambda.$$

Writing $\alpha = \frac{1}{2}(\tau_0 + 1)$, $\beta = \frac{1}{2}(\tau_2 - \frac{\tau_1^2}{\tau_0})$, $\gamma = \tau_1/\tau_0$, and comparing with the definition of a normal-gamma density, we have

$$\begin{aligned} p(\mu, \lambda \mid \tau_0, \tau_1, \tau_2) &= p(\mu, \lambda \mid \alpha, \beta, \gamma) \\ &= \text{Ng}(\mu, \lambda \mid \alpha, \beta, \gamma) \\ &= \text{N}\Big(\mu \mid \gamma, \lambda(2\alpha - 1)\Big)\text{Ga}(\lambda \mid \alpha, \beta), \end{aligned}$$

with $\alpha > \frac{1}{2}$, $\beta > 0$, $\gamma \in \Re$.

5.2.2 Canonical Conjugate Analysis

Conjugate prior density families were motivated by considerations of tractability in implementing the Bayesian paradigm. The following proposition demonstrates that, in the case of regular exponential family likelihoods and conjugate prior densities, the analytic forms of the joint posterior and predictive densities which underlie any form of inference summary or decision making are easily identified.

Proposition 5.5. (***Conjugate analysis for regular exponential families***). *For the exponential family likelihood and conjugate prior density of Proposition 5.4:*

(i) *the posterior density for* $\boldsymbol{\theta}$ *is*

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{\tau}) \underset{n}{=} p(\boldsymbol{\theta} \mid \boldsymbol{\tau} + \boldsymbol{t}_n(\boldsymbol{x}))$$

where

$$\boldsymbol{\tau} + \boldsymbol{t}_n(\boldsymbol{x}) = \left(\tau_0 + n,\ \tau_1 + \sum_{j=1}^{n} h_1(x_j), \ldots, \tau_k + \sum_{j=1}^{n} h_k(x_j) \right);$$

(ii) *the predictive density for future observables* $\boldsymbol{y} = (y_1, \ldots, y_m)$ *is*

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\tau}) &= p(\boldsymbol{y} \mid \boldsymbol{\tau} + \boldsymbol{t}_n(\boldsymbol{x})) \\ &= \prod_{l=1}^{m} f(y_l) \frac{K(\boldsymbol{\tau} + \boldsymbol{t}_n(\boldsymbol{x}) + \boldsymbol{t}_m(\boldsymbol{y}))}{K(\boldsymbol{\tau} + \boldsymbol{t}_n(\boldsymbol{x}))}, \end{aligned}$$

where $\boldsymbol{t}_m(\boldsymbol{y}) = [m,\ \sum_{l=1}^{m} h_1(y_l), \ldots, \sum_{l=1}^{m} h_k(y_l)]$.

Proof. By Bayes' theorem,

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{\tau}) &\propto p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\tau}) \\ &\propto [g(\boldsymbol{\theta})]^{\tau_0 + n} \exp\left\{ \sum_{i=1}^{k} c_i \phi_i(\boldsymbol{\theta}) \left(\tau_i + \sum_{j=1}^{n} h_i(x_j) \right) \right\} \\ &\propto p(\boldsymbol{\theta} \mid \boldsymbol{\tau} + \boldsymbol{t}_n(\boldsymbol{x})), \end{aligned}$$

which proves (i). Moreover,

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\tau}) &= \int_{\Theta} p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{\theta} \\ &= \prod_{l=1}^{m} f(y_l) \cdot [K(\boldsymbol{\tau} + \boldsymbol{t}_n(\boldsymbol{x}))]^{-1} \int_{\Theta} [g(\boldsymbol{\theta})]^{\tau_0 + n + m} \\ &\qquad \times \exp\left\{ \sum_{i=1}^{k} c_i \phi_i(\boldsymbol{\theta}) \left(\tau_i + \sum_{j=1}^{n} h_i(x_j) + \sum_{l=1}^{m} h_i(y_l) \right) \right\} d\boldsymbol{\theta} \\ &= \prod_{l=1}^{m} f(y_l) \frac{K\big(\boldsymbol{\tau} + \boldsymbol{t}_n(\boldsymbol{x}) + \boldsymbol{t}_m(\boldsymbol{y})\big)}{K\big(\boldsymbol{\tau} + \boldsymbol{t}_n(\boldsymbol{x})\big)}, \end{aligned}$$

which proves (ii). ◁

Proposition 5.5(i) establishes that the conjugate family is *closed under sampling*, with respect to the corresponding exponential family likelihood, a concept which seems to be due to G. A. Barnard. This means that both the joint prior and posterior densities belong to the same, simply defined, family of distributions, the inference process being totally defined by the mapping $\boldsymbol{\tau} \to (\boldsymbol{\tau} + \boldsymbol{t}_n(\boldsymbol{x}))$, under which the labelling parameters of the prior density are simply modified by the addition of the values of the sufficient statistic to form the labelling parameter of the posterior distribution. The inference process defined by Bayes' theorem is therefore reduced from the essentially infinite-dimensional problem of the transformation of density functions, to a simple, additive finite-dimensional transformation. Proposition 5.5(ii) establishes that a similar, simplifying closure property holds for predictive densities.

The forms arising in the conjugate analysis of a number of standard exponential family forms are summarised in Appendix A. However, to provide some preliminary insights into the prior → posterior → predictive process described by Proposition 5.5, we shall illustrate the general results by reconsidering Example 5.4.

Example 5.4. ***(continued)***. With the Bernoulli likelihood written in its explicit exponential family form, and writing $r_n = x_1 + \cdots + x_n$, the posterior density corresponding to the conjugate prior density, $p(\theta \mid \tau_0, \tau_1)$, is given by

$$\begin{aligned} p(\theta \mid \boldsymbol{x}, \tau_0, \tau_1) &\propto p(\boldsymbol{x} \mid \theta) p(\theta \mid \tau_0, \tau_1) \\ &\propto (1-\theta)^n \exp\left\{ \log\left(\frac{\theta}{1-\theta}\right) r_n \right\} (1-\theta)^{\tau_0} \\ &\qquad \times \exp\left\{ \log\left(\frac{\theta}{1-\theta}\right) \tau_1 \right\} \\ &= \frac{\Gamma(\tau_0(n)+2)}{\Gamma(\tau_1(n)+1)\Gamma(\tau_0(n)-\tau_1(n)+1)} (1-\theta)^{\tau_0(n)} \\ &\qquad \times \exp\left\{ \log\left(\frac{\theta}{1-\theta}\right) \tau_1(n) \right\}, \end{aligned}$$

where $\tau_0(n) = \tau_0 + n$, $\tau_1(n) = \tau_1 + r_n$, showing explicitly how the inference process reduces to the updating of the prior to posterior hyperparameters by the addition of the sufficient statistics, n and r_n.

Alternatively, we could proceed on the basis of the original representation of the Bernoulli likelihood, combining it directly with the familiar beta prior density, $\text{Be}(\theta \mid \alpha, \beta)$, so that

$$\begin{aligned} p(\theta \mid \boldsymbol{x}, \alpha, \beta) &\propto p(\boldsymbol{x} \mid \theta) p(\theta \mid \alpha, \beta) \\ &\propto \theta^{r_n} (1-\theta)^{n-r_n} \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &= \frac{\Gamma(\alpha_n + \beta_n)}{\Gamma(\alpha_n)\Gamma(\beta_n)} \theta^{\alpha_n - 1} (1-\theta)^{\beta_n - 1}, \end{aligned}$$

where $\alpha_n = \alpha + r_n$, $\beta_n = \beta + n - r_n$ and, again, the process reduces to the updating of the prior to posterior hyperparameters.

Clearly, the two notational forms and procedures used in the example are equivalent. Using the standard exponential family form has the advantage of displaying the simple hyperparameter updating by the addition of the sufficient statistics. However, the second form seems much less cumbersome notationally and is more transparently interpretable and memorable in terms of the beta density.

In general, when analysing particular models we shall work in terms of whatever functional representation seems best suited to the task in hand.

Example 5.4. ***(continued)***. Instead of working with the original Bernoulli likelihood, $p(x_1, \ldots, x_n|\theta)$, we could, of course, work with a likelihood defined in terms of the sufficient statistic (n, r_n). In particular, if either n or r_n were ancillary, we would use one or other of $p(r_n \,|\, n, \theta)$ or $p(n \,|\, r_n, \theta)$ and, in either case,

$$p(\theta \,|\, n, r_n, \alpha, \beta) \propto \theta^{r_n}(1-\theta)^{n-r_n}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Taking the binomial form, $p(r_n \,|\, n, \theta)$, the prior to posterior operation defined by Bayes' theorem can be simply expressed, in terms of the notation introduced in Section 3.2.4, as

$$\mathrm{Bi}(r_n \,|\, \theta, n) \otimes \mathrm{Be}(\theta \,|\, \alpha, \beta) \equiv \mathrm{Be}(\theta \,|\, \alpha + r_n, \beta + n - r_n).$$

The predictive density for future Bernoulli observables, which we denote by

$$\boldsymbol{y} = (y_1, \ldots, y_m) = (x_{n+1}, \ldots, x_{n+m}),$$

is also easily derived. Writing $r'_m = y_1 + \cdots + y_m$, we see that

$$\begin{aligned} p(\boldsymbol{y} \,|\, \boldsymbol{x}, \alpha, \beta) &= p(\boldsymbol{y} \,|\, \alpha_n, \beta_n) \\ &= \int_0^1 p(\boldsymbol{y} \,|\, \theta) p(\theta \,|\, \alpha_n, \beta_n)\, d\theta \\ &= \frac{\Gamma(\alpha_n + \beta_n)}{\Gamma(\alpha_n)\Gamma(\beta_n)} \int_0^1 \theta^{\alpha_n + r'_m - 1}(1-\theta)^{\beta_n + m - r'_m - 1}\, d\theta \\ &= \frac{\Gamma(\alpha_n + \beta_n)}{\Gamma(\alpha_n)\Gamma(\beta_n)} \frac{\Gamma(\alpha_{n+m})\Gamma(\beta_{n+m})}{\Gamma(\alpha_{n+m} + \beta_{n+m})}\,, \end{aligned}$$

where

$$\begin{aligned} \alpha_{n+m} &= \alpha_n + r'_m = \alpha + r_n + r'_m, \\ \beta_{n+m} &= \beta_n + m - r'_m = \beta + (n+m) - (r_n + r'_m), \end{aligned}$$

a result which also could be obtained directly from Proposition 5.5(ii).

If, instead, we were interested in the predictive density for r'_m, it easily follows that

$$\begin{aligned} p(r'_m \,|\, \alpha_n, \beta_n, m) &= \int_0^1 p(r'_m \,|\, m, \theta) p(\theta \,|\, \alpha_n, \beta_n)\; d\theta \\ &= \int_0^1 \binom{m}{r'_m} p(\boldsymbol{y} \,|\, \theta) p(\theta \,|\, \alpha_n, \beta_n)\, d\theta \\ &= \binom{m}{r'_m} p(\boldsymbol{y} \,|\, \alpha_n, \beta_n). \end{aligned}$$

Comparison with Section 3.2.2 reveals this predictive density to have the binomial-beta form, $\text{Bb}(r'_m \mid \alpha_n, \beta_n, m)$.

The particular case $m = 1$ is of some interest, since $p(r'_m = 1 \mid \alpha_n, \beta_n, m = 1)$ is then the predictive probability assigned to a success on the $(n+1)$th trial, given r_n observed successes in the first n trials and an initial $\text{Be}(\theta \mid \alpha, \beta)$ belief about the limiting relative frequency of successes, θ.

We see immediately, on substituting into the above, that

$$p(r'_m = 1 \mid \alpha_n, \beta_n, m = 1) = \frac{\alpha_n}{\alpha_n + \beta_n} = E(\theta \mid \alpha_n, \beta_n),$$

using the fact that $\Gamma(t+1) = t\Gamma(t)$ and recalling, from Section 3.2.2, the form of the mean of a beta distribution.

With respect to quadratic loss, $E(\theta \mid \alpha_n, \beta_n) = (\alpha + r_n)/(\alpha + \beta + n)$ is the optimal estimate of θ given current information, and the above result demonstrates that this should serve as the evaluation of the probability of a success on the next trial. In the case $\alpha = \beta = 1$ this evaluation becomes $(r_n + 1)/(n + 2)$, which is the celebrated *Laplace's rule of succession* (Laplace, 1812), which has served historically to stimulate considerable philosophical debate about the nature of inductive inference. We shall consider this problem further in Example 5.16 of Section 5.4.4. For an elementary, but insightful, account of Bayesian inference for the Bernoulli case, see Lindley and Phillips (1976).

In presenting the basic ideas of conjugate analysis, we used the following notation for the k-parameter exponential family and corresponding prior form:

$$p(x \mid \boldsymbol{\theta}) = f(x)g(\boldsymbol{\theta}) \exp\left\{\sum_{i=1}^{k} c_i \phi_i(\boldsymbol{\theta}) h_i(x)\right\}, \quad x \in X,$$

and

$$p(\boldsymbol{\theta} \mid \boldsymbol{\tau}) = [K(\boldsymbol{\tau})]^{-1}[g(\boldsymbol{\theta})]^{\tau_0} \exp\left\{\sum_{i=1}^{k} \phi_i(\boldsymbol{\theta})\tau_i\right\}, \quad \theta \in \Theta,$$

the latter being defined for $\boldsymbol{\tau}$ such that $K(\boldsymbol{\tau}) < \infty$.

From a notational perspective (cf. Definition 4.12), we can obtain considerable simplification by defining $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_k)$, $\boldsymbol{y} = (y_1, \ldots, y_k)$, where $\psi_i = c_i\phi_i(\boldsymbol{\theta})$ and $y_i = h(x_i)$, $i = 1, \ldots, k$, together with prior hyperparameters n_0, $\boldsymbol{y}_0$, so that these forms become

$$p(\boldsymbol{y} \mid \boldsymbol{\psi}) = a(\boldsymbol{y}) \exp\left\{\boldsymbol{y}^t\boldsymbol{\psi} - b(\boldsymbol{\psi})\right\}, \quad \boldsymbol{y} \in Y,$$
$$p(\boldsymbol{\psi} \mid n_0, \boldsymbol{y}_0) = c(n_0, \boldsymbol{y}_0) \exp\left\{n_0\boldsymbol{y}_0^t\boldsymbol{\psi} - n_0 b(\boldsymbol{\psi})\right\}, \quad \boldsymbol{\psi} \in \Psi,$$

for appropriately defined Y, Ψ and real-valued functions a, b and c. We shall refer to these (Definition 4.12) as the *canonical (or natural) forms* of the exponential family and its conjugate prior family. If $\Psi = \Re^k$, we require $n_0 > 0$, $\boldsymbol{y}_0 \in Y$

in order for $p(\boldsymbol{\psi} \mid n_0, \boldsymbol{y}_0)$ to be a proper density; for $\Psi \neq \Re^k$, the situation is somewhat more complicated (see Diaconis and Ylvisaker, 1979, for details). We shall typically assume that Ψ consists of all $\boldsymbol{\psi}$ such that $\int_Y p(\boldsymbol{y} \mid \boldsymbol{\psi}) d\boldsymbol{y} = 1$ and that $b(\boldsymbol{\psi})$ is continuously differentiable and strictly convex throughout the interior of Ψ.

The motivation for choosing $n_0, \boldsymbol{y}_0$ as notation for the prior hyperparameter is partly clarified by the following proposition and becomes even clearer in the context of Proposition 5.7.

Proposition 5.6. (***Canonical conjugate analysis***). *If $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are the values of $\boldsymbol{y}$ resulting from a random sample of size n from the canonical exponential family parametric model, $p(\boldsymbol{y} \mid \boldsymbol{\psi})$, then the posterior density corresponding to the canonical conjugate form, $p(\boldsymbol{\psi} \mid n_0, \boldsymbol{y}_0)$, is given by*

$$p(\boldsymbol{\psi} \mid n_0, \boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = p\left(\boldsymbol{\psi} \;\middle|\; n_0 + n,\ \frac{n_0 \boldsymbol{y}_0 + n\overline{\boldsymbol{y}}_n}{(n + n_0)}\right),$$

where $\overline{\boldsymbol{y}}_n = \sum_{i=1}^n \boldsymbol{y}_i / n$.

Proof.

$$\begin{aligned} p(\psi \mid n_0, \boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) &\propto \prod_{i=1}^n p(\boldsymbol{y}_i \mid \boldsymbol{\psi}) p(\boldsymbol{\psi} \mid n_0, \boldsymbol{y}_0) \\ &\propto \exp\left\{n\overline{\boldsymbol{y}}_n^t \boldsymbol{\psi} - n b(\boldsymbol{\psi})\right\} \\ &\quad \times \exp\left\{n_0 \boldsymbol{y}_0^t \boldsymbol{\psi} - n_0 b(\boldsymbol{\psi})\right\} \\ &\propto \exp\left\{(n_0 \boldsymbol{y}_0 + n\overline{\boldsymbol{y}}_n)^t \boldsymbol{\psi} - (n_0 + n) b(\boldsymbol{\psi})\right\}, \end{aligned}$$

and the result follows. ◁

Example 5.4. ***(continued)***. In the case of the Bernoulli parametric model, we have seen earlier that the pairing of the parametric model and conjugate prior can be expressed as

$$p(x \mid \theta) = (1 - \theta) \exp\left\{x \log\left(\frac{\theta}{1 - \theta}\right)\right\}$$

$$p(\theta \mid \tau_0, \tau_1) = [K(\boldsymbol{\tau})]^{-1} (1 - \theta)^{\tau_0} \exp\left\{\tau_1 \log\left(\frac{\theta}{1 - \theta}\right)\right\},$$

The canonical forms in this case are obtained by setting

$$y = x, \quad \psi = \log\left(\frac{\theta}{1 - \theta}\right), \quad a(y) = 1, \quad b(\psi) = \log(1 + e^{\psi}),$$

$$c(n_0, y_0) = \frac{\Gamma(n_0 + 2)}{\Gamma(n_0 y_0 + 1)\Gamma(n_0 - n_0 y_0 + 1)},$$

and, hence, the posterior distribution of the canonical parameter ψ is given by

$$p(\psi \mid n_0, y_0, y_1, \ldots, y_n) \propto \exp\left[(n_0 + n)\left\{\frac{n_0 y_0 + n\overline{y}_n}{n + n_0} \psi - b(\psi)\right\}\right].$$

Example 5.5. ***(continued)***. In the case of the Poisson parametric model, we have seen earlier that the pairings of the parametric model and conjugate form can be expressed as

$$p(x \mid \theta) = \frac{1}{x!} \exp(-\theta) \exp(x \log \theta)$$
$$p(\theta \mid \tau_0, \tau_1) = [K(\boldsymbol{\tau})]^{-1} \exp(-\tau_0 \theta) \exp(\tau_1 \log \theta),$$

The canonical forms in this case are obtained by setting

$$y = x, \quad \psi = \log \theta, \quad a(y) = \frac{1}{y!}\,, \quad b(\psi) = e^{\psi}, \quad c(n_0, y_0) = \frac{n_0^{y_0+1}}{\Gamma(y_0+1)}\,.$$

The posterior distribution of the canonical parameter ψ is now immediately given by Proposition 5.6.

Example 5.6. ***(continued)***. In the case of the normal parametric model, we have seen earlier that the pairings of the parametric model and conjugate form can be expressed as

$$p(x \mid \mu, \lambda) = (2\pi)^{-1/2} \left[\lambda^{1/2} \exp\left(-\frac{1}{2}\lambda\mu^2\right)\right] \exp\left\{x(\lambda\mu) - \frac{1}{2}x^2\lambda\right\}$$
$$p(\mu, \lambda \mid \tau_0, \tau_1, \tau_2) = [K(\tau)]^{-1} \left[\lambda^{1/2} \exp\left(-\frac{1}{2}\lambda\mu^2\right)\right]^{\tau_0} \exp\left\{\tau_1(\lambda\mu) - \frac{1}{2}\tau_2\lambda\right\}\,.$$

The canonical forms in this case are obtained by setting

$$\boldsymbol{y} = (y_1, y_2) = (x, x^2), \quad \boldsymbol{\psi} = (\psi_1, \psi_2) = \left(\lambda\mu, -\frac{1}{2}\lambda\right)\,,$$
$$a(\boldsymbol{y}) = (2\pi)^{-1/2}, \quad b(\boldsymbol{\psi}) = \log(-2\psi_2)^{-1/2} - \frac{\psi_1^2}{4\psi_2}\,,$$
$$c(n_0, \boldsymbol{y}_0) = \left(\frac{2\pi}{n_0}\right)^{1/2} \frac{\left(\frac{1}{2}(n_0 y_{02})\right)^{(n_0 y_{01}+1)/2}}{\Gamma\left(\frac{1}{2}(n_0+1)\right)}\,.$$

Again, the posterior distribution of the canonical parameters $\boldsymbol{\psi} = (\psi_1, \psi_2)$ is now immediately given by Proposition 5.6.

For specific applications, the choice of the representation of the parametric model and conjugate prior forms is typically guided by the ease of interpretation of the parametrisations adopted. Example 5.6 above suffices to demonstrate that the canonical forms may be very unappealing. From a *theoretical* perspective, however, the canonical representation often provides valuable unifying insight, as in Proposition 5.6, where the economy of notation makes it straightforward to demonstrate that the learning process just involves a simple weighted average,

$$\frac{n_0 \boldsymbol{y}_0 + n\bar{\boldsymbol{y}}_n}{n_0 + n}\,,$$

of prior and sample information. Again using the canonical forms, we can give a more precise characterisation of this weighted average.

Proposition 5.7. (***Weighted average form of posterior expectation***).
If $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are the values of $\boldsymbol{y}$ resulting from a random sample of size n from the canonical exponential family parametric model,

$$p(\boldsymbol{y} \,|\, \boldsymbol{\psi}) = a(\boldsymbol{y}) \exp\left\{\boldsymbol{y}^t \boldsymbol{\psi} - b(\boldsymbol{\psi})\right\},$$

with canonical conjugate prior $p(\boldsymbol{\psi} \,|\, n_0, \boldsymbol{y}_0)$, then

$$E\left[\nabla b(\boldsymbol{\psi}) \,|\, n_0, \boldsymbol{y}_0, \boldsymbol{y}\right] = \pi \overline{\boldsymbol{y}}_n + (1 - \pi)\boldsymbol{y}_0,$$

where

$$\pi = \frac{n}{n_0 + n}\,, \quad \left[\nabla b(\boldsymbol{\psi})\right]_i = \frac{\partial}{\partial \psi_i} b(\boldsymbol{\psi}).$$

Proof. By Proposition 5.6, it suffices to prove that $E(\nabla b(\psi) \,|\, n_0, \boldsymbol{y}_0) = \boldsymbol{y}_0$. But

$$\begin{aligned} n_0\big[\boldsymbol{y}_0 - E(\nabla b(\boldsymbol{\psi}) \,|\, n_0, \boldsymbol{y}_0)\big] &= \int_{\Psi} n_0\big(\boldsymbol{y}_0 - \nabla b(\boldsymbol{\psi})\big)\, p(\boldsymbol{\psi} \,|\, n_0, \boldsymbol{y}_0)\, d\boldsymbol{\psi} \\ &= \int_{\Psi} \nabla p(\boldsymbol{\psi} \,|\, n_0, \boldsymbol{y}_0)\, d\boldsymbol{\psi}. \end{aligned}$$

This establishes the result. ◁

Proposition 5.7 reveals, in this natural conjugate setting, that the posterior expectation of $\nabla b(\boldsymbol{\psi})$, that is its Bayes estimate with respect to quadratic loss (see Proposition 5.2), is a weighted average of $\boldsymbol{y}_0$ and $\overline{\boldsymbol{y}}_n$. The former is the prior estimate of $\nabla b(\boldsymbol{\psi})$; the latter can be viewed as an intuitively "natural" sample-based estimate of $\nabla b(\boldsymbol{\psi})$, since

$$\begin{aligned} E(\boldsymbol{y} \,|\, \boldsymbol{\psi}) - \nabla b(\boldsymbol{\psi}) &= \int \big(\boldsymbol{y} - \nabla b(\boldsymbol{\psi})\big) p(\boldsymbol{y} \,|\, \boldsymbol{\psi}) d\boldsymbol{y} \\ &= \int \nabla p(\boldsymbol{y} \,|\, \boldsymbol{\psi})\, d\boldsymbol{y} = \nabla \int p(\boldsymbol{y} \,|\, \boldsymbol{\psi}) d\boldsymbol{y} = 0 \end{aligned}$$

and hence $E(\boldsymbol{y} \,|\, \boldsymbol{\psi}) = E(\overline{\boldsymbol{y}}_n \,|\, \boldsymbol{\psi}) = \nabla b(\boldsymbol{\psi})$.

For any given prior hyperparameters, $(n_0, \boldsymbol{y}_0)$, as the sample size n becomes large, the weight, π, tends to one and the sample-based information dominates the posterior. In this context, we make an important point alluded to in our discussion of "objectivity and subjectivity", in Section 4.8.2. Namely, that in the stylised setting of a group of individuals agreeing on an exponential family parametric form, but assigning different conjugate priors, a sufficiently large sample will lead to more or less identical posterior beliefs. Statements based on the latter might well, in common parlance, be claimed to be "objective". One should always be aware, however, that this is no more than a conventional way of indicating a subjective consensus, resulting from a large amount of data processed in the light of a central core of shared assumptions.

Proposition 5.7 shows that conjugate priors for exponential family parameters imply that posterior expectations are linear functions of the sufficient statistics. It is interesting to ask whether other forms of prior specification can also lead to linear posterior expectations. Or, more generally, whether knowing or constraining posterior moments to be of some simple algebraic form suffices to characterise possible families of prior distributions. These kinds of questions are considered in detail in, for example, Diaconis and Ylvisaker (1979) and Goel and DeGroot (1980). In particular, it can be shown, under some regularity conditions, that, for continuous exponential families, linearity of the posterior expectation does imply that the prior must be conjugate.

The weighted average form of posterior mean,

$$E[\nabla b(\boldsymbol{\psi}) \,|\, n_0, \boldsymbol{y}_0, \boldsymbol{y}] = \frac{n_0 \boldsymbol{y}_0 + n\overline{\boldsymbol{y}}_n}{n_0 + n} \,,$$

obtained in Proposition 5.7, and also appearing explicitly in the prior to posterior updating process given in Proposition 5.6 makes clear that the prior parameter, n_0, attached to the prior mean, $\boldsymbol{y}_0$ for $\nabla b(\boldsymbol{\psi})$, plays an analogous role to the sample size, n, attached to the data mean $\overline{\boldsymbol{y}}_n$. The choice of an n_0 which is large relative to n thus implies that the prior will dominate the data in determining the posterior (see, however, Section 5.6.3 for illustration of why a weighted-average form might not be desirable). Conversely, the choice of an n_0 which is small relative to n ensures that the form of the posterior is essentially determined by the data. In particular, this suggests that a tractable analysis which "lets the data speak for themselves" can be obtained by letting $n_0 \to 0$. Clearly, however, this has to be regarded as simply a convenient approximation to the posterior that would have been obtained from the choice of a prior with small, but positive n_0. The choice $n_0 = 0$ typically implies a form of $p(\boldsymbol{\psi} \,|\, n_0, \boldsymbol{y}_0)$ which does not integrate to unity (a so-called *improper* density) and thus cannot be interpreted as representing an actual prior belief. The following example illustrates this use of limiting, improper conjugate priors in the context of the Bernoulli parametric model with beta conjugate prior, using standard rather than canonical forms for the parametric models and prior densities.

Example 5.4. ***(continued)***. We have seen that if $r_n = x_1 + \cdots + x_n$ denotes the number of successes in n Bernoulli trials, the conjugate beta prior density, $\text{Be}(\theta \,|\, \alpha, \beta)$, for the limiting relative frequency of successes, θ, leads to a $\text{Be}(\theta \,|\, \alpha + r_n, \beta + n - r_n)$ posterior for θ, which has expectation

$$\frac{\alpha + r_n}{\alpha + \beta + n} = \pi \left(\frac{r_n}{n}\right) + (1 - \pi) \left(\frac{\alpha}{\alpha + \beta}\right),$$

where $\pi = (\alpha + \beta + n)^{-1} n$, providing a weighted average between the prior mean for θ and the frequency estimate provided by the data. In this notation, $n_0 \to 0$ corresponds to $\alpha \to 0$,

$\beta \to 0$, which implies a Be$(\theta \mid r_n, n - r_n)$ approximation to the posterior distribution, having expectation r_n/n. The limiting prior form, however, would be

$$p(\theta \mid \alpha = 0, \beta = 0) \propto \theta^{-1}(1-\theta)^{-1},$$

which is not a proper density. As a technique for arriving at the approximate posterior distribution, it is certainly convenient to make formal use of Bayes' theorem with this improper form playing the role of a prior, since

$$\begin{aligned} p(\theta \mid \alpha = 0, \beta = 0, n, r_n) &\propto p(r_n \mid n\theta) p(\theta \mid \alpha = 0, \beta = 0) \\ &\propto \theta^{r_n}(1-\theta)^{n-r_n}\theta^{-1}(1-\theta)^{-1} \\ &\propto \text{Be}(\theta \mid r_n, n - r_n). \end{aligned}$$

It is important to recognise, however, that this is merely an approximation device and in no way justifies regarding $p(\theta \mid \alpha = 0, \beta = 0)$ as having any special significance as a representation of "prior ignorance". Clearly, *any* choice of α, β small compared with r_n, $n - r_n$ (for example, $\alpha = \beta = \frac{1}{2}$ or $\alpha = \beta = 1$ for typical values of $r_n, n - r_n$) will lead to an almost identical posterior distribution for θ.

A further problem of interpretation arises if we consider inferences for functions of θ. Consider, for example, the choice $\alpha = \beta = 1$, which implies a uniform prior density for θ. At an intuitive level, it might be argued that this represents "complete ignorance" about θ, which should, presumably, entail "complete ignorance" about any function, $g(\theta)$, of θ. However, $p(\theta)$ uniform implies that $p(g(\theta))$ is not uniform. This makes it clear that *ad hoc* intuitive notions of "ignorance, or of what constitutes a "non-informative" prior distribution (in some sense), cannot be relied upon. There is a need for a more formal analysis of the concept and this will be given in Section 5.4, with further discussion in Section 5.6.2.

Proposition 5.2 established the general forms of Bayes estimates for some commonly used loss functions. Proposition 5.7 provided further insight into the (posterior mean) form arising from quadratic loss in the case of an exponential family parametric model with conjugate prior. Within this latter framework, the following development, based closely on Gutiérrez-Peña (1992), provides further insight into how the posterior mode can be justified as a Bayes estimate.

We recall, from the discussion preceding Proposition 5.6, the canonical forms of the k-parameter exponential family and its corresponding conjugate prior:

$$p(\boldsymbol{y}|\boldsymbol{\psi}) = a(\boldsymbol{y}) \exp\left\{\boldsymbol{y}^t\boldsymbol{\psi} - b(\boldsymbol{\psi})\right\}, \quad \boldsymbol{y} \in Y$$

and

$$p(\boldsymbol{\psi}|n_0, \boldsymbol{y}_0) = c(n_0, \boldsymbol{y}_0) \exp\left\{n_0\boldsymbol{y}_0^t\boldsymbol{\psi} - n_0 b(\boldsymbol{\psi})\right\}, \quad \boldsymbol{\psi} \in \Psi,$$

for appropriately defined Y, Ψ and real-valued functions a, b and c.

Consider $p(\boldsymbol{\psi}|n_0, \boldsymbol{y}_0)$ and define $d(s, \boldsymbol{t}) = -\log c(s, s^{-1}\boldsymbol{t})$, with $s > 0$ and $\boldsymbol{t} \in Y$. Further define

$$\begin{aligned}\nabla d(s, \boldsymbol{t}) &= \left[\frac{\partial d(s, \boldsymbol{t})}{\partial t_1}, \cdots, \frac{\partial d(s, \boldsymbol{t})}{\partial t_k}\right]^t \\ &= [d_1(s, \boldsymbol{t}), \ldots, d_k(s, \boldsymbol{t})]^t\end{aligned}$$

and $d_0(s, \boldsymbol{t}) = \partial d(s, \boldsymbol{t})/\partial s$. As a final preliminary, recall the logarithmic divergence measure

$$\delta(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_0) = \int p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\theta})}{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0)}\, d\boldsymbol{x}$$

between two distributions $p(\boldsymbol{x}|\boldsymbol{\theta})$ and $p(\boldsymbol{x}|\boldsymbol{\theta}_0)$. We can now establish the following technical results.

Proposition 5.8. (***Logarithmic divergence between conjugate distributions***). *With respect to the canonical form of the k-parameter exponential family and its corresponding conjugate prior:*

(i) $\delta(\boldsymbol{\psi}|\boldsymbol{\psi}_0) = b(\boldsymbol{\psi}_0) - b(\boldsymbol{\psi}) + (\boldsymbol{\psi} - \boldsymbol{\psi}_0)^t \nabla b(\boldsymbol{\psi})$*;*

(ii) $E[\delta(\boldsymbol{\psi}|\boldsymbol{\psi}_0)] = d_0(n_0, n_0\boldsymbol{y}_0) + b(\boldsymbol{\psi}_0)$
$+n_0{}^{-1}\{k + [\nabla d(n_0, n_0\boldsymbol{y}_0) - \boldsymbol{\psi}_0]^t n_0\boldsymbol{y}_0\}$.

Proof. From the definition of logarithmic divergence we see that

$$\delta(\boldsymbol{\psi}|\boldsymbol{\psi}_0) = b(\boldsymbol{\psi}_0) - b(\boldsymbol{\psi}) + (\boldsymbol{\psi} - \boldsymbol{\psi}_0)^t E_{\boldsymbol{y}|\boldsymbol{\psi}}[\boldsymbol{y}],$$

and (i) follows. Moreover,

$$E[\delta(\boldsymbol{\psi}|\boldsymbol{\psi}_0)] = b(\boldsymbol{\psi}_0) - E[b(\boldsymbol{\psi})] + E[\boldsymbol{\psi}^t \nabla b(\boldsymbol{\psi})] - \boldsymbol{\psi}_0^t E[\nabla b(\boldsymbol{\psi})].$$

Differentiation of the identity

$$\log \int \exp\{\boldsymbol{t}^t\boldsymbol{\psi} - sb(\boldsymbol{\psi})\} d\boldsymbol{\psi} = d(s, \boldsymbol{t}),$$

with respect to s, establishes straightforwardly that

$$E[b(\boldsymbol{\psi})] = -d_0(n_0, n_0\boldsymbol{y}_0).$$

Recalling that $E[\nabla b(\boldsymbol{\psi})] = \boldsymbol{y}_0$, we can write, for $i = 1, \ldots, k$,

$$\log \int b_i(\boldsymbol{\psi}) \exp\{\boldsymbol{t}^t\boldsymbol{\psi} - sb(\boldsymbol{\psi})\} d\boldsymbol{\psi} = \log t_i - \log c(s, s^{-1}\boldsymbol{t}) - \log s.$$

Differentiating this identity with respect to t_i, and interchanging the order of differentiation and integration, we see that

$$\int \psi_i b_i(\boldsymbol{\psi}) c(s, s^{-1}\boldsymbol{t}) \exp\{\boldsymbol{t}^t\boldsymbol{\psi} - sb(\boldsymbol{\psi})\} d\boldsymbol{\psi} = s^{-1}[1 + d_i(s, \boldsymbol{t})t_i],$$

for $i = 1, \ldots, k$, so that

$$E[\boldsymbol{\psi}^t \nabla b(\boldsymbol{\psi})] = n_0^{-1}[k + \nabla d(n_0, n_0\boldsymbol{y}_0)^t (n_0\boldsymbol{y}_0)] - n_0^{-1}\boldsymbol{\psi}_0^t(n_0, \boldsymbol{y}_0)$$

and (ii) follows. ◁

This result now enables us to establish easily the main result of interest.

Proposition 5.9. (***Conjugate posterior modes as Bayes estimates***).
With respect to the loss function $l(\boldsymbol{a}, \boldsymbol{\psi}) = \delta(\boldsymbol{\psi}|\boldsymbol{a})$, the Bayes estimate for $\boldsymbol{\psi}$, derived from independent observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ from the canonical k-parameter exponential family $p(\boldsymbol{y}|\boldsymbol{\psi})$ and corresponding conjugate prior $p(\boldsymbol{\psi}|n_0, \boldsymbol{y}_0)$, is the posterior mode, $\boldsymbol{\psi}^*$, which satisfies

$$\nabla b(\boldsymbol{\psi}^*) = (n_0 + n)^{-1}(n_0 \boldsymbol{y}_0 + n\bar{\boldsymbol{y}}_n),$$

with $\bar{\boldsymbol{y}}_n = n^{-1}(\boldsymbol{y}_1 + \cdots + \boldsymbol{y}_n)$.

Proof. We note first (see the proof of Proposition 5.6) that the logarithm of the posterior density is given by

$$\text{constant } + (n_0 \boldsymbol{y}_0 + n\bar{\boldsymbol{y}}_n)^t \boldsymbol{\psi} - (\boldsymbol{n}_0 + n) b(\boldsymbol{\psi}),$$

from which the claimed estimating equation for the posterior mode, $\boldsymbol{\psi}^*$, is immediately obtained. The result now follows by noting that the same equation arises in the minimisation of (ii) of Proposition 5.8, with $n_0 + n$ replacing n_0, and $n_0 \boldsymbol{y}_0 + n\bar{\boldsymbol{y}}_n$ replacing $n_0 \boldsymbol{y}_0$. $\triangleleft$

For a recent discussion of conjugate priors for exponential families, see Consonni and Veronese (1992b). In complex problems, conjugate priors may have strong, unsuspected implications; for an example, see Dawid (1988a).

5.2.3 Approximations with Conjugate Families

Our main motivation in considering conjugate priors for exponential families has been to provide tractable prior to posterior (or predictive) analysis. At the same time, we might hope that the conjugate family for a particular parametric model would contain a sufficiently rich range of prior density "shapes" to enable one to approximate reasonably closely any particular actual prior belief function of interest. The next example shows that might well not be the case. However, it also indicates how, with a suitable extension of the conjugate family idea, we can achieve both tractability and the ability to approximate closely any actual beliefs.

Example 5.7. ***(The spun coin)***. Diaconis and Ylvisaker (1979) highlight the fact that, whereas a tossed coin typically generates equal long-run frequencies of heads and tails, this is not at all the case if a coin is spun on its edge. Experience suggests that these long-run frequencies often turn out for some coins to be in the ratio 2:1 or 1:2, and for other coins even as extreme as 1:4. In addition, some coins do appear to behave symmetrically.

Let us consider the repeated spinning under perceived "identical conditions" of a given coin, about which we have no specific information beyond the general background set out

above. Under the circumstances specified, suppose we judge the sequence of outcomes to be exchangeable, so that a Bernoulli parametric model, together with a prior density for the long-run frequency of heads, completely specifies our belief model. How might we represent this prior density mathematically?

We are immediately struck by two things: first, in the light of the information given, any realistic prior shape will be at least bimodal, and possibly trimodal; secondly, the conjugate family for the Bernoulli parametric model is the beta family (see Example 5.4), which does not contain bimodal densities. It appears, therefore, that an insistence on tractability, in the sense of restricting ourselves to conjugate priors, would preclude an honest prior specification.

However, we can easily generate multimodal shapes by considering *mixtures* of beta densities,

$$p(\theta \mid \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{m} \pi_i \text{Be}(\theta \mid \alpha_i, \beta_i),$$

with mixing weights $\pi_i > 0$, $\pi_1 + \cdots + \pi_m = 1$, attached to a selection of conjugate densities, $\text{Be}(\theta \mid \alpha_i, \beta_i)$, $i = 1, \ldots, m$. Figure 5.2 displays the prior density resulting from the mixture

$$0.5\ \text{Be}(\theta \mid 10, 20) + 0.2\ \text{Be}(\theta \mid 15, 15) + 0.3\ \text{Be}(\theta \mid 20, 10),$$

which, among other things, reflects a judgement that about 20% of coins seem to behave symmetrically and most of the rest tend to lead to 2:1 or 1:2 ratios, with somewhat more of the latter than the former.

Suppose now that we observe n outcomes $\boldsymbol{x} = (x_1, \ldots, x_n)$ and that these result in $r_n = x_1 + \cdots + x_n$ heads, so that

$$p(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{r_n} (1-\theta)^{n-r_n}.$$

Considering the general mixture prior form

$$p(\theta \mid \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{m} \pi_i\ \text{Be}(\theta \mid \alpha_i, \beta_i),$$

we easily see from Bayes' theorem that

$$p(\theta \mid \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}) = p(\theta \mid \boldsymbol{\pi}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*),$$

where

$$\alpha_i^* = \alpha_i + r_n, \quad \beta_i^* = \beta_i + n - r_n$$

and

$$\begin{aligned} \pi_i^* &\propto \pi_i \int_0^1 \theta^{r_n} (1-\theta)^{n-r_n} \text{Be}(\theta \mid \alpha_i, \beta_i)\, d\theta \\ &\propto \pi_i \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \cdot \frac{\Gamma(\alpha_i^*)\Gamma(\beta_i^*)}{\Gamma(\alpha_i^* + \beta_i^*)}\ , \end{aligned}$$

so that the resulting posterior density,

$$p(\theta \mid \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}) = \sum_{i=1}^{m} \pi_i^* \ \mathrm{Be}(\theta \mid \alpha_i^*, \beta_i^*),$$

is itself a mixture of m beta components. This establishes that the general mixture class of beta densities is *closed under sampling* with respect to the Bernoulli model.

In the case considered above, suppose that the spun coin results in 3 heads after 10 spins and 14 heads after 50 spins. The suggested prior density corresponds to $m = 3$,

$$\boldsymbol{\pi} = (0.5, 0.2, 0.3), \quad \boldsymbol{\alpha} = (10, 15, 20), \quad \boldsymbol{\beta} = (20, 15, 10).$$

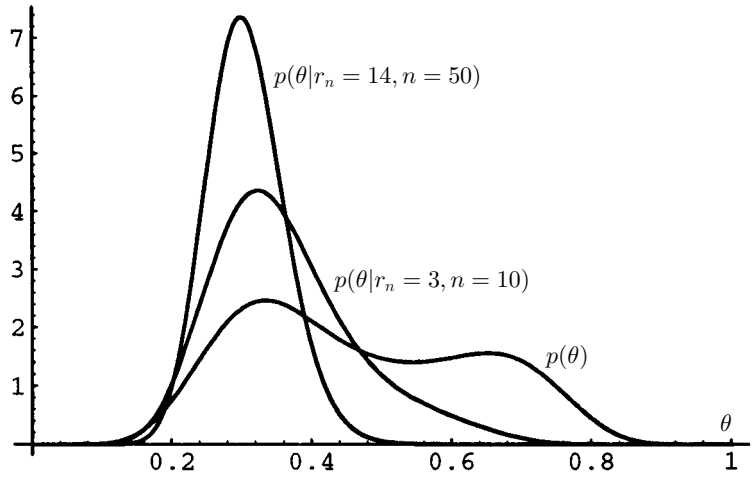

Figure 5.2 *Prior and posteriors from a three-component beta mixture prior density*

Detailed calculation yields:

$$\begin{aligned}
&\text{for } n = 10,\ r_n = 3;\ \boldsymbol{\pi}^* = (0.77, 0.16, 0.07),\\
&\qquad\qquad \boldsymbol{\alpha}^* = (13, 18, 23),\ \boldsymbol{\beta}^* = (27, 22, 17)\\
&\text{for } n = 50,\ r_n = 14;\ \boldsymbol{\pi}^* = (0.90, 0.09, 0.006),\\
&\qquad\qquad \boldsymbol{\alpha}^* = (24, 29, 34),\ \boldsymbol{\beta}^* = (56, 51, 46),
\end{aligned}$$

and the resulting posterior densities are shown in Figure 5.2.

This example demonstrates that, at least in the case of the Bernoulli parametric model and the beta conjugate family, the use of mixtures of conjugate densities both maintains the tractability of the analysis and provides a great deal of flexibility in approximating actual forms of prior belief. In fact, the same is true for any exponential family model and corresponding conjugate family, as we show in the following.

Proposition 5.10. **(*Mixtures of conjugate priors*)**. *Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a random sample from a regular exponential family distribution such that*

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \prod_{j=1}^{n} f(x_j)[g(\boldsymbol{\theta})]^n \exp\left\{\sum_{i=1}^{m} c_i \phi_i(\boldsymbol{\theta}) \left(\sum_{j=1}^{n} h_i(x_j)\right)\right\}$$

and let

$$p(\theta \mid \boldsymbol{\pi}, \boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_m) = \sum_{l=1}^{m} \pi_l p(\boldsymbol{\theta} \mid \boldsymbol{\tau}_l),$$

where, for $l = 1, \ldots, m$,

$$p(\theta \mid \boldsymbol{\tau}_l) = [K(\boldsymbol{\tau}_l)]^{-1}[g(\boldsymbol{\theta})]^{\tau_{l0}} \exp\left\{\sum_{i=1}^{k} c_i \phi_i(\boldsymbol{\theta}) \tau_{li}\right\}$$

are elements of the conjugate family. Then

$$p(\theta \mid \boldsymbol{\pi}, \boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_m, \boldsymbol{x}) = p(\theta \mid \boldsymbol{\pi}^*, \boldsymbol{\tau}_1^*, \ldots, \boldsymbol{\tau}_m^*) = \sum_{l=1}^{m} \pi_l^* p(\theta \mid \boldsymbol{\tau}_l^*),$$

where, with $\boldsymbol{t}_n(\boldsymbol{x}) = \left\{n, \sum_{j=1}^{n} h_1(x_j), \ldots, \sum_{j=1}^{n} h_k(j)\right\}$,

$$\boldsymbol{\tau}_l^* = \boldsymbol{\tau}_l + \boldsymbol{t}_n(\boldsymbol{x}),$$

and

$$\pi_l^* \propto \pi_l \prod_{j=1}^{n} f(x_j) \frac{K(\boldsymbol{\tau}_l^*)}{K(\boldsymbol{\tau}_l)} \,.$$

Proof. The results follows straightforwardly from Bayes' theorem and Proposition 5.5. ◁

It is interesting to ask just how flexible mixtures of conjugate prior are. The answer is that *any* prior density for an exponential family parameter can be approximated arbitrarily closely by such a mixture, as shown by Dalal and Hall (1983), and Diaconis and Ylvisaker (1985). However, their analyses do not provide a constructive mechanism for building up such a mixture. In practice, we are left with having to judge when a particular tractable choice, typically a conjugate form, a limiting conjugate form, or a mixture of conjugate forms, is "good enough, in the sense that probability statements based on the resulting posterior will not differ radically from the statements that would have resulted from using a more honest, but difficult to specify or intractable, prior.

The following result provides some guidance, in a much more general setting than that of conjugate mixtures, as to when an "approximate" (possibly improper) prior may be safely used in place of an "honest" prior.

Proposition 5.11. (***Prior approximation***). *Suppose that a belief model is defined by* $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ *and* $p(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$ *and that* $q(\boldsymbol{\theta})$ *is a non-negative function such that* $q(\boldsymbol{x}) = \int_\Theta p(\boldsymbol{x} \mid \boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta} < \infty$, *where, for some* $\Theta_0 \subseteq \Theta$ *and* $\alpha, \beta \in \Re^*$,

(a) $1 \leq p(\boldsymbol{\theta})/q(\boldsymbol{\theta}) \leq 1 + \alpha$, *for all* $\boldsymbol{\theta} \in \Theta_0$,

(b) $p(\boldsymbol{\theta})/q(\boldsymbol{\theta}) \leq \beta$, *for all* $\boldsymbol{\theta} \in \Theta$.

Let $p = \int_{\Theta_0} p(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{\theta}$, $q = \int_{\Theta_0} q(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{\theta}$, *and* $q(\boldsymbol{\theta} \mid \boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{\theta}) q(\boldsymbol{\theta}) / \int p(\boldsymbol{x} \mid \boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta}$. *Then,*

(i) $(1-p)/p \leq \beta(1-q)/q$

(ii) $q \leq p(\boldsymbol{x})/q(\boldsymbol{x}) \leq (1+\alpha)/p$

(iii) *for all* $\boldsymbol{\theta} \in \Theta$, $p(\boldsymbol{\theta} \mid \boldsymbol{x})/q(\boldsymbol{\theta} \mid \boldsymbol{x}) \leq [p(\boldsymbol{\theta})/q(\boldsymbol{\theta})]/q \leq \beta/q$

(iv) *for all* $\boldsymbol{\theta} \in \Theta_0$, $p/(1+\alpha) \leq p(\boldsymbol{\theta} \mid \boldsymbol{x})/q(\boldsymbol{\theta} \mid \boldsymbol{x}) \leq (1+\alpha)/q$

(v) *for* $\varepsilon = \max\{(1-p), (1-q)\}$ *and* $f : \Theta \to \Re$ *such that* $\mid f(\boldsymbol{\theta}) \mid \leq m$,

$$m^{-1} \left| \int_\Theta f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{\theta} - \int_\Theta f(\boldsymbol{\theta}) q(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{\theta} \right| \leq \alpha + 3\varepsilon$$

Proof. (Dickey, 1976). Part (i) clearly follows from

$$\frac{1-p}{p} = \frac{\int_{\Theta_0^c} p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Theta_0} p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \leq \beta \frac{1-q}{q} \,.$$

Clearly,

$$p(\boldsymbol{x}) \geq \int_{\Theta_0} p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \geq \int_{\Theta_0} p(\boldsymbol{x} \mid \boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta} = q \cdot q(\boldsymbol{x}),$$

$$q(\boldsymbol{x}) \geq \int_{\Theta_0} q(\boldsymbol{x} \mid \boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta} \geq \frac{1}{1+\alpha} \int_{\Theta_0} p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{p}{1+\alpha} p(\boldsymbol{x}),$$

which establishes (ii). Part (iii) follows from (b) and (ii), and part (iv) follows from (a) and (ii). Finally,

$$\begin{aligned}
m^{-1} \left| \int_{\Theta} f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{\theta} - \int_{\Theta} f(\boldsymbol{\theta}) q(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{\theta} \right| &\leq \int \left| p(\boldsymbol{\theta} \mid \boldsymbol{x}) - q(\boldsymbol{\theta} \mid \boldsymbol{x}) \right| d\boldsymbol{\theta} \\
&\leq \int_{\Theta_0} \left| p(\boldsymbol{\theta} \mid \boldsymbol{x}) - q(\boldsymbol{\theta} \mid \boldsymbol{x}) \right| d\boldsymbol{\theta} + \int_{\Theta_0^c} \left| p(\boldsymbol{\theta} \mid \boldsymbol{x}) - q(\boldsymbol{\theta} \mid \boldsymbol{x}) \right| d\boldsymbol{\theta} \\
&\leq \int_{\Theta_0} \left| q(\boldsymbol{\theta} \mid \boldsymbol{x}) \left(\frac{p(\boldsymbol{\theta} \mid \boldsymbol{x})}{q(\boldsymbol{\theta} \mid \boldsymbol{x})} - 1 \right) \right| d\boldsymbol{\theta} + \int_{\Theta_0^c} \left| p(\boldsymbol{\theta} \mid \boldsymbol{x}) \right| d\boldsymbol{\theta} + \int_{\Theta_0^c} \left| q(\boldsymbol{\theta} \mid \boldsymbol{x}) \right| d\boldsymbol{\theta} \\
&\leq \int_{\Theta_0} \left| q(\boldsymbol{\theta} \mid \boldsymbol{x}) \left(\frac{1+\alpha}{q} - 1 \right) \right| d\boldsymbol{\theta} + (1-p) + (1-q) \qquad \text{(by iv)} \\
&= (1+\alpha-q) + (1-p) + (1-q) \leq \alpha + 3\varepsilon,
\end{aligned}$$

which proves (v). ◁

If, in the above, Θ_0 is a subset of Θ with high probability under $q(\boldsymbol{\theta} \mid \boldsymbol{x})$ and α is chosen to be small and β not too large, so that $q(\boldsymbol{\theta})$ provides a good approximation to $p(\boldsymbol{\theta})$ within Θ_0 and $p(\boldsymbol{\theta})$ is nowhere much greater than $q(\boldsymbol{\theta})$, then (i) implies that Θ_0 has high probability under $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ and (ii), (iv) and (v) establish that both the respective predictive and posterior distributions, within Θ_0, and also the posterior expectations of bounded functions are very close. More specifically, if f is taken to be the indicator function of any subset $\Theta^* \subseteq \Theta$, (v) implies that

$$\left| \int_{\Theta^*} p(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{\theta} - \int_{\Theta^*} q(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{\theta} \right| \leq \alpha + 3\varepsilon,$$

providing a bound on the inaccuracy of the posterior probability statement made using $q(\boldsymbol{\theta} \mid \boldsymbol{x})$ rather than $p(\boldsymbol{\theta} \mid \boldsymbol{x})$.

Proposition 5.11 therefore asserts that if a mathematically convenient alternative, $q(\boldsymbol{\theta})$, to the would-be honest prior, $p(\boldsymbol{\theta})$, can be found, giving high posterior probability to a set $\Theta_0 \subseteq \Theta$ within which it provides a good approximation to $p(\boldsymbol{\theta})$ and such that it is nowhere orders of magnitude smaller than $p(\boldsymbol{\theta})$ outside Θ_0, then $q(\boldsymbol{\theta})$ may reasonably be used in place of $p(\boldsymbol{\theta})$.

In the case of $\Theta = \Re$, Figure 5.3 illustrates, in stylised form, a frequently occurring situation, where the choice $q(\boldsymbol{\theta}) = c$, for some constant c, provides

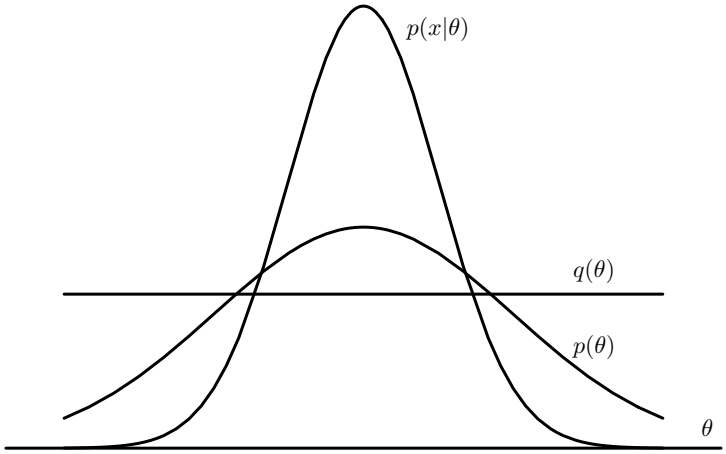

Figure 5.3 *Typical conditions for precise measurement*

a convenient approximation. In qualitative terms, the likelihood is highly peaked relative to $p(\boldsymbol{\theta})$, which has little curvature in the region of non-negligible likelihood.

In this situation of "precise measurement" (Savage, 1962), the choice of the function $q(\boldsymbol{\theta}) = c$, for an appropriate constant c, clearly satisfies the conditions of Proposition 5.10 and we obtain

$$p(\theta \mid \boldsymbol{x}) \simeq q(\theta \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \theta)c}{\int_{\Re} p(\boldsymbol{x} \mid \theta)c\, d\theta} = \frac{p(\boldsymbol{x} \mid \theta)}{\int_{\Re} p(\boldsymbol{x} \mid \theta)\, d\theta},$$

the normalised likelihood function.

The second of the implementation questions posed at the end of Section 5.1.6 concerned the possibility of avoiding the need for precise mathematical representation of the prior density in situations where the information provided by the data is far greater than that implicit in the prior. The above analysis goes some way to answering that question; the following section provides a more detailed analysis.

5.3 ASYMPTOTIC ANALYSIS

In Chapter 4, we saw that in representations of belief models for observables involving a parametric model $p(x \mid \boldsymbol{\theta})$ and a prior specification $p(\boldsymbol{\theta})$, the parameter $\boldsymbol{\theta}$ acquired an operational meaning as some form of strong law limit of observables. Given observations $\boldsymbol{x} = (x_1, \ldots, x_n)$, the posterior distribution, $p(\boldsymbol{\theta} \mid \boldsymbol{x})$, then describes beliefs about that strong law limit in the light of the information provided by $x_1, \ldots, x_n$. To answer the second question posed at the end of Section 5.1.6, we

now wish to examine various properties of $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ as the number of observations increases; i.e., as $n \to \infty$. Intuitively, we would hope that beliefs about $\boldsymbol{\theta}$ would become more and more concentrated around the "true" parameter value; i.e., the corresponding strong law limit. Under appropriate conditions, we shall see that this is, indeed, the case.

5.3.1 Discrete Asymptotics

We begin by considering the situation where $\Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \}$ consists of a countable (possibly finite) set of values, such that the parametric model corresponding to the true parameter, $\boldsymbol{\theta}_t$, is "distinguishable" from the others, in the sense that the logarithmic divergences, $\int p(x \mid \boldsymbol{\theta}_t) \log[p(x \mid \boldsymbol{\theta}_t)/p(x \mid \boldsymbol{\theta}_i)]\, dx$ are strictly larger than zero, for all $i \neq t$.

Proposition 5.12. (***Discrete asymptotics***). *Let $\boldsymbol{x} = (x_1, \dots, x_n)$ be observations for which a belief model is defined by the parametric model $p(x \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots\}$, and the prior $p(\boldsymbol{\theta}) = \{p_1, p_2, \dots\}$, $p_i > 0$, $\sum_i p_i = 1$. Suppose that $\boldsymbol{\theta}_t \in \Theta$ is the true value of $\boldsymbol{\theta}$ and that, for all $i \neq t$,*

$$\int p(x \mid \boldsymbol{\theta}_t) \log \left[\frac{p(x \mid \boldsymbol{\theta}_t)}{p(x \mid \boldsymbol{\theta}_i)} \right] dx > 0;$$

then

$$\lim_{n \to \infty} p(\boldsymbol{\theta}_t \mid \boldsymbol{x}) = 1, \quad \lim_{n \to \infty} p(\boldsymbol{\theta}_i \mid \boldsymbol{x}) = 0, \ i \neq t.$$

Proof. By Bayes' theorem, and assuming that $p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta})$,

$$\begin{aligned} p(\boldsymbol{\theta}_i \mid \boldsymbol{x}) &= p_i \, \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}_i)}{p(\boldsymbol{x})} \\ &= \frac{p_i \, \{p(\boldsymbol{x} \mid \boldsymbol{\theta}_i)/p(\boldsymbol{x} \mid \boldsymbol{\theta}_t)\}}{\sum_i p_i \, \{p(\boldsymbol{x} \mid \boldsymbol{\theta}_i)/p(\boldsymbol{x} \mid \boldsymbol{\theta}_t)\}} \\ &= \frac{\exp\{\log p_i + S_i\}}{\sum_i \exp\{\log p_i + S_i\}}, \end{aligned}$$

where

$$S_i = \sum_{j=1}^{n} \log \frac{p(x_j \mid \boldsymbol{\theta}_i)}{p(x_j \mid \boldsymbol{\theta}_t)} \, .$$

Conditional on $\boldsymbol{\theta}_t$, the latter is the sum of n independent identically distributed random quantities and hence, by the strong law of large numbers (see Section 3.2.3),

$$\lim_{n \to \infty} \frac{1}{n} S_i = \int p(x \mid \boldsymbol{\theta}_t) \log \left[\frac{p(x \mid \boldsymbol{\theta}_i)}{p(x \mid \boldsymbol{\theta}_t)} \right] dx.$$

The right-hand side is negative for all $i \neq t$, and equals zero for $i = t$, so that, as $n \to \infty$, $S_t \to 0$ and $S_i \to -\infty$ for $i \neq t$, which establishes the result. ◁

An alternative way of expressing the result of Proposition 5.12, established for countable Θ, is to say that the posterior distribution function for $\boldsymbol{\theta}$ ultimately degenerates to a step function with a single (unit) step at $\boldsymbol{\theta} = \boldsymbol{\theta}_t$. In fact, this result can be shown to hold, under suitable regularity conditions, for much more general forms of Θ. However, the proofs require considerable measure-theoretic machinery and the reader is referred to Berk (1966, 1970) for details.

A particularly interesting result is that if the true $\boldsymbol{\theta}$ is *not* in Θ, the posterior degenerates onto the value in Θ which gives the parametric model closest in logarithmic divergence to the true model.

5.3.2 Continuous Asymptotics

Let us now consider what can be said in the case of general Θ about the forms of probability statements implied by $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ for large n. Proceeding heuristically for the moment, without concern for precise regularity conditions, we note that, in the case of a parametric representation for an exchangeable sequence of observables,

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{x}) &\propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(x_i \mid \boldsymbol{\theta}) \\ &\propto \exp\{\log p(\boldsymbol{\theta}) + \log p(\boldsymbol{x} \mid \boldsymbol{\theta})\}. \end{aligned}$$

If we now expand the two logarithmic terms about their respective maxima, $\boldsymbol{m}_0$ and $\hat{\boldsymbol{\theta}}_n$, assumed to be determined by setting $\nabla \log p(\boldsymbol{\theta}) = 0$, $\nabla \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) = 0$, respectively, we obtain

$$\begin{aligned} \log p(\boldsymbol{\theta}) &= \log p(\boldsymbol{m}_0) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{m}_0)^t \boldsymbol{H}_0 (\boldsymbol{\theta} - \boldsymbol{m}_0) + R_0 \\ \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) &= \log p(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_n) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)^t \boldsymbol{H}(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) + R_n, \end{aligned}$$

where R_0, R_n denote remainder terms and

$$\boldsymbol{H}_0 = \left(-\frac{\partial^2 \log p(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right) \Bigg|_{\boldsymbol{\theta} = \boldsymbol{m}_0} \qquad \boldsymbol{H}(\hat{\boldsymbol{\theta}}_n) = \left(-\frac{\partial^2 \log p(\boldsymbol{x} \mid \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right) \Bigg|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_n}.$$

Assuming regularity conditions which ensure that R_0, R_n are small for large n, and ignoring constants of proportionality, we see that

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{x}) &\propto \exp\left\{ -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{m}_0)^t \boldsymbol{H}_0 (\boldsymbol{\theta} - \boldsymbol{m}_0) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)^t H(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) \right\} \\ &\propto \exp\left\{ -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{m}_n)^t \boldsymbol{H}_n (\boldsymbol{\theta} - \boldsymbol{m}_n) \right\}, \end{aligned}$$

with

$$\boldsymbol{H}_n = \boldsymbol{H}_0 + \boldsymbol{H}(\hat{\boldsymbol{\theta}}_n)$$
$$\boldsymbol{m}_n = \boldsymbol{H}_n^{-1}\left(\boldsymbol{H}_0\boldsymbol{m}_0 + \boldsymbol{H}(\hat{\boldsymbol{\theta}}_n)\hat{\boldsymbol{\theta}}_n\right),$$

where $\boldsymbol{m}_0$ (the *prior mode*) maximises $p(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}_n$ (the *maximum likelihood estimate*) maximises $p(\boldsymbol{x} \mid \boldsymbol{\theta})$. The Hessian matrix, $\boldsymbol{H}(\hat{\boldsymbol{\theta}}_n)$, measures the local curvature of the log-likelihood function at its maximum, $\hat{\boldsymbol{\theta}}_n$, and is often called the *observed information matrix*.

This heuristic development thus suggests that $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ will, for large n, tend to resemble a multivariate normal distribution, $\mathrm{N}_k(\boldsymbol{\theta} \mid \boldsymbol{m}_n, \boldsymbol{H}_n)$ (see Section 3.2.5) whose mean is a matrix weighted average of a prior (modal) estimate and an observation-based (maximum likelihood) estimate, and whose precision matrix is the sum of the prior precision matrix and the observed information matrix.

Other approximations suggest themselves: for example, for large n the prior precision will tend to be small compared with the precision provided by the data and could be ignored. Also, since, by the strong law of large numbers, for all i, j,

$$\lim_{n\to\infty}\left\{\frac{1}{n}\left(-\frac{\partial^2 \log p(\boldsymbol{x} \mid \boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)\right\} = \lim_{n\to\infty}\left\{\frac{1}{n}\sum_{l=1}^{n}\left(-\frac{\partial^2 \log p(x_l \mid \boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)\right\}$$
$$= \int p(x \mid \boldsymbol{\theta})\left(-\frac{\partial^2 \log p(x \mid \boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)dx$$

we see that $H(\hat{\boldsymbol{\theta}}_n) \to n\boldsymbol{I}(\hat{\boldsymbol{\theta}}_n)$, where $\boldsymbol{I}(\boldsymbol{\theta})$, defined by

$$(\boldsymbol{I}(\boldsymbol{\theta}))_{ij} = \int p(x \mid \boldsymbol{\theta})\left(-\frac{\partial^2 \log p(x \mid \boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)dx,$$

is the so-called *Fisher (or expected) information matrix*. We might approximate $p(\boldsymbol{\theta} \mid \boldsymbol{x})$, therefore, by either $\mathrm{N}_k(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}_n, H(\hat{\boldsymbol{\theta}}_n))$ or $\mathrm{N}_k(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}_n, nI(\hat{\boldsymbol{\theta}}_n))$, where k is the dimension of $\boldsymbol{\theta}$.

In the case of $\theta \in \Theta \subseteq \Re$,

$$H(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2}\log p(\boldsymbol{x} \mid \theta),$$

so that the approximate posterior variance is the negative reciprocal of the rate of change of the first derivative of $\log p(\boldsymbol{x} \mid \theta)$ in the neighbourhood of its maximum. Sharply peaked log-likelihoods imply small posterior uncertainty and vice-versa.

There is a large literature on the regularity conditions required to justify mathematically the heuristics presented above. Those who have contributed to the field include: Laplace (1812), Jeffreys (1939/1961, Chapter 4), LeCam (1953, 1956, 1958,

1966, 1970, 1986), Lindley (1961b), Freedman (1963b, 1965), Walker (1969), Chao (1970), Dawid (1970), DeGroot (1970, Chapter 10), Ibragimov and Hasminski (1973), Heyde and Johnstone (1979), Hartigan (1983, Chapter 4), Bermúdez (1985), Chen (1985), Sweeting and Adekola (1987), Fu and Kass (1988), Fraser and McDunnough (1989), Sweeting (1992) and Ghosh *et al.* (1994). Related work on higher-order expansion approximations in which the normal appears as a leading term includes that of Hartigan (1965), Johnson (1967, 1970), Johnson and Ladalla (1979) and Crowder (1988). The account given below is based on Chen (1985).

In what follows, we assume that $\boldsymbol{\theta} \in \Theta \subseteq \Re^k$ and that $\{p_n(\boldsymbol{\theta}), n = 1, 2, \ldots\}$ is a sequence of posterior densities for $\boldsymbol{\theta}$, typically of the form $p_n(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \,|\, x_1, \ldots, x_n)$, derived from an exchangeable sequence with parametric model $p(x \,|\, \boldsymbol{\theta})$ and prior $p(\boldsymbol{\theta})$, although the mathematical development to be given does not require this. We define $L_n(\boldsymbol{\theta}) = \log p_n(\boldsymbol{\theta})$, and assume throughout that, for every n, there is a strict local maximum, $\boldsymbol{m}_n$, of p_n (or, equivalently, L_n) satisfying:

$$\boldsymbol{L}'_n(\boldsymbol{m}_n) = \nabla L_n(\boldsymbol{\theta}) \,|_{\,\boldsymbol{\theta}=\boldsymbol{m}_n} = 0$$

and implying the existence and positive-definiteness of

$$\Sigma_n = \left(-\boldsymbol{L}''_n(\boldsymbol{m}_n)\right)^{-1},$$

where $[\boldsymbol{L}''_n(\boldsymbol{m}_n)]_{ij} = \left(\partial^2 L_n(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j\right) \,|_{\,\boldsymbol{\theta}=\boldsymbol{m}_n}$.

Defining $|\,\boldsymbol{\theta}\,| = (\boldsymbol{\theta}^t\boldsymbol{\theta})^{1/2}$ and $B_\delta(\boldsymbol{\theta}^*) = \{\boldsymbol{\theta} \in \Theta;\, |\,\boldsymbol{\theta} - \boldsymbol{\theta}^*\,| < \delta\}$, we shall show that the following three basic conditions are sufficient to ensure a valid normal approximation for $p_n(\boldsymbol{\theta})$ in a small neighbourhood of $\boldsymbol{m}_n$ as n becomes large.

(c1) "*Steepness*". $\overline{\sigma}_n^2 \to 0$ as $n \to \infty$, where $\overline{\sigma}_n^2$ is the largest eigenvalue of Σ_n.

(c2) "*Smoothness*". For any $\varepsilon > 0$, there exists N and $\delta > 0$ such that, for any $n > N$ and $\boldsymbol{\theta} \in B_\delta(\boldsymbol{m}_n)$, $\boldsymbol{L}''_n(\boldsymbol{\theta})$ exists and satisfies

$$\boldsymbol{I} - \boldsymbol{A}(\varepsilon) \leq \boldsymbol{L}''_n(\boldsymbol{\theta})\{\boldsymbol{L}''(\boldsymbol{m}_n)\}^{-1} \leq \boldsymbol{I} + \boldsymbol{A}(\varepsilon),$$

where $\boldsymbol{I}$ is the $k \times k$ identity matrix and $\boldsymbol{A}(\varepsilon)$ is a $k \times k$ symmetric positive-semidefinite matrix whose largest eigenvalue tends to zero as $\varepsilon \to 0$.

(c3) "*Concentration*". For any $\delta > 0$, $\int_{B_\delta(\boldsymbol{m}_n)} p_n(\boldsymbol{\theta}) d\boldsymbol{\theta} \to 1$ as $n \to \infty$.

Essentially, we shall see that (c1), (c2) together ensure that, for large n, inside a small neighbourhood of $\boldsymbol{m}_n$ the function p_n becomes highly peaked and behaves like the multivariate normal density kernel $\exp\{-\frac{1}{2}\,(\boldsymbol{\theta}-\boldsymbol{m}_n)^t\,\Sigma_n^{-1}\,(\boldsymbol{\theta}-\boldsymbol{m}_n)\}$. The final condition (c3) ensures that the probability outside any neighbourhood of $\boldsymbol{m}_n$ becomes negligible. We do not require any assumption that the $\boldsymbol{m}_n$ themselves converge, nor do we need to insist that $\boldsymbol{m}_n$ be a global maximum of p_n. We implicitly assume, however, that the limit of $p_n(\boldsymbol{m}_n)\,|\,\Sigma_n\,|^{\,1/2}$ exists as $n \to \infty$, and we shall now establish a bound for that limit.

Proposition 5.13. (***Bounded concentration***).
The conditions (c1), (c2) imply that

$$\lim_{n\to\infty} p_n(\boldsymbol{m}_n)\,|\Sigma_n|^{1/2} \le (2\pi)^{-k/2},$$

with equality if and only if (c3) holds.

Proof. Given $\varepsilon > 0$, consider $n > N$ and $\delta > 0$ as given in (c2). Then, for any $\boldsymbol{\theta} \in B_\delta(\boldsymbol{m}_n)$, a simple Taylor expansion establishes that

$$\begin{aligned} p_n(\boldsymbol{\theta}) &= p_n(\boldsymbol{m}_n)\exp\{L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{m}_n)\} \\ &= p_n(\boldsymbol{m}_n)\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{m}_n)^t(\boldsymbol{I}+\boldsymbol{R}_n)\Sigma_n^{-1}(\boldsymbol{\theta}-\boldsymbol{m}_n)\right\}, \end{aligned}$$

where

$$\boldsymbol{R}_n = \boldsymbol{L}_n''(\boldsymbol{\theta}^+)\{\boldsymbol{L}_n''(\boldsymbol{m}_n)\}^{-1}(\boldsymbol{m}_n) - \boldsymbol{I},$$

for some $\boldsymbol{\theta}^+$ lying between $\boldsymbol{\theta}$ and $\boldsymbol{m}_n$. It follows that

$$P_n(\delta) = \int_{B_\delta(\boldsymbol{m}_n)} p_n(\boldsymbol{\theta})d\boldsymbol{\theta}$$

is bounded above by

$$P_n^+(\delta) = p_n(\boldsymbol{m}_n)\,|\,\Sigma_n\,|^{1/2}\,|\,\boldsymbol{I}-\boldsymbol{A}(\varepsilon)\,|^{-1/2}\int_{|\,\boldsymbol{z}\,|<s_n}\exp\left\{-\tfrac{1}{2}\boldsymbol{z}^t\boldsymbol{z}\right\}d\boldsymbol{z}$$

and below by

$$P_n^-(\delta) = p_n(\boldsymbol{m}_n)\,|\,\Sigma_n\,|^{1/2}\,|\,\boldsymbol{I}+\boldsymbol{A}(\varepsilon)\,|^{-1/2}\int_{|\,\boldsymbol{z}\,|<t_n}\exp\left\{-\tfrac{1}{2}\boldsymbol{z}^t\boldsymbol{z}\right\}d\boldsymbol{z},$$

where $s_n = \delta(1-\underline{\alpha}(\varepsilon))^{1/2}/\underline{\sigma}_n$ and $t_n = \delta(1+\underline{\alpha}(\varepsilon))^{1/2}/\overline{\sigma}_n$, with $\overline{\sigma}_n^2(\underline{\sigma}_n^2)$ and $\overline{\alpha}(\varepsilon)(\underline{\alpha}(\varepsilon))$ the largest (smallest) eigenvalues of Σ_n and $\boldsymbol{A}(\varepsilon)$, respectively, since, for any $k\times k$ matrix $\boldsymbol{V}$,

$$B_{\delta/\overline{V}}(0) \subseteq \left\{\boldsymbol{z};(\boldsymbol{z}^t\boldsymbol{V}\boldsymbol{z})^{1/2}<\delta\right\} \subseteq B_{\delta/\underline{V}}(0),$$

where $\overline{V}^2(\underline{V}^2)$ are the largest (smallest) eigenvalues of $\boldsymbol{V}$.

Since (c1) implies that both s_n and t_n tend to infinity as $n\to\infty$, we have

$$\begin{aligned} |\boldsymbol{I}-\boldsymbol{A}(\varepsilon)|^{1/2}\lim_{n\to\infty}P_n(\delta) &\le \lim_{n\to\infty}p_n(\boldsymbol{m}_n)|\Sigma_n|^{1/2}(2\pi)^{k/2} \\ &\le |\boldsymbol{I}+\boldsymbol{A}(\varepsilon)|^{1/2}\lim_{n\to\infty}P_n(\delta), \end{aligned}$$

and the required inequality follows from the fact that $|\boldsymbol{I}\pm\boldsymbol{A}(\varepsilon)|\to 1$ as $\varepsilon\to 0$ and $P_n(\delta)\le 1$ for all n. Clearly, we have equality if and only if $\lim_{n\to\infty}P_n(\delta)=1$, which is condition (c3). ◁

We can now establish the main result, which may colloquially be stated as "$\boldsymbol{\theta}$ has an asymptotic posterior $\mathrm{N}_k(\boldsymbol{\theta}|\boldsymbol{m}_n, \Sigma_n^{-1})$ distribution, where $\boldsymbol{L}_n'(\boldsymbol{m}_n) = 0$ and $\Sigma_n^{-1} = -\boldsymbol{L}_n''(\boldsymbol{m}_n)$."

Proposition 5.14. (***Asymptotic posterior normality***). *For each n, consider $p_n(\cdot)$ as the density function of a random quantity $\boldsymbol{\theta}_n$, and define, using the notation above, $\boldsymbol{\phi}_n = \Sigma_n^{-1/2}(\boldsymbol{\theta}_n - \boldsymbol{m}_n)$. Then, given (c1) and (c2), (c3) is a necessary and sufficient condition for $\boldsymbol{\phi}_n$ to converge in distribution to $\boldsymbol{\phi}$, where $p(\boldsymbol{\phi}) = (2\pi)^{-k/2} \exp\left\{-\frac{1}{2}\boldsymbol{\phi}^t\boldsymbol{\phi}\right\}$.*

Proof. Given (c1) and (c2), and writing $\boldsymbol{b} \geq \boldsymbol{a}$, for $\boldsymbol{a}, \boldsymbol{b} \in \Re^k$, to denote that all components of $\boldsymbol{b} - \boldsymbol{a}$ are non-negative, it suffices to show that, as $n \to \infty$, $P_n(\boldsymbol{a} \leq \boldsymbol{\phi}_n \leq \boldsymbol{b}) \to P(\boldsymbol{a} \leq \boldsymbol{\phi} \leq \boldsymbol{b})$ if and only if (c3) holds.

We first note that

$$P_n(\boldsymbol{a} \leq \boldsymbol{\phi}_n \leq \boldsymbol{b}) = \int_{\Theta_n} p_n(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

where, by (c1), for any $\delta > 0$ and sufficiently large n,

$$\Theta_n = \left\{\boldsymbol{\theta}; \Sigma_n^{1/2}\boldsymbol{a} \leq (\boldsymbol{\theta} - \boldsymbol{m}_n) \leq \Sigma_n^{1/2}\boldsymbol{b}\right\} \subset B_\delta(\boldsymbol{m}_n).$$

It then follows, by a similar argument to that used in Proposition 5.13, that, for any $\varepsilon > 0$, $P_n(\boldsymbol{a} \leq \boldsymbol{\phi}_n \leq \boldsymbol{b})$ is bounded above by

$$P_n(\boldsymbol{m}_n) \left|\boldsymbol{I} - \boldsymbol{A}(\varepsilon)\right|^{-1/2} \left|\Sigma_n\right|^{1/2} \int_{Z(\varepsilon)} \exp\left\{-\tfrac{1}{2}\boldsymbol{z}^t\boldsymbol{z}\right\} d\boldsymbol{z},$$

where

$$Z(\varepsilon) = \left\{\boldsymbol{z}; \left[\boldsymbol{I} - \boldsymbol{A}(\varepsilon)\right]^{1/2} \boldsymbol{a} \leq \boldsymbol{z} \leq \left[\boldsymbol{I} - \boldsymbol{A}(\varepsilon)\right]^{1/2} \boldsymbol{b}\right\},$$

and is bounded below by a similar quantity with $+\boldsymbol{A}(\varepsilon)$ in place of $-\boldsymbol{A}(\varepsilon)$.

Given (c1), (c2), as $\varepsilon \to 0$ we have

$$\lim_{n\to\infty} P_n(\boldsymbol{a} \leq \boldsymbol{\phi}_n \leq \boldsymbol{b}) = \lim_{n\to\infty} p_n(\boldsymbol{m}_n) \left|\,\Sigma_n\,\right|^{1/2} \int_{Z(0)} \exp\left\{-\tfrac{1}{2}\boldsymbol{z}^t\boldsymbol{z}\right\} d\boldsymbol{z},$$

where $Z(0) = \{\boldsymbol{z}; \boldsymbol{a} \leq \boldsymbol{z} \leq \boldsymbol{b}\}$. The result follows from Proposition 5.13. ◁

Conditions (c1) and (c2) are often relatively easy to check in specific applications, but (c3) may not be so directly accessible. It is useful therefore to have available alternative conditions which, given (c1), (c2), imply (c3). Two such are provided by the following:

(c4) For any $\delta > 0$, there exists an integer N and $c,\, d \in \Re^+$ such that, for any $n > N$ and $\boldsymbol{\theta} \notin B_\delta(\boldsymbol{m}_n)$,

$$L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{m}_n) < -c\left\{(\boldsymbol{\theta} - \boldsymbol{m}_n)^t \Sigma_n^{-1} (\boldsymbol{\theta} - \boldsymbol{m}_n)\right\}^d.$$

(c5) As (c4), but, with $G(\boldsymbol{\theta}) = \log g(\boldsymbol{\theta})$ for some density (or normalisable positive function) $g(\boldsymbol{\theta})$ over Θ,

$$L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{m}_n) < -c\,|\Sigma_n|^{-d} + G(\boldsymbol{\theta}).$$

Proposition 5.15. (***Alternative conditions***). *Given (c1), (c2), either (c4) or (c5) implies (c3).*

Proof. It is straightforward to verify that

$$\int_{\Theta - B_\delta(\boldsymbol{m}_n)} p_n(\boldsymbol{\theta}) d\boldsymbol{\theta} \leq p_n(\boldsymbol{m}_n)\,|\Sigma_n|^{1/2} \int_{|\,z\,| > \delta/\overline{\sigma}_n} \exp\left\{-c(\boldsymbol{z}^t\boldsymbol{z})^d\right\} d\boldsymbol{z},$$

given (c4), and similarly, that

$$\int_{\Theta - B_\delta(\boldsymbol{m}_n)} p_n(\boldsymbol{\theta}) d\boldsymbol{\theta} \leq p_n(\boldsymbol{m}_n)\,|\Sigma_n|^{1/2}\,|\Sigma_n|^{-1/2} \exp\left\{-c\,|\Sigma_n|^{-d}\right\},$$

given (c4).

Since $p_n(\boldsymbol{m}_n)\,|\,\Sigma_n\,|^{1/2}$ is bounded (Proposition 5.11) and the remaining terms or the right-hand side clearly tend to zero, it follows that the left-hand side tends to zero as $n \to \infty$. $\triangleleft$

To understand better the relative ease of checking (c4) or (c5) in applications, we note that, if $p_n(\boldsymbol{\theta})$ is based on data $\boldsymbol{x}$,

$$L_n(\boldsymbol{\theta}) = \log p(\boldsymbol{\theta}) + \log p(\boldsymbol{x}\,|\,\boldsymbol{\theta}) - \log p(\boldsymbol{x}),$$

so that $L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{m}_n)$ does not involve the, often intractable, normalising constant $p(\boldsymbol{x})$. Moreover, (c4) does not even require the use of a proper prior for the vector $\boldsymbol{\theta}$.

We shall illustrate the use of (c4) for the general case of canonical conjugate analysis for exponential families.

Proposition 5.16. (***Asymptotic normality under conjugate analysis***). *Suppose that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are data resulting from a random sample of size n from the canonical exponential family form*

$$p(\boldsymbol{y} \,|\, \boldsymbol{\psi}) = a(\boldsymbol{y}) \exp\left\{\boldsymbol{y}^t \boldsymbol{\psi} - b(\boldsymbol{\psi})\right\}$$

with canonical conjugate prior density

$$p(\boldsymbol{\psi} \,|\, n_0, \boldsymbol{y}_0) = c(n_0, \boldsymbol{y}_0) \exp\left\{n_0 \boldsymbol{y}_0^t \boldsymbol{\psi} - n_0 b(\boldsymbol{\psi})\right\}.$$

For each n, consider the posterior density

$$p_n(\boldsymbol{\psi}) = p(\boldsymbol{\psi} \,|\, n_0 + n, n_0 \boldsymbol{y}_0 + n\overline{\boldsymbol{y}}_n),$$

with $\overline{\boldsymbol{y}}_n = \sum_{i=1}^n \boldsymbol{y}_i / n$, to be the density function for a random quantity $\boldsymbol{\psi}_n$, and define $\boldsymbol{\phi}_n = \Sigma_n^{-1/2}(\boldsymbol{\psi}_n - \boldsymbol{b}'(\boldsymbol{m}_n))$, where

$$\boldsymbol{b}'(\boldsymbol{m}_n) = \nabla b(\boldsymbol{\psi})\Big|_{\boldsymbol{\psi}=\boldsymbol{m}_n} = \frac{n_0 \boldsymbol{y}_0 + n\overline{\boldsymbol{y}}_n}{n_0 + n}$$

$$\left(\boldsymbol{b}''(\boldsymbol{m}_n)\right)_{ij} = \left(\frac{\partial^2 b(\boldsymbol{\psi})}{\partial \psi_i \partial \psi_j}\right)\Bigg|_{\boldsymbol{\psi}=\boldsymbol{m}_n} = (n_0 + n)\,(\Sigma_n)_{ij}^{-1}.$$

Then $\boldsymbol{\phi}_n$ converges in distribution to $\boldsymbol{\phi}$, where

$$p(\boldsymbol{\phi}) = (2\pi)^{-k/2} \exp\left\{-\tfrac{1}{2}\boldsymbol{\phi}^t \boldsymbol{\phi}\right\}.$$

Proof. Colloquially, we have to prove that $\boldsymbol{\psi}$ has an asymptotic posterior $\mathrm{N}_k(\boldsymbol{\psi} \,|\, \boldsymbol{b}'(\boldsymbol{m}_n), \Sigma_n^{-1})$ distribution, where $\boldsymbol{b}'(\boldsymbol{m}_n') = (n_0 + n)^{-1}(n_0 \boldsymbol{y}_0 + n\overline{\boldsymbol{y}}_n)$ and $\Sigma_n^{-1} = (n_0 + n)^{-1}\boldsymbol{b}''(\boldsymbol{m}_n)$. From a mathematical perspective,

$$p_n(\boldsymbol{\psi}) \propto \exp\left\{(n_0 + n)h(\boldsymbol{\psi})\right\},$$

where $h(\boldsymbol{\psi}) = [\boldsymbol{b}'(\boldsymbol{m}_n)]^t \boldsymbol{\psi} - b(\boldsymbol{\psi})$, with $b(\boldsymbol{\psi})$ a continuously differentiable and strictly convex function (see Section 5.2.2). It follows that, for each n, $p_n(\boldsymbol{\psi})$ is unimodal with a maximum at $\boldsymbol{\psi} = \boldsymbol{m}_n$ satisfying $\nabla h(\boldsymbol{m}_n) = 0$. By the strict concavity of $h(\cdot)$, for any $\delta > 0$ and $\theta \notin B_\delta(\boldsymbol{m}_n)$, we have, for some $\boldsymbol{\psi}^+$ between $\boldsymbol{\psi}$ and $\boldsymbol{m}_n$, with angle θ between $\boldsymbol{\psi} - \boldsymbol{m}_n$ and $\nabla h(\boldsymbol{\psi}^+)$,

$$\begin{aligned} h(\boldsymbol{\psi}) - h(\boldsymbol{m}_n) &= (\boldsymbol{\psi} - \boldsymbol{m}_n)^t \nabla h(\boldsymbol{\psi}^+) \\ &= |\,\boldsymbol{\psi} - \boldsymbol{m}_n\,|\;\;|\,\nabla h(\boldsymbol{\psi}^+)\,|\,\cos\theta \\ &< -c\,|\,\boldsymbol{\psi} - \boldsymbol{m}_n\,|, \end{aligned}$$

for $c = \inf\left\{ \mid \nabla h(\boldsymbol{\psi}^+) \mid ; \boldsymbol{\psi} \notin B_\delta(\boldsymbol{m}_n)\right\} > 0$. It follows that

$$\begin{aligned} L_n(\boldsymbol{\psi}) - L_n(\boldsymbol{m}_n) &< -(n_0+n) \mid \boldsymbol{\psi} - \boldsymbol{m}_n \mid \\ &< -c_1 \left\{(\boldsymbol{\psi} - \boldsymbol{m}_n)^t \Sigma_n^{-1} (\boldsymbol{\psi} - \boldsymbol{m}_n)\right\}^{1/2}, \end{aligned}$$

where $c_1 = c\lambda^{-1}$, with λ^2 the largest eigenvalue of $\boldsymbol{b}''(\boldsymbol{m}_n)$, and hence that (c4) is satisfied. Conditions (c1), (c2) follows straightforwardly from the fact that

$$\begin{aligned} (n_0+n)\Sigma_n^{-1} &= \boldsymbol{b}''(m_n), \\ \boldsymbol{L}_n''(\boldsymbol{\psi})\{\boldsymbol{L}_n''(\boldsymbol{m}_n)\}^{-1} &= \boldsymbol{b}''(\boldsymbol{\psi})\{\boldsymbol{b}''(\boldsymbol{m}_n)\}^{-1}, \end{aligned}$$

the latter not depending on $n_0 + n$, and so the result follows by Propositions 5.12 and 5.13. ◁

Example 5.4. ***(continued)***. Suppose that $\mathrm{Be}(\theta \mid \alpha_n, \beta_n)$, where $\alpha_n = \alpha + r_n$, and $\beta_n = \beta + n - r_n$, is the posterior derived from n Bernoulli trials with r_n successes and a $\mathrm{Be}(\theta \mid \alpha, \beta)$ prior. Proceeding directly,

$$\begin{aligned} L_n(\theta) = \log p_n(\theta) &= \log p(\boldsymbol{x} \mid \theta) + \log p(\boldsymbol{\theta}) - \log p(\boldsymbol{x}) \\ &= (\alpha_n - 1)\log\theta + (\beta_n - 1)\log(1-\theta) - \log p(\boldsymbol{x}) \end{aligned}$$

so that

$$L_n'(\theta) = \frac{(\alpha_n - 1)}{\theta} - \frac{(\beta_n - 1)}{1-\theta}$$

and

$$L_n''(\theta) = -\frac{(\alpha_n - 1)}{\theta^2} - \frac{(\beta_n - 1)}{(1-\theta)^2}\,.$$

It follows that

$$m_n = \frac{\alpha_n - 1}{(\alpha_n + \beta_n - 2)}\,, \quad (-L_n''(m_n))^{-1} = \frac{(\alpha_n - 1)(\beta_n - 1)}{(\alpha_n + \beta_n - 2)^3}\,.$$

Condition (c1) is clearly satisfied since $(-L_n''(m_n))^{-1} \to 0$ as $n \to \infty$; condition (c2) follows from the fact that $L_n''(\theta)$ is a continuous function of θ. Finally, (c4) may be verified with an argument similar to the one used in the proof of Proposition 5.16.

Taking $\alpha = \beta = 1$ for illustration, we see that

$$m_n = \frac{r_n}{n}\,, \quad (-L_n''(m_n))^{-1} = \frac{1}{n}\cdot\frac{r_n}{n}\left(1 - \frac{r_n}{n}\right)\,,$$

and hence that the asymptotic posterior for θ is

$$\mathrm{N}\left(\theta \,\middle|\, \frac{r_n}{n}, \left\{\frac{1}{n}\cdot\frac{r_n}{n}\left(1-\frac{r_n}{n}\right)\right\}^{-1}\right).$$

(As an aside, we note the interesting "duality" between this asymptotic form for θ given n, r_n, and the asymptotic distribution for r_n/n given θ, which, by the central limit theorem, has the form

$$\mathrm{N}\left(\frac{r_n}{n} \,\middle|\, \theta, \left\{\frac{1}{n}\theta(1-\theta)\right\}^{-1}\right).$$

Further reference to this kind of "duality" will be given in Appendix B.)

5.3.3 Asymptotics under Transformations

The result of Proposition 5.16 is given in terms of the canonical parametrisation of the exponential family underlying the conjugate analysis. This prompts the obvious question as to whether the asymptotic posterior normality "carries over, with appropriate transformations of the mean and covariance, to an arbitrary (one-to-one) reparametrisation of the model. More generally, we could ask the same question in relation to Proposition 5.14. A partial answer is provided by the following.

Proposition 5.17. (***Asymptotic normality under transformation***).
With the notation and background of Proposition 5.14, suppose that $\boldsymbol{\theta}$ has an asymptotic $\mathrm{N}_k(\boldsymbol{\theta}|\boldsymbol{m}_n, \Sigma_n^{-1})$ distribution, with the additional assumptions that, with respect to a parametric model $p(\boldsymbol{x}|\boldsymbol{\theta}_0)$, $\bar{\sigma}_n^2 \to 0$ and $\boldsymbol{m}_n \to \boldsymbol{\theta}_0$ in probability, and that given any $\delta > 0$, there is a constraint $c(\delta)$ such that $P(\bar{\sigma}_n^2 \underline{\sigma}_n^{-2} \leq c(\delta)) \geq 1 - \delta$ for all sufficiently large n, where $\bar{\sigma}_n^2$ ($\underline{\sigma}_n^2$) is the largest (smallest) eigenvalue of Σ_n^2. Then, if $\boldsymbol{\nu} = \boldsymbol{g}(\boldsymbol{\theta})$ is a transformation such that, at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$,

$$\boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

is non-singular with continuous entries, $\boldsymbol{\nu}$ has an asymptotic distribution

$$\mathrm{N}_k\left(\boldsymbol{\nu} \;\middle|\; \boldsymbol{g}(\boldsymbol{m}_n), [\boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{m}_n)\Sigma_n \boldsymbol{J}_{\boldsymbol{g}}^t(\boldsymbol{m}_n)]^{-1}\right).$$

Proof. This is a generalization and Bayesian reformulation of classical results presented in Serfling (1980, Section 3.3). For details, see Mendoza (1994). ◁

For any finite n, the adequancy of the normal approximation provided by Proposition 5.17 may be highly dependent on the particular transformation used. Anscombe (1964a, 1964b) analyses the choice of transformations which improve asymptotic normality. A related issue is that of selecting appropriate parametrisations for various numerical approximation methods (Hills and Smith, 1992, 1993).

The expression for the asymptotic posterior precision matrix (inverse covariance matrix) given in Proposition 5.17 is often rather cumbersome to work with. A simpler, alternative form is given by the following.

Corollary 1. (***Asymptotic precision after transformation***).
In Proposition 5.10, if $\boldsymbol{H}_n = \Sigma_n^{-1}$ denotes the asymptotic precision matrix for $\boldsymbol{\theta}$, then the asymptotic precision matrix for $\boldsymbol{\nu} = \boldsymbol{g}(\boldsymbol{\theta})$ has the form

$$\boldsymbol{J}_{\boldsymbol{g}^{-1}}^t\big(\boldsymbol{g}(\boldsymbol{m}_n)\big)\boldsymbol{H}_n \boldsymbol{J}_{\boldsymbol{g}^{-1}}\big(\boldsymbol{g}(\boldsymbol{m}_n)\big),$$

where

$$\boldsymbol{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{\nu}) = \frac{\partial \boldsymbol{g}^{-1}(\boldsymbol{\nu})}{\partial \boldsymbol{\nu}}$$

is the Jacobian of the inverse transformation.

Proof. This follows immediately by reversing of the roles of $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$. ◁

In many applications, we simply wish to consider one-to-one transformations of a single parameter. The next result provides a convenient summary of the required transformation result.

Corollary 2. (***Asymptotic normality after scalar transformation***).
Suppose that given the conditions of Propositions 5.14, 5.17 with scalar θ, the sequence m_n tends in probability to θ_0 under $p(\boldsymbol{x}|\theta_0)$, and that $L_n''(m_n) \to 0$ in probability as $n \to \infty$. Then, if $\nu = g(\theta)$ is such that $g'(\theta) = dg(\theta)/\, d\theta$ is continuous and non-zero at $\theta = \theta_0$, the asymptotic posterior distribution for ν is

$$\mathrm{N}(\nu|g(m_n), -L_n''(m_n)[g'(m_n)]^{-2}).$$

Proof. The conditions ensure, by Proposition 5.14, that θ has an asymptotic posterior distribution of the form $\mathrm{N}(\theta|m_n, -L_n''(m_n))$, so that the result follows from Proposition 5.17. ◁

Example 5.4. ***(continued)***. Suppose, again, that $\mathrm{Be}(\theta \mid \alpha_n, \beta_n)$, where $\alpha_n = \alpha + r_n$, and $\beta_n = \beta + n - r_n$, is the posterior distribution of the parameter of a Bernoulli distribution afte n trials, and suppose now that we are interested in the asymptotic posterior distribution of the variance stabilising transformation (recall Example 3.3)

$$\nu = g(\theta) = 2\sin^{-1}\sqrt{\theta}\,.$$

Straightforward application of Corollary 2 to Proposition 5.17, leads to the asymptotic distribution

$$\mathrm{N}(\nu|2\sin^{-1}(\sqrt{r_n/n}), n),$$

whose mean and variance can be compared with the forms given in Example 3.3.

It is clear from the presence of the term $[g'(m_n)]^{-2}$ in the form of the asymptotic precision given in Corollary 2 to Proposition 5.17 that things will go wrong if $g'(m_n) \to 0$ as $n \to \infty$. This is dealt with in the result presented by the requirement that $g'(\theta_0) \neq 0$, where $m_n \to \theta_0$ in probability. A concrete illustration of the problems that arise when such a condition is not met is given by the following.

Example 5.8. ***(Non-normal asymptotic posterior)***. Suppose that the asymptotic posterior for a parameter $\theta \in \Re$ is given by $\mathrm{N}(\theta|\bar{x}_n, n)$, $n\bar{x}_n = x_1 + \cdots + x_n$, perhaps derived from $\mathrm{N}(x_i|\theta, 1)$, $i = 1, \ldots, n$, with $\mathrm{N}(\theta|0, h)$, having $h \approx 0$. Now consider the transformation $\nu = g(\theta) = \theta^2$, and suppose that the actual value of θ generating the x_i through $\mathrm{N}(x_i|\theta, 1)$ is $\theta = 0$.

Intuitively, it is clear that ν cannot have an asymptotic normal distribution since the sequence $\bar{x}_n^2$ is converging in probability to 0 through *strictly positive* values. Technically, $g'(0) = 0$ and the condition of the corollary is not satisfied. In fact, it can be shown that the asymptotic posterior distribution of $n\nu$ is χ^2 in this case.

One attraction of the availability of the results given in Proposition 5.17 and Corollary 1 is that verification of the conditions for asymptotic posterior normality (as in, for example, Proposition 5.14) may be much more straightforward under one choice of parametrisation of the likelihood than under another. The result given enables us to identify the posterior normal form for any convenient choice of parameters, subsequently deriving the form for the parameters of interest by straightforward transformation. An indication of the usefulness of this result is given in the following example (and further applications can be found in Section 5.4).

Example 5.9. ***(Asymptotic posterior normality for a ratio)***. Suppose that we have a random sample $x_1, \ldots, x_n$ from the model $\{\prod_{i=1}^{n} \mathrm{N}(x_i|\theta_1, 1), \mathrm{N}(\theta_1|0, \lambda_1)\}$ and, independently, another random sample $y_1, \cdots, y_n$ from the model $\{\prod_{i=1}^{n} \mathrm{N}(y_i|\theta_2, 1), \mathrm{N}(\theta_2|0, \lambda_2)\}$, where $\lambda_1 \approx 0$, $\lambda_2 \approx 0$ and $\theta_2 \neq 0$. We are interested in the posterior distribution of $\phi_1 = \theta_1/\theta_2$ as $n \to \infty$.

First, we note that, for large n, it is very easily verified that the joint posterior distribution for $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is given by

$$\mathrm{N}_2\left\{ \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \middle| \begin{pmatrix} \bar{x}_n \\ \bar{y}_n \end{pmatrix}, \begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix} \right\},$$

where $n\bar{x}_n = x_1 + \cdots + x_n$, $n\bar{y}_n = y_1 + \cdots + y_n$. Secondly, we note that the marginal asymptotic posterior for ϕ_1 can be obtained by defining an appropriate ϕ_2 such that $(\theta_1, \theta_2) \to (\phi_1, \phi_2)$ is a one-to-one transformation, obtaining the distribution of $\boldsymbol{\phi} = (\phi_1, \phi_2)$ using Proposition 5.17, and subsequently marginalising to ϕ_1.

An obvious choice for ϕ_2 is $\phi_2 = \theta_2$, so that, in the notation of Proposition 5.17, $\boldsymbol{g}(\theta_1, \theta_2) = (\phi_1, \phi_2)$ and

$$\boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{\theta}) = \begin{pmatrix} \partial\phi_1/\partial\theta_1 & \partial\phi_1/\partial\theta_2 \\ \partial\phi_2/\partial\theta_1 & \partial\phi_2/\partial\theta_2 \end{pmatrix} = \begin{pmatrix} \theta_2^{-1} & -\theta_1\theta_2^{-2} \\ 0 & 1 \end{pmatrix}.$$

The determinant of this, θ_2^{-1}, is non-zero for $\theta_2 \neq 0$, and the conditions of Proposition 5.17 are clearly satisfied. It follows that the asymptotic posterior of $\boldsymbol{\phi}$ is

$$\mathrm{N}_2\left(\begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} \middle| \begin{pmatrix} \bar{x}_n/\bar{y}_n \\ \bar{y}_n \end{pmatrix}, \quad n\bar{y}_n^2 \begin{pmatrix} 1 + (\bar{x}_n/\bar{y}_n)^2 & -\bar{x}_n \\ -\bar{x}_n & \bar{y}_n^2 \end{pmatrix}^{-1} \right),$$

so that the required asymptotic posterior for $\phi_1 = \theta_1/\theta_2$ is

$$\mathrm{N}\left(\phi_1 \middle| \frac{\bar{x}_n}{\bar{y}_n}, \quad n\bar{y}_n^2 \left(\frac{\bar{y}_n^2}{\bar{x}_n^2 + \bar{y}_n^2} \right) \right).$$

Any reader remaining unappreciative of the simplicity of the above analysis may care to examine the form of the likelihood function, etc., corresponding to an initial parametrisation directly in terms of ϕ_1, ϕ_2, and to contemplate verifying directly the conditions of Proposition 5.14 using the ϕ_1, ϕ_2 parametrisation.

5.4 REFERENCE ANALYSIS

In the previous section, we have examined situations where data corresponding to large sample sizes come to dominate prior information, leading to inferences which are negligibly dependent on the initial state of information. The third of the questions posed at the end of Section 5.1.6 relates to specifying prior distributions in situations where it is felt that, *even for moderate sample sizes*, the data should be expected to dominate prior information because of the "vague" nature of the latter.

However, the problem of characterising a "*non-informative*" or "*objective*" prior distribution, representing "*prior ignorance*", "*vague prior knowledge*" and "*letting the data speak for themselves*" is far more complex than the apparent intuitive immediacy of these words and phrases would suggest.

In Section 5.6.2, we shall provide a brief review of the fascinating history of the quest for this "baseline", limiting prior form. However, it is as well to make clear straightaway our own view—very much in the operationalist spirit with which we began our discussion of uncertainty in Chapter 2—that "mere words" are an inadequate basis for clarifying such a slippery concept. Put bluntly: data cannot ever speak entirely for themselves; every prior specification has *some* informative posterior or predictive implications; and "vague" is itself much too vague an idea to be useful. There is no "objective" prior that represents ignorance.

On the other hand, we recognise that there *is* often a pragmatically important need for a form of prior to posterior analysis capturing, *in some well-defined sense*, the notion of the prior having a minimal effect, relative to the data, on the final inference. Such a *reference analysis* might be required as an approximation to actual individual beliefs; more typically, it might be required as a limiting "what if?" baseline in considering a range of prior to posterior analyses, or as a *default* option when there are insufficient resources for detailed elicitation of actual prior knowledge.

In line with the unified perspective we have tried to adopt throughout this volume, the setting for our development of such a reference analysis will be the general decision-theoretic framework, together with the specific information-theoretic tools that have emerged in earlier chapters as key measures of the discrepancies (or "distances") between belief distributions. From the approach we adopt, it will be clear that the *reference prior* component of the analysis is simply a mathematical tool. It has considerable pragmatic importance in implementing a *reference analysis*, whose role and character will be precisely defined, but it is not a privileged, "uniquely non-informative" or "objective" prior. Its main use will be to provide a "conventional" prior, to be used when a default specification having a claim to being *non-influential* in the sense described above is required. We seek to move away, therefore, from the rather philosophically muddled debates about "prior ignorance" that have all too often confused these issues, and towards well-defined decision-theoretic and information-theoretic procedures.

5.4.1 Reference Decisions

Consider a specific form of decision problem with possible decisions $d \in \mathcal{D}$ providing possible answers, $a \in \mathcal{A}$, to an inference problem, with unknown state of the world $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$, utilities for consequences $(a, \boldsymbol{\omega})$ given by $u\big(d(\boldsymbol{\omega}_1)\big) = u(a, \boldsymbol{\omega}_1)$ and the availability of an experiment e which consists of obtaining an observation $\boldsymbol{x}$ having parametric model $p(\boldsymbol{x} \,|\, \boldsymbol{\omega}_2)$ and a prior probability density $p(\boldsymbol{\omega}) = p(\boldsymbol{\omega}_1 \,|\, \boldsymbol{\omega}_2)p(\boldsymbol{\omega}_2)$ for the unknown state of the world, $\boldsymbol{\omega}$. This general structure describes a situation where practical consequences depend directly on the $\boldsymbol{\omega}_1$ component of $\boldsymbol{\omega}$, whereas inference from data $\boldsymbol{x} \in X$ provided by experiment e takes place indirectly, through the $\boldsymbol{\omega}_2$ component of $\boldsymbol{\omega}$ as described by $p(\boldsymbol{\omega}_1 \,|\, \boldsymbol{\omega}_2)$. If $\boldsymbol{\omega}_1$ is a function of $\boldsymbol{\omega}_2$, the prior density is, of course, simply $p(\boldsymbol{\omega}_2)$.

To avoid subscript proliferation, let us now, without any risk of confusion, indulge in a harmless abuse of notation by writing $\boldsymbol{\omega}_1 = \boldsymbol{\omega}, \boldsymbol{\omega}_2 = \boldsymbol{\theta}$. This both simplifies the exposition and has the mnemonic value of suggesting that $\boldsymbol{\omega}$ is the state of the world of ultimate interest (since it occurs in the utility function), whereas $\boldsymbol{\theta}$ is a parameter in the usual sense (since it occurs in the probability model). Often $\boldsymbol{\omega}$ is just some function $\boldsymbol{\omega} = \boldsymbol{\phi}(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$; if $\boldsymbol{\omega}$ is not a function of $\boldsymbol{\theta}$, the relationship between $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$ is that described in their joint distribution $p(\boldsymbol{\omega}, \boldsymbol{\theta}) = p(\boldsymbol{\omega} \,|\, \boldsymbol{\theta})p(\boldsymbol{\theta})$.

Now, for given conditional prior $p(\boldsymbol{\omega} \,|\, \boldsymbol{\theta})$ and utility function $u(a, \boldsymbol{\omega})$, let us examine, *in utility terms*, the influence of the prior $p(\boldsymbol{\theta})$, relative to the observational information provided by e. We note that if a_0^* denotes the optimal answer under $p(\boldsymbol{\omega})$ and $a_{\boldsymbol{x}}^*$ denotes the optimal answer under $p(\boldsymbol{\omega} \,|\, \boldsymbol{x})$, then, using Definition 3.13 (ii), with appropriate notational changes, and noting that

$$\int p(\boldsymbol{x}) \int p(\boldsymbol{\omega} \,|\, \boldsymbol{x}) u(a_0^*, \boldsymbol{\omega})\, d\boldsymbol{\omega} d\boldsymbol{x} = \int p(\boldsymbol{\omega}) u(a_0^*, \boldsymbol{\omega})\, d\boldsymbol{\omega},$$

the expected (utility) value of the experiment e, given the prior $p(\boldsymbol{\theta})$, is

$$v_u\{e, p(\boldsymbol{\theta})\} = \int p(\boldsymbol{x}) \int p(\boldsymbol{\omega} \,|\, \boldsymbol{x}) u(a_{\boldsymbol{x}}^*, \boldsymbol{\omega})\, d\boldsymbol{\omega} d\boldsymbol{x} - \int p(\boldsymbol{\omega}) u(a_0^*, \boldsymbol{\omega})\, d\boldsymbol{\omega},$$

where, assuming $\boldsymbol{\omega}$ is independent of $\boldsymbol{x}$, given $\boldsymbol{\theta}$,

$$p(\boldsymbol{\omega}) = \int p(\boldsymbol{\omega} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \qquad p(\boldsymbol{\omega} \,|\, \boldsymbol{x}) = \int \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\omega} \,|\, \boldsymbol{\theta})}{p(\boldsymbol{x})} p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

and

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

If $e(k)$ denotes the experiment consisting of k independent replications of e, that is yielding observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ with joint parametric model $\prod_{i=1}^k p(\boldsymbol{x}_i \,|\, \boldsymbol{\theta})$, then $v_u\{e(k), p(\boldsymbol{\theta})\}$, the expected utility value of the experiment $e(k)$, has the same mathematical form as $v_u\{e, p(\boldsymbol{\theta})\}$, but with $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ and $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^k p(\boldsymbol{x}_i \,|\, \boldsymbol{\theta})$. Intuitively, at least in suitably regular cases, as $k \to \infty$ we obtain,

from $e(\infty)$, perfect (i.e., complete) information about $\boldsymbol{\theta}$, so that, assuming the limit to exist,

$$v_u\{e(\infty), p(\boldsymbol{\theta})\} = \lim_{k\to\infty} v_u\{e(k), p(\boldsymbol{\theta})\}$$

is the expected (utility) *value of perfect information*, about $\boldsymbol{\theta}$, given $p(\boldsymbol{\theta})$.

Clearly, the more valuable the information contained in $p(\boldsymbol{\theta})$, the less will be the expected value of perfect information about $\boldsymbol{\theta}$; conversely, the less valuable the information contained in the prior, the more we would expect to gain from exhaustive experimentation. This, then, suggests a well-defined "thought experiment" procedure for characterising a "minimally valuable prior": choose, from the class of priors which has been identified as compatible with other assumptions about $(\boldsymbol{\omega}, \boldsymbol{\theta})$, that prior, $\pi(\boldsymbol{\theta})$, say, which *maximises the expected value of perfect information about* $\boldsymbol{\theta}$. Such a prior will be called a u*-reference prior*; the posterior distributions,

$$\pi(\boldsymbol{\omega} \mid \boldsymbol{x}) = \int p(\boldsymbol{\omega} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{x})d\boldsymbol{\theta}$$
$$\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

derived from combining $\pi(\boldsymbol{\theta})$ with actual data $\boldsymbol{x}$, will be called *u-reference posteriors*; and the optimal decision derived from $\pi(\boldsymbol{\omega} \mid \boldsymbol{x})$ and $u(a, \boldsymbol{\omega})$ will be called a u*-reference decision*.

It is important to note that the limit above is *not* taken in order to obtain some form of asymptotic "approximation" to reference distributions; the "exact" reference prior is *defined* as that which maximises the value of *perfect* information about θ, *not* as that which maximises the expected value of the experiment.

Example 5.10. ***(Prediction with quadratic loss)***. Suppose that beliefs about a sequence of observables, $\boldsymbol{x} = (x_1, \ldots, x_n)$, correspond to assuming the latter to be a random sample from an $N(x \mid \mu, \lambda)$ parametric model, with known precision λ, together with a prior for μ to be selected from the class $\{N(\mu \mid \mu_0, \lambda_0), \mu_0 \in \Re, \lambda_0 \geq 0\}$. Assuming a quadratic loss function, the decision problem is to provide a point estimate for x_{n+1}, given $x_1, \ldots, x_n$. We shall derive a reference analysis of this problem, for which $\mathcal{A} = \Re$, $\omega = x_{n+1}$, and $\theta = \mu$. Moreover,

$$u(a, \omega) = -(a - x_{n+1})^2, \quad p(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} N(x_i \mid \mu, \lambda)$$

and, for given μ_0, λ_0, we have

$$p(\omega, \theta) = p(x_{n+1}, \mu) = p(x_{n+1} \mid \mu)p(\mu) = N(x_{n+1} \mid \mu, \lambda)N(\mu \mid \mu_0, \lambda_0).$$

For the purposes of the "thought experiment", let $\boldsymbol{z}_k = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ denote the (imagined) outcomes of k replications of the experiment yielding the observables $(x_1, \ldots, x_{kn})$, say,

and let us denote the future observation to be predicted (x_{kn+1}) simply by x. Then

$$v_u\{e(k), N(\mu \mid \mu_0, \lambda_0)\} = -\int p(\boldsymbol{z}_k) \inf_a \int p(x \mid \boldsymbol{z}_k)(a-x)^2 dx d\boldsymbol{z}_k + \inf_a \int p(x)(a-x)^2 dx.$$

However, we know from Proposition 5.3 that optimal estimates with respect to quadratic loss functions are given by the appropriate means, so that

$$\begin{aligned} v_u\{e(k), N(\mu \mid \mu_0, \lambda_0)\} &= -\int p(\boldsymbol{z}_k) V[x \mid \boldsymbol{z}_k] d\boldsymbol{z}_k + V[x] \\ &= -V[x \mid \boldsymbol{z}_k] + V[x], \end{aligned}$$

since, by virtue of the normal distributional assumptions, the predictive variance of x given $\boldsymbol{z}_k$ does not depend explicitly on $\boldsymbol{z}_k$. In fact, straightforward manipulations reveal that

$$\begin{aligned} v_u\{e(\infty), N(\mu \mid \mu_0, \lambda_0)\} &= \lim_{k\to\infty} v_u\{e(k), N(\mu \mid \mu_0, \lambda_0)\} \\ &= \lim_{k\to\infty} \left\{ -\left[\lambda^{-1} + (\lambda_0 + kn\lambda)^{-1}\right] + (\lambda^{-1} + \lambda_0^{-1}) \right\} = \lambda_0^{-1}, \end{aligned}$$

so that the *u-reference prior* corresponds to the choice $\lambda_0 = 0$, with μ_0 arbitrary.

Example 5.11. ***(Variance estimation)***. Suppose that beliefs about $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ correspond to assuming $\boldsymbol{x}$ to be a random sample from $N(x \mid 0, \lambda)$ together with a gamma prior for λ centred on λ_0, so that $p(\lambda) = \text{Ga}(\lambda \mid \alpha, \alpha\lambda_0^{-1})$, $\alpha > 0$. The decision problem is to provide a point estimate for $\sigma^2 = \lambda^{-1}$, assuming a standardised quadratic loss function, so that

$$u(a, \sigma^2) = -\left[\frac{(a-\sigma^2)}{\sigma^2}\right]^2 = -(a\lambda - 1)^2.$$

Thus, we have $\mathcal{A} = \Re^+$, $\theta = \lambda$, $w = \sigma^2$, and

$$p(\boldsymbol{x}, \lambda) = \prod_{i=1}^{n} N(x_i \mid 0, \lambda)\, \text{Ga}(\lambda \mid \alpha, \alpha\lambda_0^{-1}).$$

Let $\boldsymbol{z}_k = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ denote the outcome of k replications of the experiment. Then

$$v_u\{e(k), p(\lambda)\} = -\int p(\boldsymbol{z}_k) \inf_a \int p(\lambda \mid \boldsymbol{z}_k)\, (a\lambda - 1)^2\, d\lambda\, dz_k + \inf_a \int p(\lambda)\, (a\lambda - 1)^2\, d\lambda,$$

where

$$p(\lambda) = \text{Ga}(\lambda \mid \alpha, \alpha\lambda_0^{-1}), \quad p(\lambda \mid z_k) = \text{Ga}\left(\lambda \mid \alpha + \frac{kn}{2},\, \alpha\lambda_0^{-1} + \frac{kns^2}{2}\right),$$

and $kns^2 = \sum_i \sum_j x_{ij}^2$. Since

$$\inf_a \int \text{Ga}(\lambda \,|\, \alpha, \beta)\,(a\lambda - 1)^2\, d\lambda = \frac{1}{\alpha + 1}\,,$$

and this is attained when $a = \beta/(\alpha + 1)$, one has

$$\begin{aligned} v_u\{e(\infty), p(\lambda)\} &= \lim_{k\to\infty} v_u\{e(k), p(\lambda)\} \\ &= \lim_{k\to\infty} \left\{ -\frac{1}{1 + \alpha + (kn)/2} + \frac{1}{1+\alpha} \right\} = \frac{1}{1+\alpha}\,. \end{aligned}$$

This is maximised when $\alpha = 0$ and, hence, the *u-reference prior* corresponds to the choice $\alpha = 0$, with λ_0 arbitrary. Given *actual* data, $\boldsymbol{x} = (x_1, \ldots, x_n)$, the *$u$-reference posterior* for λ is $\text{Ga}(\lambda \,|\, n/2, ns^2/2)$, where $ns^2 = \sum_i x_i^2$ and, thus, the *u-reference decision* is to give the estimate

$$\hat{\sigma^2} = \frac{ns^2/2}{(n/2) + 1} = \frac{\Sigma x_i^2}{n+2}\,.$$

Hence, the reference estimator of σ^2 with respect to *standardised* quadratic loss is *not* the usual s^2, but a slightly smaller multiple of s^2.

It is of interest to note that, from a frequentist perspective, $\hat{\sigma}^2$ is the best invariant estimator of σ^2 and is admissible. Indeed, $\hat{\sigma}^2$ dominates s^2 or any smaller multiple of s^2 in terms of frequentist risk (cf. Example 45 in Berger, 1985a, Chapter 4). Thus, the u-reference approach has led to the "correct" multiple of s^2 as seen from a frequentist perspective.

Explicit reference decision analysis is possible when the parameter space $\Theta = \{\theta_1, \ldots, \theta_M\}$ is finite. In this case, the expected value of perfect information (cf. Definition 2.19) may be written as

$$v_u\{e(\infty), p(\theta)\} = \sum_{i=1}^{M} p(\theta_i) \sup_{\mathcal{D}} u(d(\theta_i)) - \sup_{\mathcal{D}} \sum_{i=1}^{M} p(\theta_i)\, u(d(\theta_i)),$$

and the u-reference prior, which is that $\pi(\theta)$ which maximises $v_u\{e(\infty), p(\theta)\}$, may be explicitly obtained by standard algebraic manipulations. For further information, see Bernardo (1981a) and Rabena (1998).

5.4.2 One-dimensional Reference Distributions

In Sections 2.7 and 3.4, we noted that reporting beliefs is itself a decision problem, where the "inference answer" space consists of the class of possible belief distributions that could be reported about the quantity of interest, and the utility function is a proper scoring rule which—in pure inference problems—may be identified with the logarithmic scoring rule.

Our development of reference analysis from now on will concentrate on this case, for which we simply denote $v_u\{\cdot\}$ by $v\{\cdot\}$, and replace the term "u-reference" by "reference".

In discussing reference decisions, we have considered a rather general utility structure where practical interest centred on a quantity $\boldsymbol{\omega}$ related to the $\boldsymbol{\theta}$ of an experiment by a conditional probability specification, $p(\boldsymbol{\omega} \mid \boldsymbol{\theta})$. Here, we shall consider the case where the quantity of interest is θ itself, with $\theta \in \Theta \subset \Re$. More general cases will be considered later.

If an experiment e consists of an observation $\boldsymbol{x} \in X$ having parametric model $p(\boldsymbol{x} \mid \theta)$, with $\boldsymbol{\omega} = \theta, \mathcal{A} = \{q(\cdot); q(\theta) > 0, \int_\Theta q(\theta) d\theta = 1\}$ and the utility function is the logarithmic scoring rule

$$u\{q(\cdot), \theta\} = A \log q(\theta) + B(\theta),$$

the expected utility value of the experiment e, given the prior density $p(\theta)$, is

$$v\{e, p(\theta)\} = \int p(\boldsymbol{x}) \int u\{q_x(\cdot), \theta\} p(\theta \mid \boldsymbol{x})\, d\theta d\boldsymbol{x} - \int u\{q_0(\cdot), \theta\}\, p(\theta)\, d\theta,$$

where $q_0(\cdot), q_x(\cdot)$ denote the optimal choices of $q(\cdot)$ with respect to $p(\theta)$ and $p(\theta \mid \boldsymbol{x})$, respectively. Noting that u is a proper scoring rule, so that, for any $p(\theta)$,

$$\sup_q \int u\{q(\cdot), \theta\} p(\theta)\, d\theta = \int u\{p(\cdot), \theta\} p(\theta)\, d\theta,$$

it is easily seen that

$$v\{e, p(\theta)\} \propto \int p(\boldsymbol{x}) \int p(\theta \mid \boldsymbol{x}) \log \frac{p(\theta \mid \boldsymbol{x})}{p(\theta)} d\theta d\boldsymbol{x} = I\{e, p(\theta)\}$$

the *amount of information* about θ which e may be expected to provide.

The corresponding expected information from the (hypothetical) experiment $e(k)$ yielding the (imagined) observation $\boldsymbol{z}_k = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ with parametric model

$$p(\boldsymbol{z}_k \mid \theta) = \prod_{i=1}^{k} p(\boldsymbol{x}_i \mid \theta)$$

is given by

$$I\{e(k), p(\theta)\} = \int p(\boldsymbol{z}_k) \int p(\theta \mid \boldsymbol{z}_k) \log \frac{p(\theta \mid \boldsymbol{z}_k)}{p(\theta)}\, d\theta d\boldsymbol{z}_k,$$

and so the expected (utility) value of perfect information about θ is

$$I\{e(\infty), p(\theta)\} = \lim_{k \to \infty} I\{e(k), p(\theta)\},$$

provided that this limit exists. This quantity measures the *missing information* about θ as a function of the prior $p(\theta)$.

The *reference prior* for θ, denoted by $\pi(\theta)$, is thus defined to be that prior which maximises the missing information functional. Given actual data $\boldsymbol{x}$, the *reference posterior* $\pi(\theta \mid \boldsymbol{x})$ to be reported is simply derived from Bayes' theorem, as $\pi(\theta \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \theta)\pi(\theta)$.

Unfortunately, $\lim_{k\to\infty} I\{e(k), p(\theta)\}$ is typically infinite (unless θ can only take a finite range of values) and a direct approach to deriving $\pi(\theta)$ along these lines cannot be implemented. However, a natural way of overcoming this technical difficulty is available: we derive the sequence of priors $\pi_k(\theta)$ which maximise $I\{e(k), p(\theta)\}, k = 1, 2, \ldots$, and subsequently take $\pi(\theta)$ to be a suitable limit. This approach will now be developed in detail.

Let e be the experiment which consists of one observation $\boldsymbol{x}$ from $p(\boldsymbol{x} \mid \theta)$, $\theta \in \Theta \subseteq \Re$. Suppose that we are interested in reporting inferences about θ and that no restrictions are imposed on the form of the prior distribution $p(\theta)$. It is easily verified that the amount of information about θ which k independent replications of e may be expected to provide may be rewritten as

$$I^{\theta}\{e(k), p(\theta)\} = \int p(\theta) \log \frac{f_k(\theta)}{p(\theta)}\, d\theta,$$

where

$$f_k(\theta) = \exp\left\{ \int p(\boldsymbol{z}_k \mid \theta) \log p(\theta \mid \boldsymbol{z}_k) d\boldsymbol{z}_k \right\}$$

and $\boldsymbol{z}_k = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ is a possible outcome from $e(k)$, so that

$$p(\theta \mid \boldsymbol{z}_k) \propto \prod_{i=1}^{k} p(\boldsymbol{x}_i \mid \theta) p(\theta)$$

is the posterior distribution for θ after $\boldsymbol{z}_k$ has been observed. Moreover, for any prior $p(\theta)$ one must have the constraint $\int p(\theta)\, d\theta = 1$ and, therefore, the prior $\pi_k(\theta)$ which maximises $I^{\theta}\{e(k), p(\theta)\}$ must be an extremal of the functional

$$F\{p(\cdot)\} = \int p(\theta) \log \frac{f_k(\theta)}{p(\theta)} d\theta + \lambda \left\{ \int p(\theta)\, d\theta - 1 \right\}.$$

Since this is of the form $F\{p(\cdot)\} = \int g\{p(\cdot)\}\, d\theta$, where, as a functional of $p(\cdot)$, g is twice continuously differentiable, any function $p(\cdot)$ which maximises F must satisfy the condition

$$\frac{\partial}{\partial \varepsilon} F\{p(\cdot) + \varepsilon\tau(\cdot)\}\Big|_{\varepsilon=0} = 0, \quad \text{for all } \tau.$$

It follows that, for any function τ,

$$\int \left\{ \tau(\theta) \log f_k(\theta) + \frac{p(\theta)}{f_k(\theta)} f_k'(\theta) - \tau(\theta) \left(1 + \log p(\theta)\right) + \tau(\theta)\lambda \right\} d\theta = 0,$$

where, after some algebra,

$$\begin{aligned} f_k'(\theta) &= \frac{\partial}{\partial \varepsilon} \left\{ \exp \left[\int p(\boldsymbol{z}_k \,|\, \theta) \log \frac{p(\boldsymbol{z} \,|\, \theta)\{p(\theta) + \varepsilon\tau(\theta)\}}{\int p(\boldsymbol{z}_k \,|\, \theta)\{p(\theta) + \varepsilon\tau(\theta)\}\, d\theta} d\boldsymbol{z}_k \right] \right\} \Big|_{\varepsilon=0} \\ &= f_k(\theta) \frac{\tau(\theta)}{p(\theta)} \,. \end{aligned}$$

Thus, the required condition becomes

$$\int \tau(\theta) \left\{ \log f_k(\theta) - \log p(\theta) + \lambda \right\} d\theta = 0, \quad \text{for all } \tau(\theta),$$

which implies that the desired extremal should satisfy, for all $\theta \in \Theta$,

$$\log f_k(\theta) - \log p(\theta) + \lambda = 0$$

and hence that $p(\theta) \propto f_k(\theta)$.

Note that, for each k, this only provides an *implicit* solution for the prior which maximises $I^{\theta}\{e(k), p(\theta)\}$, since $f_k(\theta)$ depends on the prior through the posterior distribution $p(\theta \,|\, \boldsymbol{z}_k) = p(\theta \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$. However, for large values of k, an approximation, $p^*(\theta \,|\, \boldsymbol{z}_k)$, say, may be found to the posterior distribution of θ, which *is* independent of the prior $p(\theta)$. It follows that, under suitable regularity conditions, the sequence of positive functions

$$p_k^*(\theta) = \exp \left\{ \int p(\boldsymbol{z}_k \,|\, \theta) \log p^*(\theta \,|\, \boldsymbol{z}_k) d\boldsymbol{z}_k \right\}$$

will induce, by formal use of Bayes' theorem, a sequence of posterior distributions

$$\pi_k(\theta \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta) p_k^*(\theta)$$

with the same limiting distributions that would have been obtained from the sequence of posteriors derived from the sequence of priors $\pi_k(\theta)$ which maximise $I^{\theta}\{e(k), p(\theta)\}$. This completes our motivation for Definition 5.7. For further information see Bernardo (1979b) and ensuing discussion.

Definition 5.7. (***One-dimensional reference distributions***).
Let $\boldsymbol{x}$ be the result of an experiment e which consists of one observation from $p(\boldsymbol{x} \mid \theta), \boldsymbol{x} \in X$, $\theta \in \Theta \subseteq \Re$, let $\boldsymbol{z}_k = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ be the result of k independent replications of e, and define

$$f_k^*(\theta) = \exp\left\{ \int p(\boldsymbol{z}_k \mid \theta) \log p^*(\theta \mid \boldsymbol{z}_k) d\boldsymbol{z}_k \right\},$$

where

$$p^*(\theta \mid \boldsymbol{z}_k) = \frac{\prod_{i=1}^k p(\boldsymbol{x}_i \mid \theta)}{\int \prod_{i=1}^k p(\boldsymbol{x}_i \mid \theta)\, d\theta}\,.$$

*The **reference posterior** density of θ after $\boldsymbol{x}$ has been observed is defined to be the log-divergence limit, $\pi(\theta \mid \boldsymbol{x})$, of $\pi_k(\theta \mid \boldsymbol{x})$, assuming this limit to exist, where*

$$\pi_k(\theta \mid \boldsymbol{x}) = c_k(\boldsymbol{x})\, p(\boldsymbol{x} \mid \theta)\, f_k^*(\theta),$$

the $c_k(\boldsymbol{x})$'s are the required normalising constants and, for almost all $\boldsymbol{x}$,

$$\lim_{k \to \infty} \int \pi_k(\theta \mid \boldsymbol{x})\, \log \frac{\pi_k(\theta \mid \boldsymbol{x})}{\pi(\theta \mid \boldsymbol{x})}\, d\theta = 0.$$

Any positive function $\pi(\theta)$ such that, for some $c(\boldsymbol{x}) > 0$ and for all $\theta \in \Theta$,

$$\pi(\theta \mid \boldsymbol{x}) = c(\boldsymbol{x})\, p(\boldsymbol{x} \mid \theta)\, \pi(\theta)$$

*will be called a **reference prior** for θ relative to the experiment e.*

It should be clear from the argument which motivates the definition that any asymptotic approximation to the posterior distribution may be used in place of the asymptotic approximation $p^*(\theta \mid \boldsymbol{z}_k)$ defined above. The use of convergence in the information sense, the natural convergence in this context, rather than just pointwise convergence, is necessary to avoid possibly pathological behaviour; for details, see Berger and Bernardo (1992c).

Although most of the following discussion refers to reference priors, it must be stressed that *only reference posterior* distributions are directly interpretable in probabilistic terms. The positive functions $\pi(\theta)$ are merely pragmatically convenient *tools* for the derivation of reference posterior distributions via Bayes' theorem. An explicit form for the reference prior is immediately available from Definition 5.7, and it will be clear from later illustrative examples that the forms which arise may have no direct probabilistic interpretation.

We should stress that the definitions and "propositions" in this section are by and large heuristic in the sense that they are lacking statements of the technical conditions which would make the theory rigorous. Making the statements and

proofs precise, however, would require a different level of mathematics from that used in this book and, at the time of writing, is still an active area of research. The reader interested in the technicalities involved is referred to Berger and Bernardo (1989, 1992a, 1992b, 1992c) and Berger *et al.* (1989). So far as the contents of this section are concerned, the reader would be best advised to view the procedure as an "algorithm, which compared with other proposals—discussed in Section 5.6.2—appears to produce appealing solutions in all situations thus far examined.

Proposition 5.18. (***Explicit form of the reference prior***).
A reference prior for θ relative to the experiment which consists of one observation from $p(\boldsymbol{x} \mid \theta)$, $\boldsymbol{x} \in X$, $\theta \in \Theta \subseteq \Re$, is given, provided the limit exists, and convergence in the information sense is verified, by

$$\pi(\theta) = c \lim_{k \to \infty} \frac{f_k^*(\theta)}{f_k^*(\theta_0)} , \qquad \theta \in \Theta$$

where $c > 0$, $\theta_0 \in \Theta$,

$$f_k^*(\theta) = \exp \left\{ \int p(\boldsymbol{z}_k \mid \theta) \log p^*(\theta \mid \boldsymbol{z}_k) d\boldsymbol{z}_k \right\},$$

with $\boldsymbol{z}_k = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ a random sample from $p(\boldsymbol{x} \mid \theta)$, and $p^*(\theta \mid \boldsymbol{z}_k)$ is an asymptotic approximation to the posterior distribution of θ.

Proof. Using $\pi(\theta)$ as a formal prior,

$$\pi(\theta \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \theta)\pi(\theta) \propto p(\boldsymbol{x} \mid \theta) \lim_{k \to \infty} \frac{f_k^*(\theta)}{f_k^*(\theta_0)} \propto \lim_{k \to \infty} \frac{p(\boldsymbol{x} \mid \theta) f_k^*(\theta)}{\int p(\boldsymbol{x} \mid \theta) f_k^*(\theta)\, d\theta} ,$$

and hence

$$\pi(\theta \mid \boldsymbol{x}) = \lim_{k \to \infty} \pi_k(\theta \mid \boldsymbol{x}), \quad \pi_k(\theta \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \theta) f_k^*(\theta)$$

as required. Note that, under suitable regularity conditions, the limits above will not depend on the particular asymptotic approximation to the posterior distribution used to derive $f_k^*(\theta)$. $\triangleleft$

If the parameter space is finite, it turns out that the reference prior is uniform, independently of the experiment performed.

Proposition 5.19. (***Reference prior in the finite case***). *Let $\boldsymbol{x}$ be the result of one observation from $p(\boldsymbol{x} \mid \theta)$, where $\theta \in \Theta = \{\theta_1, \ldots, \theta_M\}$. Then, any function of the form $\pi(\theta_i) = a$, $a > 0$, $i = 1, \ldots, M$, is a reference prior and the reference posterior is*

$$\pi(\theta_i \mid \boldsymbol{x}) = c(\boldsymbol{x}) p(\boldsymbol{x} \mid \theta_i), \quad i = 1, \ldots, M$$

where $c(\boldsymbol{x})$ is the required normalising constant.

Proof. We have already established (Proposition 5.12) that if Θ is finite then, for any strictly positive prior, $p(\theta_i \mid x_1, \ldots, x_k)$ will converge to 1 if θ_i is the true value of θ. It follows that the integral in the exponent of

$$f_k(\theta_i) = \exp\left\{ \int p(\boldsymbol{z}_k \mid \theta_i) \log p(\theta_i \mid \boldsymbol{z}_k) d\boldsymbol{z}_k \right\}, \quad i = 1, \ldots, M,$$

will converge to zero as $k \to \infty$. Hence, a reference prior is given by

$$\pi(\theta_i) = \lim_{k \to \infty} \frac{f_k(\theta_i)}{f_k(\theta_j)} = 1.$$

The general form of reference prior follows immediately. ◁

The preceding result for the case of a finite parameter space is easily derived from first principles. Indeed, in this case the expected missing information is finite and equals the entropy

$$H\{p(\theta)\} = -\sum_{i=1}^{M} p(\theta_i) \log p(\theta_i)$$

of the prior. This is maximised if and only if the prior is uniform.

The technique encapsulated in Definition 5.7 for identifying the reference prior depends on the asymptotic behaviour of the posterior for the parameter of interest under (imagined) replications of the experiment to be actually performed. Thus far, our derivations have proceeded on the basis of an assumed single observation from a parametric model, $p(\boldsymbol{x} \mid \theta)$. The next proposition establishes that for experiments involving a sequence of $n \geq 1$ observations, which are to be modelled as if they are a random sample, conditional on a parametric model, the reference prior does not depend on the size of the experiment and can thus be derived on the basis of a single observation experiment. Note, however, that for experiments involving more structured designs (for example, in linear models) the situation is much more complicated.

Proposition 5.20. (***Independence of sample size***).
Let $e_n, n \geq 1$, be the experiment which consists of the observation of a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from $p(\boldsymbol{x} \mid \theta), \boldsymbol{x} \in X$, $\theta \in \Theta$, and let $\mathcal{P}_n$ denote the class of reference priors for θ with respect to e_n, derived in accordance with Definition 5.7, by considering the sample to be a single observation from $\prod_{i=1}^{n} p(\boldsymbol{x}_i \mid \theta)$. Then $\mathcal{P}_1 = \mathcal{P}_n$, for all n.

Proof. If $\boldsymbol{z}_k = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ is the result of a k-fold independent replicate of e_1, then, by Proposition 5.18, $\mathcal{P}_1$ consists of $\pi(\theta)$ of the form

$$\pi(\theta) = c \lim_{k\to\infty} \frac{f_k^*(\theta)}{f_k^*(\theta_0)} \ ,$$

with $c > 0$, θ, $\theta_0 \in \Theta$ and

$$f_k^*(\theta) = \exp\left\{ \int p(\boldsymbol{z}_k \,|\, \theta) \log p^*(\theta \,|\, \boldsymbol{z}_k)\, d\boldsymbol{z}_k \right\},$$

where $p^*(\theta \,|\, \boldsymbol{z}_k)$ is an asymptotic approximation (as $k \to \infty$) to the posterior distribution of θ given $\boldsymbol{z}_k$.

Now consider $\boldsymbol{z}_{nk} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{2n}, \ldots, \boldsymbol{x}_{kn}\}$ which can be considered as the result of a k-fold independent replicate of e_n, so that $\mathcal{P}_n$ consists of $\pi(\theta)$ of the form

$$\pi(\theta) = c \lim_{k\to\infty} \frac{f_{nk}^*(\theta)}{f_{nk}^*(\theta_0)} \ .$$

But $\boldsymbol{z}_{nk}$ can equally be considered as a nk-fold independent replicate of e_1 and so the limiting ratios are clearly identical. ◁

In considering experiments involving random samples from distributions admitting a sufficient statistic of fixed dimension, it is natural to wonder whether the reference priors derived from the distribution of the sufficient statistic are identical to those derived from the joint distribution for the sample. The next proposition guarantees us that this is indeed the case.

Proposition 5.21. (***Compatibility with sufficient statistics***).
Let $e_n, n \geq 1$, be the experiment which consists of the observation of a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from $p(\boldsymbol{x} \,|\, \theta), \boldsymbol{x} \in X, \theta \in \Theta$, where, for all n, the latter admits a sufficient statistic $\boldsymbol{t}_n = \boldsymbol{t}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Then, for any n, the classes of reference priors derived by considering replications of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $\boldsymbol{t}_n$ respectively, coincide, and are identical to the class obtained by considering replications of e_1.

Proof. If $\boldsymbol{z}_k$ denotes a k-fold replicate of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $\boldsymbol{y}_k$ denotes the corresponding k-fold replicate of $\boldsymbol{t}_n$, then, by the definition of a sufficient statistic, $p(\theta \,|\, \boldsymbol{z}_k) = p(\theta \,|\, \boldsymbol{y}_k)$, for any prior $p(\theta)$. It follows that the corresponding asymptotic distributions are identical, so that $p^*(\theta \,|\, \boldsymbol{z}_k) = p^*(\theta \,|\, \boldsymbol{y}_k)$. We thus have

$$\begin{aligned} f_k^*(\theta) &= \exp\left\{ \int p(\boldsymbol{z}_k \,|\, \theta) \log p^*(\theta \,|\, \boldsymbol{z}_k) d\boldsymbol{z}_k \right\} \\ &= \exp\left\{ \int p(\boldsymbol{z}_k \,|\, \theta) \log p^*(\theta \,|\, \boldsymbol{y}_k) d\boldsymbol{z}_k \right\} \\ &= \exp\left\{ \int p(\boldsymbol{y}_k \,|\, \theta) \log p^*(\theta \,|\, \boldsymbol{y}_k) d\boldsymbol{y}_k \right\} \end{aligned}$$

so that, by Definition 5.7, the reference priors are identical. Identity with those derived from e_1 follows from Proposition 5.20. ◁

Given a parametric model, $p(\boldsymbol{x} \mid \theta)$, $x \in X$, $\theta \in \Theta$, we could, of course, reparametrise and work instead with $p(\boldsymbol{x} \mid \phi)$, $x \in X$, $\phi = \phi(\theta)$, for any monotone one-to-one mapping $g : \Theta \to \Phi$. The question now arises as to whether reference priors for θ and ϕ, derived from the parametric models $p(\boldsymbol{x} \mid \theta)$ and $p(\boldsymbol{x} \mid \phi)$, respectively, are consistent, in the sense that their ratio is the required Jacobian element. The next proposition establishes this form of consistency and can clearly be extended to mappings which are piecewise monotone.

Proposition 5.22. (***Invariance under one-to-one transformations***).
Suppose that $\pi_\theta(\theta)$, $\pi_\phi(\phi)$ are reference priors derived by considering replications of experiments consisting of a single observation from $p(\boldsymbol{x} \mid \theta)$, with $\boldsymbol{x} \in X$, $\theta \in \Theta$ and from $p(\boldsymbol{x} \mid \phi)$, with $x \in X, \phi \in \Phi$, respectively, where $\phi = g(\theta)$ and $g : \Theta \to \Phi$ is a one-to-one monotone mapping. Then, for some $c > 0$ and for all $\phi \in \Phi$:

(i) $\pi_\phi(\phi) = c\, \pi_\theta\left(g^{-1}(\phi)\right)$, *if Θ is discrete;*

(ii) $\pi_\phi(\phi) = c\, \pi_\theta\left(g^{-1}(\phi)\right) \mid J_\phi \mid$, *if* $J_\phi = \dfrac{\partial g^{-1}(\phi)}{\partial \phi}$ *exists.*

Proof. If Θ is discrete, so is Φ and the result follows from Proposition 5.19. Otherwise, if $\boldsymbol{z}_k$ denotes a k-fold replicate of a single observation from $p(\boldsymbol{x} \mid \theta)$, then, for any proper prior $p(\theta)$, the corresponding prior for ϕ is given by $p_\phi(\phi) = p_\theta\left(g^{-1}(\phi)\right) |J_\phi|$ and hence, for all $\phi \in \Phi$,

$$p_\phi(\phi \mid \boldsymbol{z}_k) = p_\theta\left(g^{-1}(\phi) \mid \boldsymbol{z}_k\right) |J_\phi|.$$

It follows that, as $k \to \infty$, the asymptotic posterior approximations are related by the same Jacobian element and hence

$$\begin{aligned} f_k^*(\theta) &= \exp\left\{ \int p(\boldsymbol{z}_k \mid \theta) \log p^*(\theta \mid \boldsymbol{z}_k) d\boldsymbol{z}_k \right\} \\ &= |J_\phi|^{-1} \exp\left\{ \int p(\boldsymbol{z}_k \mid \phi) \log p^*(\phi \mid \boldsymbol{z}_k) d\boldsymbol{z}_k \right\} \\ &= |J_\phi|^{-1}\, f_k^*(\phi). \end{aligned}$$

The second result now follows from Proposition 5.18. ◁

The assumed existence of the asymptotic posterior distributions that would result from an imagined k-fold replicate of the experiment under consideration clearly plays a key role in the derivation of the reference prior. However, it is important to note that no assumption has thus far been required concerning the form of this asymptotic posterior distribution. As we shall see later, we shall typically consider the case of asymptotic posterior normality, but the following example shows that the technique is by no means restricted to this case.

Example 5.12. ***(Uniform model)***. Let e be the experiment which consists of observing the sequence $x_1, \ldots, x_n, n \geq 1$, whose belief distribution is represented as that of a random sample from a uniform distribution on $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, $\theta \in \Re$, together with a prior distribution $p(\theta)$ for θ. If

$$\boldsymbol{t}_n = \left[x_{\min}^{(n)}, x_{\max}^{(n)} \right], \quad x_{\min}^{(n)} = \min\{x_1, \ldots, x_n\}, \quad x_{\max}^{(n)} = \max\{x_1, \ldots, x_n\},$$

then $\boldsymbol{t}_n$ is a sufficient statistic for θ, and

$$p(\theta \,|\, \boldsymbol{x}) = p(\theta \,|\, \boldsymbol{t}_n) \propto p(\theta), \qquad x_{\max}^{(n)} - \tfrac{1}{2} \leq \theta \leq x_{\min}^{(n)} + \tfrac{1}{2} \,.$$

It follows that, as $k \to \infty$, a k-fold replicate of e with a uniform prior will result in the posterior uniform distribution

$$p^*(\theta \,|\, \boldsymbol{t}_{kn}) \propto c, \qquad x_{\max}^{(kn)} - \tfrac{1}{2} \leq \theta \leq x_{\min}^{(kn)} + \tfrac{1}{2} \,.$$

It is easily verified that

$$\int p(\boldsymbol{t}_{kn} \,|\, \theta) \log p^*(\theta \,|\, \boldsymbol{t}_{kn}) d\boldsymbol{t}_{kn} = E\left[-\log\left\{ 1 - (x_{\max}^{(kn)} - x_{\min}^{(kn)}) \right\} \middle| \theta \right] \,,$$

the expectation being with respect to the distribution of $\boldsymbol{t}_{kn}$. For large k, the right-hand side is well-approximated by

$$-\log\left\{ 1 - \left(E\left[x_{\max}^{(kn)} \right] - E\left[x_{\min}^{(kn)} \right] \right) \right\} \,,$$

and, noting that the distributions of

$$u = x_{\max}^{(kn)} - \theta - \tfrac{1}{2} \,, \quad v = x_{\min}^{(kn)} - \theta + \tfrac{1}{2}$$

are $\mathrm{Be}(u \,|\, kn, 1)$ and $\mathrm{Be}(v \,|\, 1, kn)$, respectively, we see that the above reduces to

$$-\log\left[1 - \frac{kn}{kn+1} + \frac{1}{kn+1} \right] = \log\left(\frac{kn+1}{2} \right) \,.$$

It follows that $f_{kn}^*(\theta) = (kn+1)/2$, and hence that

$$\pi(\theta) = c \lim_{k \to \infty} \frac{(kn+1)/2}{(kn+1)/2} = c \,.$$

Any reference prior for this problem is therefore a constant and, therefore, given a set of actual data $\boldsymbol{x} = (x_1, \ldots, x_n)$, the reference posterior distribution is

$$\pi(\theta \,|\, \boldsymbol{x}) \propto c \,, \qquad x_{\max}^{(n)} - \tfrac{1}{2} \leq \theta \leq x_{\min}^{(n)} + \tfrac{1}{2} \,,$$

a uniform distribution over the set of θ values which remain possible after $\boldsymbol{x}$ has been observed.

Typically, under suitable regularity conditions, the asymptotic posterior distribution $p^*(\theta \mid \boldsymbol{z}_{kn})$, corresponding to an imagined k-fold replication of an experiment e_n involving a random sample of n from $p(\boldsymbol{x} \mid \theta)$, will only depend on $\boldsymbol{z}_{kn}$ through an *asymptotically sufficient, consistent estimate of* θ, a concept which is made precise in the next proposition. In such cases, the reference prior can easily be identified from the form of the asymptotic posterior distribution.

Proposition 5.23. (***Explicit form of the reference prior when there is a consistent, asymptotically sufficient, estimator***). *Let e_n be the experiment which consists of the observation of a random sample $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from $p(\boldsymbol{x} \mid \theta), \boldsymbol{x} \in X, \theta \in \Theta \subseteq \Re$, and let $\boldsymbol{z}_{kn}$ be the result of a k-fold replicate of e_n. If there exists $\hat{\theta}_{kn} = \hat{\theta}_{kn}(\boldsymbol{z}_{kn})$ such that, with probability one*

$$\lim_{k\to\infty} \hat{\theta}_{kn} = \theta$$

and, as $k \to \infty$,

$$\int p(\boldsymbol{z}_{kn} \mid \theta) \log \frac{p^*(\theta \mid \boldsymbol{z}_{kn})}{p^*(\theta \mid \hat{\theta}_{kn})} d\boldsymbol{z}_{kn} \to 0,$$

then, for any $c > 0, \theta_0 \in \Theta$, reference priors are defined by

$$\pi(\theta) = c \lim_{k\to\infty} \frac{f^*_{kn}(\theta)}{f^*_{kn}(\theta_0)} \ ,$$

where

$$f^*_{kn}(\theta) = p^*(\theta \mid \hat{\theta}_{kn})\Big|_{\hat{\theta}_{kn}=\theta} \ .$$

Proof. As $k \to \infty$, it follows from the assumptions that

$$\begin{aligned} f^*_{kn}(\theta) &= \exp\left\{ \int p(\boldsymbol{z}_{kn} \mid \theta) \log p^*(\theta \mid \boldsymbol{z}_{kn}) d\boldsymbol{z}_{kn} \right\} \\ &= \exp\left\{ \int p(\boldsymbol{z}_{kn} \mid \theta) \log p^*(\theta \mid \hat{\theta}_{kn}) d\boldsymbol{z}_{kn} \right\} \\ &= \exp\left\{ \int p(\hat{\theta}_{kn} \mid \theta) \log p^*(\theta \mid \hat{\theta}_{kn}) d\hat{\theta}_{kn} \right\} \\ &= \exp\left\{ \log p^*(\theta \mid \hat{\theta}_{kn})\Big|_{\hat{\theta}_{kn}=\theta} \right\} = p^*(\theta \mid \hat{\theta}_{kn})\Big|_{\hat{\theta}_{kn}=\theta} \ . \end{aligned}$$

The result now follows from Proposition 5.18. ◁

Example 5.13. ***(Deviation from uniformity model)***. Let e_n be the experiment which consists of obtaining a random sample from $p(x \mid \theta), 0 \le x \le 1, \theta > 0$, where

$$p(x \mid \theta) = \begin{cases} \theta\{2x\}^{\theta-1} & \text{for} \quad 0 \le x \le \frac{1}{2} \\ \\ \theta\{2(1-x)\}^{\theta-1} & \text{for} \quad \frac{1}{2} \le x \le 1 \end{cases}$$

defines a one-parameter probability model on $[0, 1]$, which finds application (see Bernardo and Bayarri, 1985) in exploring deviations from the standard uniform model on $[0, 1]$ (given by $\theta = 1$).

It is easily verified that if $\boldsymbol{z}_{kn} = \{x_1, \ldots, x_{kn}\}$ results from a k-fold replicate of e_n, the sufficient statistic t_{kn} is given by

$$t_{kn} = -\frac{1}{nk} \sum_{i=1}^{kn} \left\{ \log\{2x_i\} 1_{[0,1/2]}(x_i) + \log\{2(1-x_i)\} 1_{]1/2,1]}(x_i) \right\}$$

and, for any prior $p(\theta)$,

$$\begin{aligned} p(\theta \mid \boldsymbol{z}_{kn}) &= p(\theta \mid t_{kn}) \\ &\propto p(\theta)\theta^{kn} \exp\{-kn(\theta-1)t_{kn}\}. \end{aligned}$$

It is also easily shown that $p(t_{kn} \mid \theta) = \text{Ga}(t_{kn} \mid kn, kn\theta)$, so that

$$E[t_{kn} \mid \theta] = \frac{1}{\theta}, \quad V[t_{kn} \mid \theta] = \frac{1}{kn\theta^2} \,,$$

from which we can establish that $\hat{\theta}_{kn} = t_{kn}^{-1}$ is a sufficient, consistent estimate of θ. It follows that

$$p^*(\theta \mid \hat{\theta}_{kn}) \propto \theta^{kn} \exp\left\{ -\frac{kn(\theta-1)}{\hat{\theta}_{kn}} \right\}$$

provides, for large k, an asymptotic posterior approximation which satisfies the conditions required in Proposition 5.23. From the form of the right-hand side, we see that

$$\begin{aligned} p^*(\theta \mid \hat{\theta}_{kn}) &= \text{Ga}(\theta \mid kn+1, kn/\hat{\theta}_{kn}) \\ &= \frac{(kn/\hat{\theta}_{kn})^{kn+1}}{\Gamma(kn+1)} \, \theta^{kn} \exp\left\{ \frac{-kn\theta}{\hat{\theta}_{kn}} \right\} \,, \end{aligned}$$

so that

$$f_{kn}^*(\theta) = p^*(\theta \mid \hat{\theta}_{kn})\Big|_{\hat{\theta}_{kn}=\theta} = \frac{(kn)^{kn+1} e^{-nk}}{\Gamma(kn+1)\theta} \,,$$

and, from Proposition 5.18, for some $c > 0$, $\theta_0 > 0$,

$$\pi(\theta) = c \lim_{k\to\infty} \frac{f_{kn}^*(\theta)}{f_{kn}^*(\theta_0)} = \frac{c\theta_0}{\theta} \propto \frac{1}{\theta} \,.$$

The reference posterior for θ having observed actual data $\boldsymbol{x} = (x_1, \ldots, x_n)$, producing the sufficient statistic $t = t(\boldsymbol{x})$, is therefore

$$\begin{aligned} \pi(\theta \mid \boldsymbol{x}) = \pi(\theta \mid t) &\propto p(\boldsymbol{x} \mid \theta) \, \frac{1}{\theta} \\ &\propto \theta^{n-1} \exp\{-n(\theta-1)t\}, \end{aligned}$$

which is a $\text{Ga}(\theta \mid n, nt)$ distribution.

Under regularity conditions similar to those described in Section 5.2.3, the asymptotic posterior distribution of θ tends to normality. In such cases, we can obtain a characterisation of the reference prior directly in terms of the parametric model in which θ appears.

Proposition 5.24. (***Reference priors under asymptotic normality***).
Let e_n be the experiment which consists of the observation of a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from $p(\boldsymbol{x} \mid \theta)$, $\boldsymbol{x} \in X$, $\theta \in \Theta \subset \Re$. Then, if the asymptotic posterior distribution of θ, given a k-fold replicate of e_n, is normal with precision $knh(\hat{\theta}_{kn})$, where $\hat{\theta}_{kn}$ is a consistent estimate of θ, reference priors have the form

$$\pi(\theta) \propto \{h(\theta)\}^{1/2}.$$

Proof. Under regularity conditions such as those detailed in Section 5.2.3, it follows that an asymptotic approximation to the posterior distribution of θ, given a k-fold replicate of e_n, is

$$p^*(\theta \mid \hat{\theta}_{kn}) = N\left(\theta \mid \hat{\theta}_{kn}, knh(\hat{\theta}_{kn})\right),$$

where $\hat{\theta}_{kn}$ is some consistent estimator of θ. Thus, by Proposition 5.23,

$$\begin{aligned} f^*_{kn}(\theta) &= p^*(\theta \mid \hat{\theta}_{kn})\Big|_{\hat{\theta}_{kn}=\theta} \\ &= \left(\frac{kn}{2\pi}\right)^{1/2} \{h(\theta)\}^{1/2}, \end{aligned}$$

and therefore, for some $c > 0$, $\theta_0 \in \Theta$,

$$\pi(\theta) = c \lim_{k\to\infty} \frac{f^*_{kn}(\theta)}{f^*_{kn}(\theta_0)} = \frac{\{h(\theta)\}^{1/2}}{\{h(\theta_0)\}^{1/2}} \propto \{h(\theta)\}^{1/2},$$

as required. ◁

The result of Proposition 5.24 is closely related to the "rules" proposed by Jeffreys (1946, 1939/1961) and by Perks (1947) to derive "non-informative" priors. Typically, under the conditions where asymptotic posterior normality obtains we find that

$$h(\theta) = \int p(\boldsymbol{x} \mid \theta) \left(-\frac{\partial^2}{\partial\theta^2} \log p(\boldsymbol{x} \mid \theta)\right) d\boldsymbol{x},$$

i.e., *Fisher's information* (Fisher, 1925), and hence the reference prior,

$$\pi(\theta) \propto h(\theta)^{1/2},$$

becomes Jeffreys' (or Perks') prior. See Polson (1992) for a related derivation.

It should be noted however that, even under conditions which guarantee asymptotic normality, Jeffreys' formula is not necessarily the easiest way of deriving a reference prior. As illustrated in Examples 5.12 and 5.13 above, it is often simpler to apply Proposition 5.18 using an asymptotic approximation to the posterior distribution.

It is important to stress that reference distributions are, by definition, a function of the *entire* probability model $p(\boldsymbol{x} \mid \theta)$, $\boldsymbol{x} \in X$, $\boldsymbol{\theta} \in \Theta$, not only of the observed likelihood. Technically, this is a consequence of the fact that the amount of information which an experiment may be *expected* to provide is the value of an integral over the entire sample space X, which, therefore, has to be specified. We have, of course, already encountered in Section 5.1.4 the idea that knowledge of the data generating mechanism may influence the prior specification.

Example 5.14. ***(Binomial and negative binomial models)***. Consider an experiment which consists of the observation of n Bernoulli trials, with n fixed in advance, so that $\boldsymbol{x} = \{x_1, \ldots, x_n\}$,

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x}, \quad x \in \{0, 1\}, \quad 0 \leq \theta \leq 1,$$

$$h(\theta) = -\sum_{x=0}^{1} p(x \mid \theta) \frac{\partial^2}{\partial \theta^2} \log p(x \mid \theta) = \theta^{-1}(1-\theta)^{-1},$$

and hence, by Proposition 5.24, the reference prior is

$$\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}.$$

If $r = \sum_{i=1}^{n} x_i$, the reference posterior,

$$\pi(\theta \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \theta)\pi(\theta) \propto \theta^{r-1/2}(1-\theta)^{n-r-1/2},$$

is the beta distribution $\text{Be}(\theta \mid r + \frac{1}{2}, n - r + \frac{1}{2})$. Note that $\pi(\theta \mid \boldsymbol{x})$ is proper, whatever the number of successes r. In particular, if $r = 0$, $\pi(\theta \mid \boldsymbol{x}) = \text{Be}(\theta \mid \frac{1}{2}, n + \frac{1}{2})$, from which sensible inference summaries can be made, *even though there are no observed successes*. (Compare this with the Haldane (1948) prior, $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$, which produces an improper posterior until at least one success is observed.)

Consider now, however, an experiment which consists of counting the number x of Bernoulli trials which it is necessary to perform in order to observe a prespecified number of successes, $r \geq 1$. The probability model for this situation is the negative binomial

$$p(x \mid \theta) = \binom{x-1}{r-1} \theta^r (1-\theta)^{x-r}, \quad x = r, r+1, \ldots$$

from which we obtain

$$h(\theta) = -\sum_{x=r}^{\infty} p(x \mid \theta) \frac{\partial^2}{\partial \theta^2} \log p(x \mid \theta) = r\theta^{-2}(1-\theta)^{-1}$$

and hence, by Proposition 5.24, the reference prior is $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$. The reference posterior is given by

$$\pi(\theta \mid x) \propto p(x \mid \theta)\pi(\theta) \propto \theta^{r-1}(1-\theta)^{x-r-1/2}, \quad x = r, r+1, \ldots,$$

which is the beta distribution $\text{Be}(\theta \mid r, x - r + \frac{1}{2})$. Again, we note that this distribution is proper, whatever the number of observations x required to obtain r successes. Note that $r = 0$ is *not* possible under this model: the use of an inverse binomial sampling design implicitly assumes that r successes *will* eventually occur *for sure*, which is not true in direct binomial sampling. This difference in the underlying assumption about θ is duly reflected in the slight difference which occurs between the respective reference prior distributions.

See Geisser (1984) and ensuing discussion for further analysis and discussion of this canonical example.

In reporting results, scientists are typically required to specify not only the data but *also* the conditions under which the data were obtained (the *design* of the experiment), so that the data analyst has available the *full* specification of the probability model $p(\boldsymbol{x} \mid \theta)$, $\boldsymbol{x} \in X$, $\theta \in \Theta$. In order to carry out the reference analysis described in this section, such a full specification is clearly required.

We want to stress, however, that the preceding argument is totally compatible with a full personalistic view of probability. A reference prior is nothing but a (limiting) form of rather *specific* beliefs; namely, those which maximise the missing information which a *particular* experiment could possibly be expected to provide. Consequently, different experiments generally define different types of limiting beliefs. To report the corresponding reference posteriors (possibly for a range of possible alternative models) is only part of the general prior-to-posterior mapping which interpersonal or sensitivity considerations would suggest should always be carried out. Reference analysis provides an answer to an important "what if?" question: namely, what can be said about the parameter of interest *if* prior information were minimal *relative* to the maximum information which a well-defined, specific experiment could be expected to provide?

5.4.3 Restricted Reference Distributions

When analysing the inferential implications of the result of an experiment for a quantity of interest, θ, where, for simplicity, we continue to assume that $\theta \in \Theta \subseteq \Re$, it is often interesting, either *per se*, or on a "what if?" basis, to *condition* on some assumed features of the prior distribution $p(\theta)$, thus defining a restricted class, Q, say, of priors which consists of those distributions compatible with such conditioning. The concept of a reference posterior may easily be extended to this situation by maximising the missing information which the experiment may possibly be expected to provide *within* this restricted class of priors.

Repeating the argument which motivated the definition of (unrestricted) reference distributions, we are led to seek the limit of the sequence of posterior distributions, $\pi_k(\theta \mid \boldsymbol{x})$, which correspond to the sequence of priors, $\pi_k(\theta)$, which are obtained by maximising, *within* Q, the amount of information

$$I\{e(k), p(\theta)\} = \int p(\theta) \log \frac{f_k(\theta)}{p(\theta)}\, d\theta,$$

where

$$f_k(\theta) = \exp\left\{ \int p(\boldsymbol{z}_k \mid \theta) \log p(\theta \mid \boldsymbol{z}_k) d\boldsymbol{z}_k \right\},$$

which could be expected from k independent replications $z = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ of the single observation experiment.

Definition 5.8. (***Restricted reference distributions***).
Let $\boldsymbol{x}$ be the result of an experiment e which consists of one observation from $p(\boldsymbol{x} \mid \theta)$, $x \in X$, with $\theta \in \Theta \subseteq \Re$, let Q be a subclass of the class of all prior distributions for θ, let $\boldsymbol{z}_k = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ be the result of k independent replications of e and define

$$f_k^*(\theta) = \exp\left\{ \int p(\boldsymbol{z}_k \mid \theta) \log p^*(\theta \mid \boldsymbol{z}_k) d\boldsymbol{z}_k \right\},$$

where

$$p^*(\theta \mid \boldsymbol{z}_k) = \frac{\prod_{i=1}^k p(\boldsymbol{x}_i \mid \theta)}{\int \prod_{i=1}^k p(\boldsymbol{x}_i \mid \theta)\, d\theta}$$

Provided it exists, the Q-reference posterior distribution of θ, after $\boldsymbol{x}$ has been observed, is defined to be $\pi^Q(\theta \mid \boldsymbol{x})$, such that

$$E[\delta\{\pi_k^Q(\theta \mid \boldsymbol{x}), \pi^Q(\theta \mid \boldsymbol{x})\}] \to 0, \quad \textit{as} \quad k \to \infty,$$

$$\pi_k^Q(\theta \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \theta) \pi_k^Q(\theta),$$

where δ is the logarithmic divergence specified in Definition 5.7, and $\pi_k^Q(\theta)$ is a prior which minimises, within Q

$$\int p(\theta) \log \frac{p(\theta)}{f_k^*(\theta)} d\theta.$$

A positive function $\pi^Q(\theta)$ in Q such that

$$\pi^Q(\theta \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \theta) \pi^Q(\theta), \quad \textit{for all } \theta \in \Theta,$$

is then called a Q-reference prior for θ relative to the experiment e.

The intuitive content of Definition 5.8 is illuminated by the following result, which essentially establishes that the Q-reference prior is the closest prior in Q to the unrestricted reference prior $\pi(\theta)$, in the sense of minimising its logarithmic divergence from $\pi(\theta)$.

Proposition 5.25. (***The restricted reference prior as an approximation***). *Suppose that an unrestricted reference prior $\pi(\theta)$ relative to a given experiment is proper; then, if it exists, a Q-reference prior $\pi_Q(\theta)$ satisfies*

$$\int \pi^Q(\theta)\, \log \frac{\pi^Q(\theta)}{\pi(\theta)}\, d\theta = \inf_{p\in\mathcal{Q}} \int p(\theta)\, \log \frac{p(\theta)}{\pi(\theta)}\, d\theta.$$

Proof. It follows from Proposition 5.18 that $\pi(\theta)$ is proper if and only if

$$\int f_k^*(\theta)\, d\theta = c_k < \infty,$$

in which case,

$$\pi(\theta) = \lim_{k\to\infty} \pi_k(\theta) = \lim_{k\to\infty} c_k^{-1} f_k^*(\theta).$$

Moreover,

$$\begin{aligned}\int p(\theta) \log \frac{f_k^*(\theta)}{p(\theta)}\, d\theta &= -\int p(\theta) \log \frac{c_k^{-1} p(\theta)}{c_k^{-1} f_k^*(\theta)} d\theta \\ &= \log c_k - \int p(\theta) \log \frac{p(\theta)}{\pi_k(\theta)} d\theta,\end{aligned}$$

which is maximised if the integral is minimised. Let $\pi_k^Q(\theta)$ be the prior which minimises the integral within Q. Then, by Definition 5.8,

$$\pi^Q(\theta \,|\, x) \propto p(x \,|\, \theta) \lim_{k\to\infty} \pi_k^Q(\theta) = p(x \,|\, \theta)\pi^Q(\theta),$$

where, by the continuity of the divergence functional, $\pi^Q(\theta)$ is the prior which minimises, within Q,

$$\int p(\theta) \log \left\{ \frac{p(\theta)}{\lim_{k\to\infty} \pi_k(\theta)} \right\} d\theta = \int p(\theta) \log \left\{ \frac{p(\theta)}{\pi(\theta)} \right\} d\theta.$$

◁

If $\pi(\theta)$ is not proper, it is necessary to apply Definition 5.8 directly in order to characterise $\pi^Q(\theta)$. The following result provides an explicit solution for the rather large class of problems where the conditions which define Q may be expressed as a collection of expected value restrictions.

Proposition 5.26. (***Explicit form of restricted reference priors***).
Let e be an experiment which provides information about θ, and, for given $\{(g_i(\cdot), \beta_i), i = 1, \ldots, m\}$, let Q be the class of prior distributions $p(\theta)$ of θ which satisfy

$$\int g_i(\theta) p(\theta) d\theta = \beta_i, \quad i = 1, \ldots, m.$$

Let $\pi(\theta)$ be an unrestricted reference prior for θ relative to e; then, a Q-reference prior of θ relative to e, if it exists, is of the form

$$\pi^Q(\theta) \propto \pi(\theta) \exp\left\{\sum_{i=1}^{m} \lambda_i g_i(\theta)\right\},$$

where the λ_i's are constants determined by the conditions which define Q.

Proof. The calculus of variations argument which underlay the derivation of reference priors may be extended to include the additional restrictions imposed by the definition of Q, thus leading us to seek an extremal of the functional

$$\int p(\theta) \log \frac{f_k^*(\theta)}{p(\theta)}\, d\theta + \lambda \left\{\int p(\theta)\, d\theta - 1\right\} + \sum_{i=1}^{m} \lambda_i \left\{\int g_i(\theta)\, p(\theta)\, d\theta - \beta_i\right\},$$

corresponding to the assumption of a k-fold replicate of e. A standard argument now shows that the solution must satisfy

$$\log f_k^*(\theta) - \log p(\theta) + \lambda + \sum_{i=1}^{m} \lambda_i g_i(\theta) \equiv 0$$

and hence that

$$p(\theta) \propto f_k^*(\theta) \exp\left\{\sum_{i=1}^{m} \lambda_i g_i(\theta)\right\}.$$

Taking $k \to \infty$, the result follows from Proposition 5.18. ◁

Example 5.15. ***(Location models)***. Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from a location model $p(x \mid \theta) = h(x - \theta)$, $x \in \Re$, $\theta \in \Re$, and suppose that the prior mean and variance of θ are restricted to be $E[\theta] = \mu_0$, $V[\theta] = \sigma_0^2$. Under suitable regularity conditions, the asymptotic posterior distribution of θ will be of the form $p^*(\theta \mid x_1, \ldots, x_n) \propto f(\hat{\theta}_n - \theta)$, where $\hat{\theta}_n$ is an asymptotically sufficient, consistent estimator of θ. Thus, by Proposition 5.23,

$$\pi(\theta) \propto p^*(\theta \mid \hat{\theta}_n)\Big|_{\hat{\theta}_n = \theta} \propto f(0),$$

which is constant, so that the unrestricted reference prior will be *uniform*. It now follows from Proposition 5.26 that the restricted reference prior will be

$$\pi^Q(\theta) \propto \exp\left\{\lambda_1 \theta + \lambda_2 (\theta - \mu_0)^2\right\},$$

with $\int \theta \pi^Q(\theta)\, d\theta = \mu_0$ and $\int (\theta - \mu_0)^2 \pi^Q(\theta)\, d\theta = \sigma_0^2$. Thus, the restricted reference prior is the *normal* distribution with the specified mean and variance.

5.4.4 Nuisance Parameters

The development given thus far has assumed that θ was one-dimensional and that interest was centred on θ or on a one-to-one transformation of θ. We shall next consider the case where $\boldsymbol{\theta}$ is two-dimensional and interest centres on reporting inferences for a one-dimensional function, $\phi = \phi(\boldsymbol{\theta})$. Without loss of generality, we may rewrite the vector parameter in the form $\boldsymbol{\theta} = (\phi, \lambda)$, $\phi \in \Phi$, $\lambda \in \Lambda$, where ϕ is the parameter of interest and λ is a nuisance parameter. The problem is to *identify a reference prior for $\boldsymbol{\theta}$, when the decision problem is that of reporting marginal inferences for ϕ*, assuming a logarithmic score (utility) function.

To motivate our approach to this problem, consider $\boldsymbol{z}_k$ to be the result of a k-fold replicate of the experiment which consists in obtaining a single observation, $\boldsymbol{x}$, from $p(\boldsymbol{x} \mid \boldsymbol{\theta}) = p(\boldsymbol{x} \mid \phi, \lambda)$. Recalling that $p(\boldsymbol{\theta})$ can be thought of in terms of the decomposition

$$p(\boldsymbol{\theta}) = p(\phi, \lambda) = p(\phi) p(\lambda \mid \phi),$$

suppose, for the moment, that a *suitable reference form*, $\pi(\lambda \mid \phi)$, for $p(\lambda \mid \phi)$ has been specified and that only $\pi(\phi)$ remains to be identified. Proposition 5.18 then implies that the "marginal reference prior" for ϕ is given by

$$\pi(\phi) \propto \lim_{k \to \infty} \left[f_k^*(\phi) / f_k^*(\phi_0)\right], \quad \phi, \phi_0 \in \Phi,$$

where

$$f_k^*(\phi) = \exp\left\{\int p(\boldsymbol{z}_k \mid \phi) \log p^*(\phi \mid \boldsymbol{z}_k) d\boldsymbol{z}_k\right\},$$

$p^*(\phi \mid \boldsymbol{z}_k)$ is an asymptotic approximation to the marginal posterior for ϕ, and

$$\begin{aligned} p(\boldsymbol{z}_k \mid \phi) &= \int p(\boldsymbol{z}_k \mid \phi, \lambda)\pi(\lambda \mid \phi)\, d\lambda \\ &= \int \prod_{i=1}^{k} p(\boldsymbol{x}_i \mid \phi, \lambda)\pi(\lambda \mid \phi)\, d\lambda. \end{aligned}$$

By conditioning throughout on ϕ, we see from Proposition 5.18 that the "conditional reference prior" for λ given ϕ has the form

$$\pi(\lambda \mid \phi) \propto \lim_{k\to\infty} \left[\frac{f_k^*(\lambda \mid \phi)}{f_k^*(\lambda_0 \mid \phi)} \right], \quad \lambda, \lambda_0 \in \Lambda, \phi \in \Phi,$$

where

$$f_k^*(\lambda \mid \phi) = \exp\left\{ \int p(\boldsymbol{z}_k \mid \phi, \lambda) \log p^*(\lambda \mid \phi, \boldsymbol{z}_k) d\boldsymbol{z}_k \right\},$$

$p^*(\lambda \mid \phi, \boldsymbol{z}_k)$ is an asymptotic approximation to the conditional posterior for λ given ϕ, and

$$p(\boldsymbol{z}_k \mid \phi, \lambda) = \prod_{i=1}^{k} p(\boldsymbol{x}_i \mid \phi, \lambda).$$

Given actual data $\boldsymbol{x}$, the marginal reference posterior for ϕ, corresponding to the reference prior

$$\pi(\boldsymbol{\theta}) = \pi(\phi, \lambda) = \pi(\phi)\pi(\lambda \mid \phi)$$

derived from the above procedure, would then be

$$\begin{aligned} \pi(\phi \mid \boldsymbol{x}) &\propto \int \pi(\phi, \lambda \mid \boldsymbol{x})\, d\lambda \\ &\propto \pi(\phi) \int p(\boldsymbol{x} \mid \phi, \lambda)\pi(\lambda \mid \phi) d\lambda. \end{aligned}$$

This would appear, then, to provide a straightforward approach to deriving reference analysis procedures in the presence of nuisance parameters. *However, there is a major difficulty*.

In general, as we have already seen, reference priors are typically *not* proper probability densities. This means that the integrated form derived from $\pi(\lambda \mid \phi)$,

$$p(\boldsymbol{z}_k \mid \phi) = \int p(\boldsymbol{z}_k \mid \phi, \lambda)\pi(\lambda \mid \phi)\, d\lambda,$$

which plays a key role in the above derivation of $\pi(\phi)$, will typically not be a proper probability model. The above approach will fail in such cases.

Clearly, a more subtle approach is required to overcome this technical problem. However, before turning to the details of such an approach, we present an example, involving *finite* parameter ranges, where the approach outlined above does produce an interesting solution.

Example 5.16. ***(Induction)***. Consider a large, finite dichotomised population, all of whose elements individually may or may not have a specified property. A random sample is taken without replacement from the population, the sample being large in absolute size, but still relatively small compared with the population size. *All* the elements sampled turn out to have the specified property. Many commentators have argued that, in view of the large absolute size of the sample, one should be led to believe quite strongly that all elements of the *population* have the property, irrespective of the fact that the population size is greater still, an argument related to Laplace's rule of succession. (See, for example, Wrinch and Jeffreys, 1921, Jeffreys, 1939/1961, pp. 128–132 and Geisser, 1980a.)

Let us denote the population size by N, the sample size by n, the observed number of elements having the property by x, and the actual number of elements in the population having the property by θ. The probability model for the sampling mechanism is then the hypergeometric, which, for possible values of x, has the form

$$p(x \mid \theta) = \frac{\binom{\theta}{x}\binom{N-\theta}{n-x}}{\binom{N}{n}} .$$

If $p(\theta = r)$, $r = 0, \ldots, N$ defines a prior distribution for θ, the posterior probability that $\theta = N$, having observed $x = n$, is given by

$$p(\theta = N \mid x = n) = \frac{p(x = n \mid \theta = N)p(\theta = N)}{\sum_{r=n}^{N} p(x = n \mid \theta = r)p(\theta = r)} .$$

Suppose we considered θ to be the parameter of interest, and wished to provide a reference analysis. Then, since the set of possible values for θ is finite, Proposition 5.19 implies that

$$p(\theta = r) = \frac{1}{N+1}, \quad r = 0, 1, \ldots, N,$$

is a reference prior. Straightforward calculation then establishes that

$$p(\theta = N \mid x = n) = \frac{n+1}{N+1} ,$$

which is *not* close to unity when n is large but n/N is small.

However, careful consideration of the problem suggests that it is *not* θ which is the parameter of interest: rather it is the parameter

$$\phi = \begin{cases} 1 & \text{if } \theta = N \\ 0 & \text{if } \theta \neq N. \end{cases}$$

To obtain a representation of θ in the form (ϕ, λ), let us define

$$\lambda = \begin{cases} 1 & \text{if } \theta = N \\ \theta & \text{if } \theta \neq N. \end{cases}$$

By Proposition 5.19, the reference priors $\pi(\phi)$ and $\pi(\lambda \mid \phi)$ are both uniform over the appropriate ranges, and are given by

$$\pi(\phi = 0) = \pi(\phi = 1) = \tfrac{1}{2}\,,$$

$$\pi(\lambda = 1 \mid \phi = 1) = 1, \quad \pi(\lambda = r \mid \phi = 0) = \frac{1}{N}\,, \quad r = 0, 1, \ldots, N-1.$$

These imply a reference prior for θ of the form

$$p(\theta) = \begin{cases} \dfrac{1}{2} & \text{if } \theta = N \\ \dfrac{1}{2N} & \text{if } \theta \neq N \end{cases}$$

and straightforward calculation establishes that

$$p(\theta = N \mid x = n) = \left[1 + \frac{1}{(n+1)}\left(1 - \frac{n}{N}\right)\right]^{-1} \approx \frac{n+1}{n+2}\,,$$

which clearly displays the irrelevance of the sampling fraction and the approach to unity for large n (see Bernardo, 1985b, for further discussion).

We return now to the general problem of defining a reference prior for $\boldsymbol{\theta} = (\phi, \lambda)$, $\phi \in \Phi$, $\lambda \in \Lambda$, where ϕ is the parameter vector of interest and λ is a nuisance parameter. We shall refer to the pair (ϕ, λ) as an *ordered parametrisation* of the model. We recall that the problem arises because in order to obtain the marginal reference prior $\pi(\phi)$ for the first parameter we need to work with the integrated model

$$p(\boldsymbol{z}_k \mid \phi) = \int p(\boldsymbol{z}_k \mid \phi, \lambda)\pi(\lambda \mid \phi)\, d\lambda.$$

However, this will only be a proper model if the conditional prior $\pi(\lambda \mid \phi)$ for the second parameter is a proper probability density and, typically, this will not be the case.

This suggests the following strategy: identify an increasing sequence $\{\Lambda_i\}$ of subsets of Λ, $\bigcup_i \Lambda_i = \Lambda$, which may depend on ϕ, such that, on each Λ_i, the conditional reference prior, $\pi(\lambda \mid \phi)$ restricted to Λ_i can be normalised to give a reference prior, $\pi_i(\lambda \mid \phi)$, which is proper. For each i, a proper integrated model can then be obtained and a marginal reference prior $\pi_i(\phi)$ identified. The required reference prior $\pi(\phi, \lambda)$ is then obtained by taking the limit as $i \to \infty$. The strategy clearly requires a choice of the Λ_i's to be made, but in any specific problem a "natural" sequence usually suggests itself. We formalise this procedure in the next definition.

Definition 5.9. (***Reference distributions given a nuisance parameter***).
Let $\boldsymbol{x}$ be the result of an experiment e which consists of one observation from the probability model $p(\boldsymbol{x} \mid \phi, \lambda)$, $\boldsymbol{x} \in X$, $(\phi, \lambda) \in \Phi \times \Lambda \subset \Re \times \Re$. The reference posterior, $\pi(\phi \mid \boldsymbol{x})$, for the parameter of interest ϕ, relative to the experiment e and to the increasing sequences of subsets of Λ, $\{\Lambda_i(\phi)\}$, $\phi \in \Phi$, $\bigcup_i \Lambda_i(\phi) = \Lambda$, is defined to be the result of the following procedure:

(i) *applying Definition 5.7 to the model $p(\boldsymbol{x} \mid \phi, \lambda)$, for fixed ϕ, obtain the conditional reference prior, $\pi(\lambda \mid \phi)$, for Λ;*

(ii) *for each ϕ, normalise $\pi(\lambda \mid \phi)$ within each $\Lambda_i(\phi)$ to obtain a sequence of proper priors, $\pi_i(\lambda \mid \phi)$;*

(iii) *use these to obtain a sequence of integrated models*

$$p_i(\boldsymbol{x} \mid \phi) = \int_{\Lambda_i(\phi)} p(\boldsymbol{x} \mid \phi, \lambda)\pi_i(\lambda \mid \phi)\, d\lambda;$$

(iv) *use those to derive the sequence of reference priors*

$$\pi_i(\phi) = c \lim_{k\to\infty} \frac{f_k^*(\phi)}{f_k^*(\phi_0)},$$
$$f_k^*(\phi) = \exp\left\{ \int p_i(\boldsymbol{z}_k \mid \phi) \log p^*(\phi \mid \boldsymbol{z}_k) d\boldsymbol{z}_k \right\},$$

and, for data $\boldsymbol{x}$, obtain the corresponding reference posteriors

$$\pi_i(\phi \mid \boldsymbol{x}) \propto \pi_i(\phi) \int_{\Lambda_i(\phi)} p(\boldsymbol{x} \mid \phi, \lambda)\pi_i(\lambda \mid \phi)\, d\lambda;$$

(v) *define $\pi(\phi \mid \boldsymbol{x})$ such that, for almost all $\boldsymbol{x}$,*

$$\lim_{i\to\infty} \int \pi_i(\phi \mid \boldsymbol{x}) \, \log \frac{\pi_i(\phi \mid \boldsymbol{x})}{\pi(\phi \mid \boldsymbol{x})} = 0.$$

The reference prior, relative to the ordered parametrisation (ϕ, λ), is any positive function $\pi(\phi, \lambda)$, such that

$$\pi(\phi \mid \boldsymbol{x}) \propto \int p(\boldsymbol{x} \mid \phi, \lambda)\pi(\phi, \lambda)\, d\lambda.$$

This will typically be simply obtained as

$$\pi(\phi, \lambda) = \lim_{i\to\infty} \frac{\pi_i(\phi)\pi_i(\lambda \mid \phi)}{\pi_i(\phi_0)\pi_i(\lambda_0 \mid \phi_0)} \, .$$

Ghosh and Mukerjee (1992) showed that, in effect, the reference prior thus defined maximises the missing information about the parameter of interest, ϕ,

subject to the condition that, given ϕ, the missing information about the nuisance parameter, λ, is maximised.

In a model involving a parameter of interest and a nuisance parameter, the form chosen for the latter is, of course, arbitrary. Thus, $p(\boldsymbol{x} \mid \phi, \lambda)$ can be written alternatively as $p(\boldsymbol{x} \mid \phi, \psi)$, for any $\psi = \psi(\phi, \lambda)$ for which the transformation $(\phi, \lambda) \rightarrow (\phi, \psi)$ is one-to-one. Intuitively, we would hope that the reference posterior for ϕ derived according to Definition 5.9 would not depend on the particular form chosen for the nuisance parameters. The following proposition establishes that this is the case.

Proposition 5.27. (***Invariance with respect to the choice of the nuisance parameter***). *Let e be an experiment which consists in obtaining one observation from $p(\boldsymbol{x} \mid \phi, \lambda)$, $(\phi, \lambda) \in \Phi \times \Lambda \subset \Re \times \Re$, and let e' be an experiment which consists in obtaining one observation from $p(\boldsymbol{x} \mid \phi, \psi)$, $(\phi, \psi) \in \Phi \times \Psi \subseteq \Re \times \Re$, where $(\phi, \lambda) \rightarrow (\phi, \psi)$ is one-to-one transformation, with $\psi = g_\phi(\lambda)$. Then, the reference posteriors for ϕ, relative to $[e, \{\Lambda_i(\phi)\}]$ and $[e', \{\Psi_i(\phi)\}]$, where $\Psi_i(\phi) = g_\phi\{\Lambda_i(\phi)\}$, are identical.*

Proof. By Proposition 5.22, for given ϕ,

$$\pi_\psi(\psi \mid \phi) = \pi_\lambda(g_\phi^{-1}(\psi) \mid \phi) \mid J_{g_\phi^{-1}}(\psi) \mid ,$$

where

$$\psi = g_\phi(\lambda), \qquad J_\psi(\phi) = \frac{\partial g_\phi^{-1}(\psi)}{\partial \psi} \cdot$$

Hence, if we define

$$\Psi_i(\phi) = \{\psi;\ \psi = g_\phi(\lambda),\ \lambda \in \Lambda_i(\phi)\}$$

and normalise $\pi_\psi(\psi \mid \phi)$ over $\Psi_i(\phi)$ and $\pi_\lambda(g_\phi^{-1}(\psi) \mid \phi)$ over $\Lambda_i(\phi)$, we see that the normalised forms are consistently related by the appropriate Jacobian element. If we denote these normalised forms, for simplicity, by $\pi_i(\lambda \mid \phi)$, $\pi_i(\psi \mid \phi)$, we see that, for the integrated models used in steps (iii) and (iv) of Definition 5.9,

$$\begin{aligned} p_i(\boldsymbol{x} \mid \phi) &= \int_{\Lambda_i(\phi)} p(\boldsymbol{x} \mid \phi, \lambda) \pi_i(\lambda \mid \phi)\, d\lambda \\ &= \int_{\Psi_i(\phi)} p(\boldsymbol{x} \mid \phi, \psi) \pi_i(\psi \mid \phi)\, d\psi, \end{aligned}$$

and hence that the procedure will lead to identical forms of $\pi(\phi \mid \boldsymbol{x})$. $\triangleleft$

Alternatively, we may wish to consider retaining the same form of nuisance parameter, λ, but redefining the parameter of interest to be a one-to-one function of ϕ. Thus, $p(\boldsymbol{x} \mid \phi, \lambda)$ might be written as $p(\boldsymbol{x} \mid \gamma, \lambda)$, where $\gamma = g(\phi)$ is now the parameter vector of interest. Intuitively, we would hope that the reference posterior for γ would be consistently related to that of ϕ by means of the appropriate Jacobian element. The next proposition establishes that this is the case.

Proposition 5.28. (***Invariance under one-to-one transformations***).
Let e be an experiment which consists in obtaining one observation from $p(\boldsymbol{x} \mid \phi, \lambda)$, $\phi \in \Phi$, $\lambda \in \Lambda$, and let e' be an experiment which consists in obtaining one observation from $p(\boldsymbol{x} \mid \gamma, \lambda)$, $\gamma \in \Gamma, \lambda \in \Lambda$, where $\gamma = g(\phi)$. Then, given data $\boldsymbol{x}$, the reference posteriors for ϕ and γ, relative to $[e, \{\Lambda_i(\phi)\}]$ and $[e', \{\Phi_i(\gamma)\}]$, $\Phi_i(\gamma) = \Lambda_i\{g(\phi)\}$ are related by:

(i) $\pi_\gamma(\gamma \mid \boldsymbol{x}) = \pi_\phi(g^{-1}(\gamma) \mid \boldsymbol{x}), \quad$ *if Φ is discrete;*

(ii) $\pi_\gamma(\gamma \mid \boldsymbol{x}) = \pi_\phi(g^{-1}(\gamma) \mid \boldsymbol{x}) \mid J_{g^{-1}}(\gamma) \mid, \quad$ *if* $J_{g^{-1}}(\gamma) = \dfrac{\partial g^{-1}(\gamma)}{\partial \gamma}$ *exists.*

Proof. In all cases, step (i) of Definition 5.9 clearly results in a conditional reference prior $\pi(\lambda \mid \phi) = \pi(\lambda \mid g^{-1}(\gamma))$. For discrete Φ, λ, $\pi_i(\phi)$ and $\pi_i(\gamma)$ defined by steps (ii)–(iv) of Definition 5.9 are both uniform distributions, by Proposition 5.18, and the result follows straightforwardly. If $J_{g^{-1}}(\gamma)$ exists, $\pi_i(\phi)$ and $\pi_i(\gamma)$ defined by steps (ii)–(iv) of Definition 5.9 are related by the claimed Jacobian element, $\mid J_{g^{-1}}(\gamma) \mid$, by Proposition 5.22, and the result follows immediately. ◁

In Proposition 5.23, we saw that the identification of explicit forms of reference prior can be greatly simplified if the approximate asymptotic posterior distribution is of the form

$$p^*(\theta \mid \boldsymbol{z}_k) = p^*(\theta \mid \hat{\theta}_k),$$

where $\hat{\theta}_k$ is an asymptotically sufficient, consistent estimate of θ. Proposition 5.24 establishes that even greater simplification results when the asymptotic distribution is normal. We shall now extend this to the nuisance parameter case.

Proposition 5.29. (***Bivariate reference priors under asymptotic normality***).
Let e_n be the experiment which consists of the observation of a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from $p(\boldsymbol{x} \mid \phi, \lambda)$, $(\phi, \lambda) \in \Phi \times \Lambda \subseteq \Re \times \Re$, and let $\{\Lambda_i(\phi)\}$ be suitably defined sequences of subsets of λ, as required by Definition 5.9. Suppose that the joint asymptotic posterior distribution of (ϕ, λ), given a k-fold replicate of e_n, is multivariate normal with precision matrix $kn\boldsymbol{H}(\hat{\phi}_{kn}, \hat{\lambda}_{kn})$, where $(\hat{\phi}_{kn}, \hat{\lambda}_{kn})$ is a consistent estimate of (ϕ, λ) and suppose that $\hat{h}_{ij} =$

$h_{ij}(\hat{\phi}_{kn}, \hat{\lambda}_{kn})$, $i = 1, 2$, $j = 1, 2$, is the partition of $\boldsymbol{H}$ corresponding to ϕ, λ. Then

$$\pi(\lambda \mid \phi) \propto \{h_{22}(\phi, \lambda)\}^{1/2};$$
$$\pi(\phi, \lambda) = \pi(\lambda \mid \phi) \lim_{i \to \infty} \left\{ \frac{\pi_i(\phi) c_i(\phi)}{\pi_i(\phi_0) c_i(\phi_0)} \right\}, \quad \phi_0 \in \Phi,$$

define a reference prior relative to the ordered parametrisation (ϕ, λ), where

$$\pi_i(\phi) \propto \exp\left\{ \int_{\Lambda_i(\phi)} \pi_i(\lambda \mid \phi) \log\left(\{h_\phi(\phi, \lambda)\}^{1/2} \right) d\lambda \right\},$$

with

$$\pi_i(\lambda \mid \phi) = c_i(\phi) \pi(\lambda \mid \phi) = \frac{\pi(\lambda \mid \phi)}{\int_{\Lambda_i(\phi)} \pi(\lambda \mid \phi)\, d\lambda},$$

and

$$h_\phi = (h_{11} - h_{12} h_{22}^{-1} h_{21}).$$

Proof. Given ϕ, the asymptotic conditional distribution of λ is normal with precision $kn h_{22}(\phi_{kn}, \hat{\lambda}_{kn})$. The first part of Proposition 5.29 then follows from Proposition 5.24.

Marginally, the asymptotic distribution of ϕ is univariate normal with precision $kn\hat{h}_\phi$, where $h_\phi = (h_{11} - h_{12} h_{22}^{-1} h_{21})$. To derive the form of $\pi_i(\phi)$, we note that if $\boldsymbol{z}_k \in Z$ denotes the result of a k-fold replication of e_n,

$$f_{kn}^*(\phi) = \exp\left\{ \int_Z \pi_i(\boldsymbol{z}_k \mid \phi) \log p^*(\phi \mid \boldsymbol{z}_k) d\boldsymbol{z}_k \right\},$$

where, with $\pi_i(\lambda \mid \phi)$ denoting the normalised version of $\pi(\lambda \mid \phi)$ over $\Lambda_i(\phi)$, the integrand has the form

$$\begin{aligned}
&\int_Z \Big[\int_{\Lambda_i(\phi)} p(\boldsymbol{z}_k \mid \phi, \lambda) \pi_i(\lambda \mid \phi)\, d\lambda \Big] \log N(\phi \mid \hat{\phi}_{kn}, kn\hat{h}_\phi) d\boldsymbol{z}_k \\
&\qquad = \int_{\Lambda_i(\phi)} \pi_i(\lambda \mid \phi) \Big[\int_Z p(\boldsymbol{z}_k \mid \phi, \lambda) \log N(\phi \mid \hat{\phi}_{kn}, kn\hat{h}_\phi) d\boldsymbol{z}_k \Big]\, d\lambda \\
&\qquad \approx \int_{\Lambda_i(\phi)} \pi_i(\lambda \mid \phi) \log \left[\frac{\{h_\phi(\phi, \lambda)\}}{2\pi} \right]^{1/2} d\lambda,
\end{aligned}$$

for large k, so that

$$\pi_i(\phi) = \lim_{k \to \infty} \frac{f_{kn}^*(\phi)}{f_{kn}^*(\phi_0)}$$

has the stated form. Since, for data $\boldsymbol{x}$, the reference prior $\pi(\phi, \lambda)$ is defined by

$$\begin{aligned}\pi(\phi \,|\, \boldsymbol{x}) &= \lim_{i\to\infty} \pi_i(\phi \,|\, \boldsymbol{x}) \propto \lim_{i\to\infty} p_i(\boldsymbol{x} \,|\, \phi)\pi_i(\phi) \\ &\propto \lim_{i\to\infty} \pi_i(\phi) \int_{\Lambda_i} p(\boldsymbol{x} \,|\, \phi, \lambda) c_i(\phi) \pi(\lambda \,|\, \phi) d\lambda \\ &\propto \int p(\boldsymbol{x} \,|\, \phi, \lambda)\pi(\phi, \lambda) d\lambda,\end{aligned}$$

the result follows. ◁

In many cases, the forms of $\{h_{22}(\phi, \lambda)\}$ and $\{h_\phi(\phi, \lambda)\}$ factorise into products of separate functions of ϕ and λ, and the subsets $\{\Lambda_i\}$ do not depend on ϕ. In such cases, the reference prior takes on a very simple form.

Corollary. *Suppose that, under the conditions of Proposition 5.29, we choose a suitable increasing sequence of subsets $\{\Lambda_i\}$ of Λ, which do not depend on ϕ, and suppose also that*

$$\{h_\phi(\phi, \lambda)\}^{1/2} = f_1(\phi)g_1(\lambda), \quad \{h_{22}(\phi, \lambda)\}^{1/2} = f_2(\phi)g_2(\lambda).$$

Then a reference prior relative to the ordered parametrisation (ϕ, λ) is

$$\pi(\phi, \lambda) \propto f_1(\phi)g_2(\lambda)$$

Proof. By Proposition 5.29, $\pi(\lambda \,|\, \phi) \propto f_2(\phi)g_2(\lambda)$, and hence

$$\pi_i(\lambda \,|\, \phi) = a_i g_2(\lambda),$$

where $a_i^{-1} = \int_{\Lambda_i} g_2(\lambda)\ d\lambda$. It then follows that

$$\begin{aligned}\pi_i(\phi) &\propto \exp\left\{ \int_{\Lambda_i} a_i g_2(\lambda) \log[f_1(\phi)g_1(\lambda)]\ d\lambda \right\} \\ &\propto b_i f_1(\phi),\end{aligned}$$

where $b_i = \int_{\Lambda_i} a_i g_2(\lambda) \log g_1(\lambda)\ d\lambda$, and the result easily follows. ◁

Example 5.17. ***(Normal mean and standard deviation)***. Let e_n be the experiment which consists in the observation of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from a normal distribution, with both mean, μ, and standard deviation, σ, unknown. We shall first obtain a reference analysis for μ, taking σ to be the nuisance parameter.

Since the distribution belongs to the exponential family, asymptotic normality obtains and the results of Proposition 5.29 can be applied. We therefore first obtain the Fisher (expected) information matrix, whose elements we recall are given by

$$h_{ij}(\mu,\sigma) = \int \mathrm{N}(x \,|\, \mu, \sigma^{-2}) \left\{ -\frac{\partial^2 \log N(x \,|\, \mu, \sigma^{-2})}{\partial\theta_i \partial\theta_j} \right\} dx,$$

from which it is easily verified that the asymptotic precision matrix as a function of $\boldsymbol{\theta} = (\mu, \sigma)$ is given by

$$\begin{aligned} \boldsymbol{H}_{\boldsymbol{\theta}}(\mu,\sigma) &= \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{pmatrix}, \\ \{h_\mu(\mu,\sigma)\}^{1/2} &= \sigma^{-1}, \\ \{h_{22}(\mu,\sigma)\}^{1/2} &= \sqrt{2}\sigma^{-1}. \end{aligned}$$

This implies that

$$\pi(\sigma \,|\, \mu) \propto \{h_{22}(\mu,\sigma)\}^{1/2} \propto \sigma^{-1},$$

so that, for example, $\Lambda_i = \{\sigma; e^{-i} \leq \sigma \leq e^i\}, i = 1, 2, \ldots$, provides a suitable sequence of subsets of $\Lambda = \Re^+$ not depending on μ, over which $\pi(\sigma \,|\, \mu)$ can be normalised and the corollary to Proposition 5.29 can be applied. It follows that

$$\pi(\mu,\sigma) = \pi(\mu)\pi(\sigma \,|\, \mu) \propto 1 \times \sigma^{-1}$$

provides a reference prior relative to the ordered parametrisation (μ, σ). The corresponding reference posterior for μ, given $\boldsymbol{x}$, is

$$\begin{aligned} \pi(\mu \,|\, \boldsymbol{x}) &\propto \int p(\boldsymbol{x} \,|\, \mu, \sigma)\pi(\mu,\sigma)\, d\sigma \\ &\propto \pi(\mu) \int \prod_{i=1}^{n} \mathrm{N}(x_i \,|\, \mu, \sigma)\pi(\sigma \,|\, \mu)\, d\sigma \\ &\propto \int \sigma^{-n} \exp\left\{ -\frac{n}{2\sigma^2} \left[(\overline{x} - \mu)^2 + s^2 \right] \right\} \sigma^{-1}\, d\sigma \\ &\propto \int \lambda^{n/2-1} \exp\left\{ -\frac{n\lambda}{2} \left[(\overline{x} - \mu)^2 + s^2 \right] \right\} d\lambda \\ &\propto \left[s^2 + (\mu - \overline{x})^2 \right]^{-n/2} \\ &= \mathrm{St}(\mu \,|\, \overline{x}, (n-1)s^{-2}, n-1), \end{aligned}$$

where $ns^2 = \Sigma(x_i - \overline{x})^2$.

If we now reverse the roles of μ and σ, so that the latter is now the parameter of interest and μ is the nuisance parameter, we obtain, writing $\boldsymbol{\phi} = (\sigma, \mu)$

$$\boldsymbol{H}_{\boldsymbol{\phi}}(\sigma,\mu) = \begin{pmatrix} 2\sigma^{-2} & 0 \\ 0 & \sigma^{-2} \end{pmatrix},$$

so that $\{h_\sigma(\sigma,\mu)\}^{1/2} = \sqrt{2}\sigma^{-1}$, $h_{22}(\sigma,\mu)\}^{1/2} = \sigma^{-1}$ and, by a similar analysis to the above,

$$\pi(\mu \,|\, \sigma) \propto \sigma^{-1}$$

so that, for example, $\Lambda_i = \{\mu; -e^i \leq \mu \leq e^i\}, i = 1, 2, \ldots$ provides a suitable sequence of subsets of $\Lambda = \Re$ not depending on σ, over which $\pi(\mu \,|\, \sigma)$ can be normalised and the corollary to Proposition 5.29 can be applied. It follows that

$$\pi(\mu, \sigma) = \pi(\sigma)\pi(\mu \,|\, \sigma) \propto 1 \times \sigma^{-1}$$

provides a reference prior relative to the ordered parametrisation (σ, μ). The corresponding reference posterior for σ, given $\boldsymbol{x}$, is

$$\begin{aligned}\pi(\sigma \,|\, \boldsymbol{x}) &\propto \int p(\boldsymbol{x} \,|\, \mu, \sigma)\ \pi(\mu, \sigma)\ \, d\mu \\ &\propto \pi(\sigma) \int \prod_{i=1}^{n} \mathrm{N}(x_i \,|\, \mu, \sigma)\ \pi(\mu \,|\, \sigma)\ \, d\mu,\end{aligned}$$

the right-hand side of which can be written in the form

$$\sigma^{-n} \exp\left\{-\frac{ns^2}{2\sigma^2}\right\} \int \sigma^{-1} \exp\left\{-\frac{n}{2\sigma^2}(\mu - \overline{x})^2\right\} d\mu.$$

Noting, by comparison with a $N(\mu \,|\, \overline{x}, n\lambda)$ density, that the integral is a constant, and changing the variable to $\lambda = \sigma^{-2}$, implies that

$$\begin{aligned}\pi(\lambda \,|\, \boldsymbol{x}) &\propto \lambda^{(n-1)/2-1} \exp\left\{\tfrac{1}{2}ns^2\lambda\right\} \\ &= \mathrm{Ga}\left(\lambda \,\middle|\, \tfrac{1}{2}(n-1),\ \tfrac{1}{2}ns^2\right),\end{aligned}$$

or, alternatively,

$$\begin{aligned}\pi(\lambda ns^2 \,|\, \boldsymbol{x}) &= \mathrm{Ga}\left(\lambda ns^2 \,\middle|\, \tfrac{1}{2}(n-1),\ \tfrac{1}{2}\right) \\ &= \chi^2(\lambda ns^2 \,|\, n-1).\end{aligned}$$

One feature of the above example is that the reference prior did not, in fact, depend on which of the parameters was taken to be the parameter of interest. In the following example the form does change when the parameter of interest changes.

Example 5.18. ***(Standardised normal mean)***. We consider the same situation as that of Example 5.17, but we now take $\phi = \mu/\sigma$ to be the parameter of interest. If σ is taken as the nuisance parameter (by Proposition 5.27 the choice is irrelevant), $\boldsymbol{\psi} = (\phi, \sigma) = \boldsymbol{g}(\mu, \sigma)$ is clearly a one-to-one transformation, with

$$\boldsymbol{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{\psi}) = \begin{pmatrix} \dfrac{\partial\mu}{\partial\phi} & \dfrac{\partial\mu}{\partial\sigma} \\ \dfrac{\partial\sigma}{\partial\phi} & \dfrac{\partial\sigma}{\partial\sigma} \end{pmatrix} = \begin{pmatrix} \sigma & \phi \\ 0 & 1 \end{pmatrix}$$

and using Corollary 1 to Proposition 5.17.

$$\boldsymbol{H}_{\boldsymbol{\psi}}(\boldsymbol{\psi}) = \boldsymbol{J}^{t}_{\boldsymbol{g}^{-1}}(\boldsymbol{\psi})\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{g}^{-1}(\boldsymbol{\psi}))\boldsymbol{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{\psi}) = \begin{pmatrix} 1 & \phi\sigma^{-1} \\ \phi\sigma^{-1} & \sigma^{-2}(2+\phi^2) \end{pmatrix}.$$

Again, the sequence $\Lambda_i = \{\sigma; e^{-i} \leq \sigma \leq e^{i}\}, i = 1, 2, \ldots$, provides a reasonable basis for applying the corollary to Proposition 5.29. It is easily seen that

$$\begin{aligned} |\, h_{\phi}(\phi,\sigma)\,|^{\,1/2} &= \frac{|\, h(\phi,\sigma)\,|^{\,1/2}}{|\, h_{22}(\phi,\sigma)\,|^{\,1/2}} \propto (2+\phi^2)^{-1/2}, \\ |\, h_{22}(\phi,\sigma)\,|^{\,1/2} &\propto (2+\phi^2)^{1/2}\sigma^{-1}, \end{aligned}$$

so that the reference prior relative to the ordered parametrisation (ϕ, σ) is given by

$$\pi(\phi,\sigma) \propto (2+\phi^2)^{-1/2}\sigma^{-1}.$$

In the (μ, σ) parametrisation this corresponds to

$$\pi(\mu,\sigma) \propto \left(2+\frac{\mu^2}{\sigma^2}\right)^{-1/2}\sigma^{-2},$$

which is clearly different from the form obtained in Example 5.17. Further discussion of this example will be provided in Example 5.26 of Section 5.6.2.

We conclude this subsection by considering a rather more involved example, where a natural choice of the required $\Lambda_i(\phi)$ subsequence *does* depend on ϕ. In this case, we use Proposition 5.29, since its corollary does not apply.

Example 5.19. ***(Product of normal means)***. Consider the case where independent random samples $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ and $\boldsymbol{y} = \{y_1, \ldots, y_m\}$ are to be taken, respectively, from $N(x \,|\, \alpha, 1)$ and $N(y \,|\, \beta, 1)$, $\alpha > 0$, $\beta > 0$, so that the complete parametric model is

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

for which, writing $\boldsymbol{\theta} = (\alpha, \beta)$ the Fisher information matrix is easily seen to be

$$\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \boldsymbol{H}(\alpha,\beta) = \begin{pmatrix} n & 0 \\ 0 & m \end{pmatrix}.$$

Suppose now that we make the one-to-one transformation $\boldsymbol{\psi} = (\phi, \lambda) = (\alpha\beta, \alpha/\beta) = \boldsymbol{g}(\alpha, \beta) = \boldsymbol{g}(\boldsymbol{\theta})$, so that $\phi = \alpha\beta$ is taken to be the parameter of interest and $\lambda = \alpha/\beta$ is taken to be the nuisance parameter. Such a parameter of interest arises, for example, when

inference about the area of a rectangle is required from data consisting of measurements of its sides.

The Jacobian of the inverse transformation is given by

$$\boldsymbol{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{\psi}) = \begin{pmatrix} \frac{\partial\alpha}{\partial\phi} & \frac{\partial\alpha}{\partial\lambda} \\ \frac{\partial\beta}{\partial\phi} & \frac{\partial\beta}{\partial\lambda} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \left(\frac{\lambda}{\phi}\right)^{1/2} & \left(\frac{\phi}{\lambda}\right)^{1/2} \\ \left(\frac{1}{\phi\lambda}\right)^{1/2} & -\frac{1}{\lambda}\left(\frac{\phi}{\lambda}\right)^{1/2} \end{pmatrix}$$

and hence, using Corollary 1 to Proposition 5.17

$$\boldsymbol{H}_{\boldsymbol{\psi}}(\boldsymbol{\psi}) = \boldsymbol{J}^{t}_{\boldsymbol{g}^{-1}}(\boldsymbol{\psi}) \boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{g}^{-1}(\boldsymbol{\psi})) \boldsymbol{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{\psi}) = \frac{nm}{4\lambda^2} \begin{bmatrix} \frac{\lambda}{\phi}\left(\frac{\lambda^2}{m} + \frac{1}{n}\right) & \left(\frac{\lambda^2}{m} - \frac{1}{n}\right) \\ \left(\frac{\lambda^2}{m} - \frac{1}{n}\right) & \phi\left(\frac{\lambda}{m} + \frac{1}{n\lambda}\right) \end{bmatrix},$$

with $|\boldsymbol{H}_{\boldsymbol{\psi}}(\boldsymbol{\psi})| = \frac{nm}{4\lambda^2}$, so that

$$\pi(\lambda \,|\, \phi) \propto |h_{22}(\phi, \lambda)|^{1/2} \propto \frac{(nm\phi)^{1/2}}{\lambda} \left(\frac{\lambda}{m} + \frac{1}{n\lambda}\right)^{1/2}.$$

The question now arises as to what constitutes a "natural" sequence $\{\lambda_i(\phi)\}$, over which to define the normalised $\pi_i(\lambda \,|\, \phi)$ required by Definition 5.9. A natural increasing sequence of subsets of the original parameter space, $\Re^+ \times \Re^+$, for (α, β) would be the sets

$$S_i = \{(\alpha, \beta); \quad 0 < \alpha < i, 0 < \beta < i\}, \quad i = 1, 2, \ldots,$$

which transform, in the space of $\lambda \in \Lambda$, into the sequence

$$\Lambda_i(\phi) = \left\{\lambda; \quad \frac{\phi}{i^2} < \lambda < \frac{i^2}{\phi}\right\}.$$

We note that unlike in the previous cases we have considered, this does depend on ϕ.

To complete the analysis, it can be shown, after some manipulation, that, for large i,

$$\pi_i(\lambda \,|\, \phi) = \frac{\sqrt{nm}}{i(\sqrt{m} + \sqrt{n})} \phi^{1/2} \lambda^{-1} \left(\frac{1}{m} + \frac{1}{n\lambda}\right)^{1/2}$$

and

$$\pi_i(\phi) = \frac{\sqrt{nm}}{i\,(\sqrt{m} + \sqrt{n})} \int_{\Lambda_i(\phi)} \left(\frac{\lambda}{m} + \frac{1}{n\lambda}\right)^{1/2} \lambda^{-1} \log\left(\frac{\lambda}{m} + \frac{1}{n\lambda}\right)^{-1/2} d\lambda,$$

which leads to a reference prior relative to the ordered parametrisation (ϕ, λ) given by

$$\pi(\phi, \lambda) \propto \phi^{1/2} \lambda^{-1} \left(\frac{\lambda}{m} + \frac{1}{n\lambda}\right)^{1/2}$$

In the original parametrisation, this corresponds to

$$\pi(\alpha, \beta) \propto (n\alpha^2 + m\beta^2)^{1/2},$$

which depends on the sample sizes through the ratio m/n and reduces, in the case $n = m$, to $\pi(\alpha, \beta) \propto (\alpha^2 + \beta^2)^{1/2}$, a form originally proposed for this problem in an unpublished 1982 Stanford University technical report by Stein, who showed that it provides approximate agreement between Bayesian credible regions and classical confidence intervals for ϕ. For a detailed discussion of this example, and of the consequences of choosing a different sequence $\Lambda_i(\phi)$, see Berger and Bernardo (1989).

We note that the preceding example serves to illustrate the fact that reference priors may depend explicitly on the sample sizes defined by the experiment. There is, of course, nothing paradoxical in this, since the underlying notion of a reference analysis is a "minimally informative" prior *relative* to the actual experiment to be performed.

5.4.5 Multiparameter Problems

The approach to the nuisance parameter case considered above was based on the use of an ordered parametrisation whose first and second components were (ϕ, λ), referred to, respectively, as the *parameter of interest* and the nuisance parameter. The reference prior for the *ordered* parametrisation (ϕ, λ) was then constructed by conditioning to give the form $\pi(\lambda \mid \phi)\pi(\phi)$.

When the model parameter vector $\boldsymbol{\theta}$ has more than two components, this successive conditioning idea can obviously be extended by considering $\boldsymbol{\theta}$ as an ordered parametrisation, $(\theta_1, \ldots, \theta_m)$, say, and generating, by successive conditioning, a reference prior, *relative to this ordered parametrisation*, of the form

$$\pi(\boldsymbol{\theta}) = \pi(\theta_m \mid \theta_1, \ldots, \theta_{m-1}) \cdots \pi(\theta_2 \mid \theta_1)\pi(\theta_1).$$

In order to describe the algorithm for producing this successively conditioned form, in the standard, regular case we shall first need to introduce some notation.

Assuming the parametric model $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, to be such that the Fisher information matrix

$$\boldsymbol{H}(\boldsymbol{\theta}) = -E_{\boldsymbol{x} \mid \boldsymbol{\theta}} \left\{ \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) \right\}$$

has full rank, we define $\boldsymbol{S}(\theta) = \boldsymbol{H}^{-1}(\theta)$, define the component vectors

$$\boldsymbol{\theta}^{[j]} = (\theta_1, \ldots, \theta_j), \quad \boldsymbol{\theta}_{[j]} = (\theta_{j+1}, \ldots, \theta_m),$$

and denote by $\boldsymbol{S}_j(\boldsymbol{\theta})$ the corresponding upper left $j \times j$ submatrix of $\boldsymbol{S}(\boldsymbol{\theta})$, and by $h_j(\theta)$ the lower right element of $\boldsymbol{S}_j^{-1}(\boldsymbol{\theta})$.

Finally, we assume that $\Theta = \Theta_1 \times \cdots \times \Theta_m$, with $\theta_i \in \Theta_i$, and, for $i = 1, 2, \ldots$, we denote by $\{\Theta_i^l\}$, $l = 1, 2, \ldots$, an increasing sequence of compact subsets of Θ_i, and define $\Theta_{[j]}^l = \Theta_{j+1}^l \times \cdots \times \Theta_m^l$.

Proposition 5.30. (***Ordered reference priors under asymptotic normality***). *With the above notation, and under regularity conditions extending those of Proposition 5.29 in an obvious way, the reference prior $\pi(\boldsymbol{\theta})$, relative to the ordered parametrisation $(\theta_1, \ldots, \theta_m)$, is given by*

$$\pi(\boldsymbol{\theta}) = \lim_{l\to\infty} \frac{\pi^l(\boldsymbol{\theta})}{\pi^l(\boldsymbol{\theta}^*)}, \quad \textit{for some } \boldsymbol{\theta}^* \in \Theta,$$

where $\pi^l(\boldsymbol{\theta})$ is defined by the following recursion:

(i) For $j = m$, and $\theta_m \in \Theta_m^l$,

$$\pi_m^l\left(\theta_{[m-1]} \,|\, \theta^{[m-1]}\right) = \pi_m^l\left(\theta_m \,|\, \theta_1, \ldots, \theta_{m-1}\right) = \frac{\{h_m(\boldsymbol{\theta})\}^{1/2}}{\int_{\Theta_m^l} \{h_m(\boldsymbol{\theta})\}^{1/2}\, d\theta_m} \cdot$$

(ii) For $j = m-1, m-2, \ldots, 2$, and $\theta_j \in \Theta_j^l$,

$$\pi_j^l\left(\boldsymbol{\theta}_{[j-1]} \,|\, \boldsymbol{\theta}^{[j-1]}\right) = \pi_{j+1}^l\left(\boldsymbol{\theta}_{[j]} \,|\, \boldsymbol{\theta}^{[j]}\right) \frac{\exp\left\{E_j^l\left[\log\{h_j(\boldsymbol{\theta})\}^{1/2}\right]\right\}}{\int_{\Theta_j^l} \exp\left\{E_j^l\left[\log\{h_j(\boldsymbol{\theta})\}^{1/2}\right]\right\} d\theta_j},$$

where

$$E_j^l\left[\log\{h_j(\boldsymbol{\theta})\}^{1/2}\right] = \int_{\Theta_{[j]}^l} \log\{h_j(\boldsymbol{\theta})\}^{1/2} \pi_{j+1}^l\left(\boldsymbol{\theta}_{[j]} \,|\, \boldsymbol{\theta}^{[j]}\right) d\boldsymbol{\theta}_{[j]}.$$

(iii) For $j = 1$, $\boldsymbol{\theta}_{[0]} = \boldsymbol{\theta}$, with $\boldsymbol{\theta}^{[0]}$ vacuous, and

$$\pi^l(\boldsymbol{\theta}) = \pi_1^l\left(\boldsymbol{\theta}_{[0]} | \boldsymbol{\theta}^{[0]}\right).$$

Proof. This follows closely the development given in Proposition 5.29. For details see Berger and Bernardo (1992a, 1992b, 1992c). ◁

The derivation of the ordered reference prior is greatly simplified if the $\{h_j(\boldsymbol{\theta})\}$ terms in the above depend only on $\theta^{[j]}$: even greater simplification obtains if $\boldsymbol{H}(\boldsymbol{\theta})$ is block diagonal, particularly, if, for $j = 1, \ldots, m$, the jth term can be factored into a product of a function of θ_j and a function not depending on θ_j.

Corollary. *If $h_j(\boldsymbol{\theta})$ depends only on $\boldsymbol{\theta}^{[j]}$, $j = 1, \ldots, m$, then*

$$\pi^l(\boldsymbol{\theta}) = \prod_{j=1}^m \frac{\{h_j(\boldsymbol{\theta})\}^{1/2}}{\int_{\Theta_j^l} \{h_j(\boldsymbol{\theta})\}^{1/2}\, d\theta_j}, \quad \boldsymbol{\theta} \in \Theta^l.$$

If $\boldsymbol{H}(\boldsymbol{\theta})$ is block diagonal (i.e., $\theta_1, \ldots, \theta_m$ are mutually orthogonal), with

$$\boldsymbol{H}(\boldsymbol{\theta}) = \begin{pmatrix} h_{11}(\boldsymbol{\theta}) & 0 & \cdots & 0 \\ 0 & h_{22}(\boldsymbol{\theta}) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & h_{mm}(\boldsymbol{\theta}) \end{pmatrix},$$

then $h_j(\boldsymbol{\theta}) = h_{jj}(\boldsymbol{\theta}), j = 1, \ldots, m$. Furthermore, if, in this latter case,

$$\{h_{jj}(\boldsymbol{\theta})\}^{1/2} = f_j(\theta_j) g_j(\boldsymbol{\theta}),$$

where $g_j(\boldsymbol{\theta})$ does not depend on θ_j, and if the Θ_j^l's do not depend on $\boldsymbol{\theta}$, then

$$\pi(\boldsymbol{\theta}) \propto \prod_{j=1}^{m} f_j(\theta_j).$$

Proof. The results follow from the recursion of Proposition 5.29. ◁

The question obviously arises as to the appropriate ordering to be adopted in any specific problem. At present, no formal theory exists to guide such a choice, but experience with a wide range of examples suggests that—at least for non-hierarchical models (see Section 4.6.5), where the parameters may have special forms of interrelationship—the best procedure is to order the components of θ on the basis of their inferential interest.

Example 5.20. ***(Reference analysis for m normal means)***. Let e_n be an experiment which consists in obtaining $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $n \geq 2$, a random sample from the multivariate normal model $N_m(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \tau \boldsymbol{I}_m)$, $m \geq 1$, for which the Fisher information matrix is easily seen to be

$$\boldsymbol{H}(\boldsymbol{\mu}, \tau) = \begin{pmatrix} \tau \boldsymbol{I}_m & 0 \\ 0 & mn/(2\tau^2) \end{pmatrix}.$$

It follows from Proposition 5.30 that the reference prior relative to the natural parametrisation $(\mu_1, \ldots, \mu_m, \tau)$, is given by

$$\pi(\mu_1, \ldots, \mu_m, \tau) \propto \tau^{-1}.$$

Clearly, in this example the result does not, in fact, depend on the order in which the parametrisation is taken, since the parameters are all mutually orthogonal.

The reference prior $\pi(\mu_1, \ldots, \mu_m, \tau) \propto \tau^{-1}$ or $\pi(\mu_1, \ldots, \mu_m, \sigma) \propto \sigma^{-1}$ if we parametrise in terms of $\sigma = \tau^{-1/2}$, is thus the appropriate reference form if we are interested in any of the individual parameters. The reference posterior for any μ_j is easily shown to be the Student density

$$\pi(\mu_j \,|\, x_1, \ldots, x_n) = \mathrm{St}\left(\mu_j \,|\, \overline{x}_j, (n-1)s^{-2}, m(n-1)\right)$$

$$n\overline{x}_j = \sum_{i=1}^{n} x_{ij}, \qquad nms^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} (x_{ij} - \overline{x}_j)^2$$

which agrees with the standard argument according to which one degree of freedom should be lost by each of the unknown means.

Example 5.21. ***(Multinomial model)***. Let $\boldsymbol{x} = \{r_1, \ldots, r_m\}$ be an observation from a multinomial distribution (see Section 3.2), so that

$$p(r_1, \ldots, r_m \,|\, \theta_1, \ldots, \theta_m) = \frac{n!}{r_1! \cdots r_m! (n - \Sigma r_i)!} \theta_1^{r_1} \cdots \theta_m^{r_m} (1 - \Sigma\theta_i)^{n - \Sigma r_i},$$

from which the Fisher information matrix

$$\boldsymbol{H}(\theta_1, \ldots, \theta_m) = \frac{n}{1 - \Sigma\theta_i} \begin{bmatrix} \dfrac{1 + \theta_1 - \Sigma\theta_i}{\theta_1} & 1 & \ldots & 1 \\ 1 & \dfrac{1 + \theta_2 - \Sigma\theta_i}{\theta_2} & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & \ldots & \dfrac{1 + \theta_m - \Sigma\theta_i}{\theta_m} \end{bmatrix}$$

is easily derived, with

$$|\boldsymbol{H}| = n^m \left[\left(1 - \sum_{i=1}^{m} \theta_1 \right) \prod_{i=1}^{m} \theta_i \right]^{-1}.$$

In this case, the conditional reference priors derived using Proposition 5.28 turn out to be proper, and there is no need to consider subset sequences $\{\Theta_i^l\}$. In fact, noting that $H^{-1}(\theta_1, \ldots, \theta_m)$ is given by

$$\frac{1}{n} \begin{bmatrix} \theta_1(1 - \theta_1) & -\theta_1\theta_2 & \cdots & -\theta_1\theta_m \\ -\theta_1\theta_2 & \theta_2(1 - \theta_2) & \cdots & -\theta_2\theta_m \\ \cdots & \cdots & \cdots & \cdots \\ -\theta_1\theta_m & -\theta_2\theta_m & \cdots & \theta_m(1 - \theta_m) \end{bmatrix},$$

we see that the conditional asymptotic precisions used in Proposition 5.29 are easily identified, and hence that

$$\pi(\theta_j \,|\, \theta_1, \ldots, \theta_{j-1}) \propto \left(\frac{1 - \sum_{i=1}^{j-1} \theta_i}{\theta_j} \right)^{1/2} \left(\frac{1}{1 - \sum_{i=1}^{j} \theta_i} \right)^{1/2}, \quad \theta_j \leq 1 - \sum_{i=1}^{j-1} \theta_i.$$

The required reference prior relative to the ordered parametrisation $(\theta_1, \ldots, \theta_m)$, say, is then given by

$$\begin{aligned} \pi(\theta_1, \ldots, \theta_m) &\propto \pi(\theta_1)\pi(\theta_2 \,|\, \theta_1) \cdots \pi(\theta_m \,|\, \theta_1, \ldots, \theta_{m-1}) \\ &\propto \theta_1^{-1/2}(1 - \theta_1)^{-1/2} \theta_2^{-1/2}(1 - \theta_1 - \theta_2)^{-1/2} \cdots \theta_m^{-1/2}(1 - \theta_1 - \cdots - \theta_m)^{-1/2}, \end{aligned}$$

and corresponding reference posterior for θ_1 is

$$\pi(\theta_1 \,|\, r_1, \ldots, r_m) \propto \int p(r_1, \ldots, r_m \,|\, \theta_1, \ldots, \theta_m)\, \pi(\theta_1, \ldots, \theta_m)\, d\theta_2 \ldots d\theta_m,$$

which is proportional to

$$\int \theta_1^{r_1-1/2}\cdots\theta_m^{r_m-1/2}(1-\Sigma\theta_i)^{n-\Sigma r_i} \times (1-\theta_1)^{-1/2}(1-\theta_1-\theta_2)^{-1/2}\cdots(1-\theta_1-\cdots-\theta_m)^{-1/2}d\theta_2\cdots d\theta_m.$$

After some algebra, this implies that

$$\pi(\theta_1 \,|\, r_1,\ldots,r_m) = \text{Be}\left(\theta_1 \,|\, r_1+\tfrac{1}{2}, n-r_1+\tfrac{1}{2}\right),$$

which, as one could expect, coincides with the reference posterior which would have been obtained had we initially collapsed the multinomial analysis to a binomial model and then carried out a reference analysis for the latter. Clearly, by symmetry considerations, the above analysis applies to any θ_i, $i=1,\ldots,m$, after appropriate changes in labelling and it is independent of the particular order in which the parameters are taken. For a detailed discussion of this example see Berger and Bernardo (1992a). Further comments on ordering of parameters are given in Section 5.6.2.

Example 5.22. ***(Normal correlation coefficient)***. Let $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ be a random sample from a bivariate normal distribution, $N_2(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\tau})$, where

$$\boldsymbol{\mu} = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \quad \boldsymbol{\tau}^{-1} = \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}.$$

Suppose that the correlation coefficient ρ is the parameter of interest, and consider the ordered parametrisation $\{\rho,\mu_1,\mu_2,\sigma_1,\sigma_2\}$. It is easily seen that

$$\boldsymbol{H}(\rho,\mu_1,\mu_2,\sigma_1,\sigma_2) = (1-\rho^2)^{-1}\begin{bmatrix} \frac{1+\rho^2}{1-\rho^2} & 0 & 0 & \frac{-\rho}{\sigma_1} & \frac{-\rho}{\sigma_2}\\ 0 & \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} & 0 & 0\\ 0 & \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} & 0 & 0\\ \frac{-\rho}{\sigma_1} & 0 & 0 & \frac{2-\rho^2}{\sigma_1^2} & \frac{-\rho^2}{\sigma_1\sigma_2}\\ \frac{-\rho}{\sigma_2} & 0 & 0 & \frac{-\rho^2}{\sigma_1\sigma_2} & \frac{2-\rho^2}{\sigma_2^2}\end{bmatrix},$$

so that

$$\boldsymbol{H}^{-1} = \begin{bmatrix} (1-\rho^2)^2 & 0 & 0 & \frac{\sigma_1}{2}\rho(1-\rho^2) & \frac{\sigma_2}{2}\rho(1-\rho^2)\\ 0 & \sigma_1^2 & \rho\sigma_1\sigma_2 & 0 & 0\\ 0 & \rho\sigma_1\sigma_2 & \sigma_2^2 & 0 & 0\\ \frac{\sigma_1}{2}\rho(1-\rho^2) & 0 & 0 & \frac{\sigma_1^2}{2} & \rho^2\frac{\sigma_1\sigma_2}{2}\\ \frac{\sigma_2}{2}\rho(1-\rho^2) & 0 & 0 & \rho^2\frac{\sigma_1\sigma_2}{2} & \frac{\sigma_2^2}{2}\end{bmatrix}.$$

After some algebra it can be shown that this leads to the reference prior

$$\pi(\rho, \mu_1, \mu_2, \sigma_1, \sigma_2) \propto (1-\rho^2)^{-1}\sigma_1^{-1}\sigma_2^{-1},$$

whatever ordering of the nuisance parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$ is taken. This agrees with Lindley's (1965, p. 219) analysis. Furthermore, as one could expect from Fisher's (1915) original analysis, the corresponding reference posterior distribution for ρ

$$\pi(\rho \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \propto \frac{(1-\rho^2)^{(n-3)/2}}{(1-\rho r)^{n-3/2}} F\left(\frac{1}{2}, \frac{1}{2}, n-\frac{1}{2}, \frac{1+\rho r}{2}\right),$$

(where F is the hypergeometric function), only depends on the data through the sample correlation coefficient r, whose sampling distribution only depends on ρ. For a detailed analysis of this example, see Bayarri (1981); further discussion will be provided in Section 5.6.2.

See, also, Hills (1987), Ye and Berger (1991) and Berger and Bernardo (1992b) for derivations of the reference distributions for a variety of other interesting models.

Infinite discrete parameter spaces

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

Example 5.23. ***(Infinite discrete case)***. In the context of capture-recapture problems, suppose it is of interest to make inferences about an integer $\theta \in \{1, 2, \ldots\}$ on the basis of a random sample $\boldsymbol{z} = \{x_1, \ldots, x_n\}$ from

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

For several plausible "diffuse looking" prior distributions for θ one finds that the corresponding posterior virtually ignores the data. Intuitively, this has to be interpreted as suggesting that such priors actually contain a large amount of information about θ compared with that provided by the data. A more careful approach to providing a "non-informative" prior is clearly required. One possibility would be to embed the discrete space $\{1, 2, \ldots\}$ in the continuous space $]0, \infty[$ since, for each $\theta > 0$, $p(x|\theta)$ is still a probability density for x. Then, using Proposition 5.24, the appropriate refrence prior is

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

and it is easily verified that this prior leads to a posterior in which the data are no longer overwhelmed. If the physical conditions of the problem require the use of discrete θ values, one could always use, for example,

$$p(\theta = 1 \,|\, \boldsymbol{z}) = \int_0^{3/2} \pi(\theta \,|\, \boldsymbol{z}) d\theta, \qquad p(\theta = j \,|\, \boldsymbol{z}) = \int_{j-1/2}^{j+1/2} \pi(\theta|\boldsymbol{z}) d\theta, \quad j > 1$$

as an approximate discrete reference posterior.

Prediction and Hierarchical Models

Two classes of problems that are not covered by the methods so far discussed are hierarchical models and prediction problems. The difficulty with these problems is that there are unknowns (typically the unknowns of interest) that have specified distributions. For instance, if one wants to predict y based on $\boldsymbol{z}$ when $(y, \boldsymbol{z})$ has density $p(y, \boldsymbol{z} \mid \boldsymbol{\theta})$, the unknown of interest is y, but its distribution is conditionally specified. One needs a reference prior for $\boldsymbol{\theta}$, not y. Likewise, in a hierarchical model with, say, $\mu_1, \mu_2, \ldots, \mu_p$ being $\mathrm{N}(\mu_i \mid \mu_0, \lambda)$, the μ_i's may be the parameters of interest but a prior is only needed for the hyperparameters μ_0 and λ.

The obvious way to approach such problems is to integrate out the variables with conditionally known distributions (y in the predictive problem and the $\{\mu_i\}$ in the hierarchical model), and find the reference prior for the remaining parameters based on this marginal model. The difficulty that arises is how to then identify parameters of interest and nuisance parameters to construct the ordering necessary for applying the reference prior method, the real parameters of interest having been integrated out.

In future work, we propose to deal with this difficulty by defining the parameter of interest in the reduced model to be the conditional mean of the original parameter of interest. Thus, in the prediction problem, $E[y|\boldsymbol{\theta}]$ (which will be either $\boldsymbol{\theta}$ or some transformation thereof) will be the parameter of interest, and in the hierarchical model $E[\mu_i \mid \mu_0, \lambda] = \mu_0$ will be defined to be the parameter of interest. This technique has so far worked well in the examples to which it has been applied, but further study is clearly needed.

5.5 NUMERICAL APPROXIMATIONS

Section 5.3 considered forms of approximation appropriate as the sample size becomes large relative to the amount of information contained in the prior distribution. Section 5.4 considered the problem of approximating a prior specification maximising the expected information to be obtained from the data. In this section, we shall consider numerical techniques for implementing Bayesian methods for arbitrary forms of likelihood and prior specification, and arbitrary sample size.

We note that the technical problem of evaluating quantities required for Bayesian inference summaries typically reduces to the calculation of a ratio of two integrals. Specifically, given a likelihood $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ and a prior density $p(\boldsymbol{\theta})$, the starting point for all subsequent inference summaries is the joint posterior density for $\boldsymbol{\theta}$ given by

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, d\boldsymbol{\theta}} \, .$$

From this, we may be interested in obtaining univariate marginal posterior densities for the components of $\boldsymbol{\theta}$, bivariate joint marginal posterior densities for pairs of

components of $\boldsymbol{\theta}$, and so on. Alternatively, we may be interested in marginal posterior densities for functions of components of $\boldsymbol{\theta}$ such as ratios or products.

In all these cases, the technical key to the implementation of the formal solution given by Bayes' theorem, for specified likelihood and prior, is the ability to perform a number of integrations. First, we need to evaluate the denominator in Bayes' theorem in order to obtain the normalising constant of the posterior density; then we need to integrate over complementary components of $\boldsymbol{\theta}$, or transformations of $\boldsymbol{\theta}$, in order to obtain marginal (univariate or bivariate) densities, together with summary moments, highest posterior density intervals and regions, or whatever. Except in certain rather stylised problems (e.g., exponential families together with conjugate priors), the required integrations will not be feasible analytically and, thus, efficient approximation strategies will be required.

In this section, we shall outline five possible numerical approximation strategies, which will be discussed under the subheadings: *Laplace Approximation; Iterative Quadrature; Importance Sampling; Sampling-importance-resampling; Markov Chain Monte Carlo*. An exhaustive account of these and other methods will be given in the second volume in this series, *Bayesian Computation*.

5.5.1 Laplace Approximation

We motivate the approximation by noting that the technical problem of evaluating quantities required for Bayesian inference summaries, is typically that of evaluating an integral of the form

$$E\left[g(\boldsymbol{\theta}) \mid \boldsymbol{x}\right] = \int g(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{\theta},$$

where $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ is derived from a predictive model with an appropriate representation as a mixture of parametric models, and $g(\boldsymbol{\theta})$ is some real-valued function of interest. Often, $g(\boldsymbol{\theta})$ is a first or second moment, and since $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ is given by

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}},$$

we see that $E[g(\boldsymbol{\theta}) \mid \boldsymbol{x})]$ has the form of a ratio of two integrals.

Focusing initially on this situation of a required inference summary for $g(\boldsymbol{\theta})$, and assuming $g(\boldsymbol{\theta})$ almost everywhere positive, we note that the posterior expectation of interest can be written in the form

$$E\left[g(\boldsymbol{\theta}) \mid \boldsymbol{x}\right] = \frac{\int \exp\{-nh^*(\boldsymbol{\theta})\} d\boldsymbol{\theta}}{\int \exp\{-nh(\boldsymbol{\theta})\} d\boldsymbol{\theta}}$$

where, with the vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ of observations fixed, the functions $h(\boldsymbol{\theta})$ and $h^*(\boldsymbol{\theta})$ are defined by

$$-nh(\boldsymbol{\theta}) = \log p(\boldsymbol{\theta}) + \log p(\boldsymbol{x} \mid \boldsymbol{\theta}),$$
$$-nh^*(\boldsymbol{\theta}) = \log g(\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + \log p(\boldsymbol{x} \mid \boldsymbol{\theta}).$$

Let us consider first the case of a single unknown parameter, $\boldsymbol{\theta} = \theta \in \Re$, and define $\hat{\theta}, \theta^*$ and $\hat{\sigma}, \sigma^*$ such that

$$-h(\hat{\theta}) = \sup_{\theta} \{-h(\theta)\}, \qquad \hat{\sigma} = [h''(\theta)]^{-1/2}\Big|_{\theta=\hat{\theta}} \quad ,$$
$$-h^*(\theta^*) = \sup_{\theta} \{-h^*(\theta)\}, \quad \sigma^* = [h^{*\prime\prime}(\theta)]^{-1/2}\Big|_{\theta=\theta^*} \quad .$$

Assuming $h(\cdot)$, $h^*(\cdot)$ to be suitably smooth functions, the *Laplace approximations* for the two integrals defining the numerator and denominator of $E[g(\theta) \mid \boldsymbol{x}]$ are given (see, for example, Jeffreys and Jeffreys, 1946) by

$$\sqrt{2\pi}\sigma^* n^{-1/2} \exp\{-nh^*(\theta^*)\},$$

and

$$\sqrt{2\pi}\hat{\sigma} n^{-1/2} \exp\left\{-nh(\hat{\theta})\right\}.$$

Essentially, the approximations consist of retaining quadratic terms in Taylor expansions of $h(\cdot)$ and $h^*(\cdot)$, and are thus equivalent to normal-like approximations to the integrands. In the context we are considering, it then follows immediately that the resulting approximation for $E[g(\theta) \mid \boldsymbol{x}]$ has the form

$$\hat{E}[g(\theta) \mid \boldsymbol{x}] = \left(\frac{\sigma^*}{\hat{\sigma}}\right) \exp\left\{-n\left[h^*(\theta^*) - h(\hat{\theta})\right]\right\},$$

and Tierney and Kadane (1986) have shown that

$$E[g(\theta) \mid \boldsymbol{x}] = \hat{E}[g(\theta) \mid \boldsymbol{x}]\left(1 + O(n^{-2})\right).$$

The Laplace approximation approach, exploiting the fact that Bayesian inference summaries typically involve ratios of integrals, is thus seen to provide a potentially very powerful general approximation technique. See, also, Tierney, Kass and Kadane (1987, 1989a, 1989b), Kass, Tierney and Kadane (1988, 1989a, 1989b, 1991) and Wong and Li (1992) for further underpinning of, and extensions to, this methodology.

Considering now the general case of $\boldsymbol{\theta} \in \Re^k$, the Laplace approximation to the denominator of $E[g(\boldsymbol{\theta}) \mid \boldsymbol{x}]$ is given by

$$\int \exp\{-nh(\boldsymbol{\theta})\} d\boldsymbol{\theta} = (2\pi)^{k/2} \left|n\nabla^2 h(\hat{\boldsymbol{\theta}})\right|^{-1/2} \exp\left\{-nh(\hat{\boldsymbol{\theta}})\right\},$$

where $\hat{\boldsymbol{\theta}}$ is defined by

$$-h(\hat{\boldsymbol{\theta}}) = \sup_{\boldsymbol{\theta}} h(\boldsymbol{\theta})$$

and

$$\left[\nabla^2 h(\hat{\boldsymbol{\theta}})\right]_{ij} = \left.\frac{\partial^2 h(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}},}$$

the Hessian matrix of h evaluated at $\hat{\boldsymbol{\theta}}$, with an exactly analogous expression for the numerator, defined in terms of $h^*(\cdot)$ and $\boldsymbol{\theta}^*$. Writing

$$\hat{\sigma} = \left|n\nabla^2 h(\hat{\boldsymbol{\theta}})\right|^{-1/2}$$

$$\sigma^* = \left|n\nabla^2 h^*(\boldsymbol{\theta}^*)\right|^{-1/2},$$

the Laplace approximation to $E[g(\boldsymbol{\theta}) \mid \boldsymbol{x}]$ is given by

$$\hat{E}[g(\boldsymbol{\theta}) \mid \boldsymbol{x}] = \left(\frac{\sigma^*}{\sigma}\right) \exp\big\{ - n[h^*(\boldsymbol{\theta}^*) - h(\hat{\boldsymbol{\theta}})\,]\big\},$$

completely analogous to the univariate case.

If $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\lambda})$ and the required inference summary is the marginal posterior density for $\boldsymbol{\phi}$, application of the Laplace approximation approach corresponds to obtaining $p(\boldsymbol{\phi} \mid \boldsymbol{x})$ pointwise by fixing $\boldsymbol{\phi}$ in the numerator and defining $g(\boldsymbol{\lambda}) = 1$. It is easily seen that this leads to

$$\begin{aligned}\hat{p}(\boldsymbol{\phi} \mid \boldsymbol{x}) &\propto \int \exp\left\{-nh_{\boldsymbol{\phi}}(\boldsymbol{\lambda})\right\} d\boldsymbol{\lambda} \\ &\propto \left|\nabla^2 h_\phi(\hat{\boldsymbol{\lambda}}_\phi)\right|^{-1/2} \exp\left\{-nh_\phi(\hat{\boldsymbol{\lambda}}_\phi)\right\}\ ,\end{aligned}$$

where

$$-nh_\phi(\boldsymbol{\lambda}) = \log p(\boldsymbol{\phi}, \boldsymbol{\lambda}) + \log p(\boldsymbol{x} \mid \boldsymbol{\phi}, \boldsymbol{\lambda}),$$

considered as a function of $\boldsymbol{\lambda}$ for fixed $\boldsymbol{\phi}$, and

$$-h_\phi(\hat{\boldsymbol{\lambda}}_\phi) = -\sup_{\boldsymbol{\lambda}} h_\phi(\boldsymbol{\lambda}).$$

The form $\hat{p}(\boldsymbol{\phi} \mid \boldsymbol{x})$ thus provides (up to proportionality) a pointwise approximation to the ordinates of the marginal posterior density for $\boldsymbol{\phi}$. Considering this form in more detail, we see that, if $p(\boldsymbol{\phi}, \boldsymbol{\lambda})$ is constant,

$$\hat{p}(\boldsymbol{\phi} \mid \boldsymbol{x}) \propto \left|-\nabla^2 \log p(\boldsymbol{x} \mid \boldsymbol{\phi}, \hat{\boldsymbol{\lambda}}_\phi)\right|\ p(\boldsymbol{x} \mid \boldsymbol{\phi}, \hat{\boldsymbol{\lambda}}_\phi).$$

The form $\nabla^2 \log p(\boldsymbol{x} \mid \boldsymbol{\phi}, \hat{\boldsymbol{\lambda}}_\phi)$ is the Hessian of the log-likelihood function, considered as a function of $\boldsymbol{\lambda}$ for fixed $\boldsymbol{\phi}$, and evaluated at the value $\hat{\boldsymbol{\lambda}}_\phi$ which maximises the log-likelihood over $\boldsymbol{\lambda}$ for fixed $\boldsymbol{\phi}$; the form $p(\boldsymbol{x} \mid \boldsymbol{\phi}, \hat{\boldsymbol{\lambda}}_\phi)$ is usually called the *profile likelihood* for $\boldsymbol{\phi}$, corresponding to the parametric model $p(\boldsymbol{x} \mid \boldsymbol{\phi}, \boldsymbol{\lambda})$. The approximation to the marginal density for $\boldsymbol{\phi}$ given by $\hat{p}(\boldsymbol{\phi} \mid \boldsymbol{x})$ has a form often referred to as the *modified profile likelihood* (see, for example, Cox and Reid, 1987, for a convenient discussion of this terminology). Approximation to Bayesian inference summaries through Laplace approximation is therefore seen to have links with forms of inference summary proposed and derived from a non-Bayesian perspective. For further references, see Appendix B, Section 4.2.

In relation to the above analysis, we note that the Laplace approximation is essentially derived by considering normal approximations to the integrands appearing in the numerator and denominator of the general form $E[g(\boldsymbol{\theta}) \mid \boldsymbol{x}]$. If the forms concerned are not well approximated by second-order Taylor expansions of the exponent terms of the integrands, which may be the case with small or moderate samples, particularly when components of $\boldsymbol{\theta}$ are constrained to ranges other than the real line, we may be able to improve substantially on this direct Laplace approximation approach.

One possible alternative, at least if $\boldsymbol{\theta} = \theta$ is a scalar parameter, is to attempt to approximate the integrands by forms other than normal, perhaps resembling more the actual posterior shapes, such as gammas or betas. Such an approach has been followed in the one-parameter case by Morris (1988), who develops a general approximation technique based around the Pearson family of densities. These are characterised by parameters m, μ_0 and a quadratic function Q, which specify a density for θ of the form

$$q_Q(\theta \mid m, \mu_0) = K_Q(m, \mu_0) \frac{p(\theta)}{Q(\theta)} \, ,$$

where

$$\begin{aligned} p(\theta) &= \exp\left\{ -m \int \left(\frac{\theta - \mu_0}{Q(\theta)} \right) d\theta \right\} \, , \\ K_Q^{-1}(m, \mu_0) &= \int \left(\frac{p(\theta)}{Q(\theta)} \right) d\theta, \\ Q(\theta) &= q_0 + q_1 \theta + q_2 \theta^2 \end{aligned}$$

and the range of θ is such that $0 < Q(\theta) < \infty$.

It is shown by Morris (1988) that, for a given choice of quadratic function Q, an analogue to the Laplace-type approximation of an integral of a unimodal function $f(\theta)$ is given by

$$\int f(\theta) \, d\theta = \frac{f(\hat{\theta})}{q_Q(\hat{\theta} \mid \hat{m}, \hat{\theta})} \, ,$$

where $\hat{m} = r''(\hat{\theta})Q(\hat{\theta})$ and $\hat{\theta}$ maximises $r(\theta) = \log[f(\theta)Q(\theta)]$. Details of the forms of K^{-1}, Q and p for familiar forms of Pearson densities are given in Morris (1988), where it is also shown that the approximation can often be further simplified to the expression

$$\int f(\theta)d\theta = \frac{f(\hat{\theta})\sqrt{2\pi}}{[-r''(\hat{\theta})]^{1/2}} \,.$$

A second alternative is to note that the version of the Laplace approximation proposed by Tierney and Kadane (1986) is not invariant to changes in the (arbitrary) parametrisation chosen when specifying the likelihood and prior density functions. It may be, therefore, that by judicious reparametrisation (of the likelihood, together with the appropriate, Jacobian adjusted, prior density) the Laplace approximation can itself be made more accurate, even in contexts where the original parametrisation does not suggest the plausibility of a normal-type approximation to the integrands. We, note, incidentally, that such a strategy is also available in multiparameter contexts, whereas the Pearson family approach does not seem so readily generalisable.

To provide a concrete illustration of these alternative analytic approximation approaches consider the following.

Example 5.24. ***(Approximating the mean of a beta distribution)***.

Suppose that a posterior beta distribution, $\text{Be}(\theta \,|\, r_n - \frac{1}{2}, n - r_n + \frac{1}{2})$, has arisen from a $\text{Bi}(r_n \,|\, n, \theta)$ likelihood, together with, $\text{Be}(\theta \,|\, \frac{1}{2}, \frac{1}{2})$ prior (the reference prior, derived in Example 5.14). Writing $r_n = x$, we can, in fact, identify the analytic form of the posterior mean in this case,

$$E[\theta \,|\, x] = \frac{x + \frac{1}{2}}{n + 1} \,,$$

but we shall ignore this for the moment and examine approximations implied by the techniques discussed above.

First, defining $g(\theta) = \theta$, we see, after some algebra, that the Tierney-Kadane form of the Laplace approximation gives the estimated posterior mean

$$\hat{E}[\theta \,|\, x] = \frac{(n-1)^{n+1/2}(x + \frac{1}{2})^{x+1}}{n^{n+3/2}(x - \frac{1}{2})^{x}} \,.$$

If, instead, we reparametrise to $\phi = \sin^{-1}\sqrt{\theta}$, the required integrals are defined in terms of

$$g(\phi) = \sin^2\phi, \quad p(\boldsymbol{x} \,|\, \phi) \propto (\sin^2\phi)^x(1 - \sin^2\phi)^{n-x}, \quad \pi(\phi) \propto 1 \,,$$

and the Laplace approximation can be shown to be

$$\tilde{E}[\theta \,|\, x] = \frac{n^{n+1/2}(x+1)^{x+1}}{(n+1)^{n+3/2}x^{x}} \,.$$

Alternatively, if we work via the Pearson family, with $Q(\theta) = \theta(1-\theta)$ as the "natural" choice for a beta-like posterior, we obtain

$$E^*[\theta \mid x] = \frac{(n+1)^{n+1/2}\,(x+3/2)^{x+1}}{(n+2)^{n+\frac{3}{2}}\,\left(x+\frac{1}{2}\right)^{x}}\,.$$

By considering the percentage errors of estimation, defined by

$$100 \times \left| \frac{\text{true} - \text{estimated}}{\text{true}} \right|\,,$$

we can study the performance of the three estimates for various values of n and x. Details are given in Achcar and Smith (1989); here, we simply summarise, in Table 5.1, the results for $n = 5$, $x = 3$, which typify the performance of the estimates for small n.

Table 5.1 *Approximation of* $E[\theta \mid x]$ *from* $\text{Be}(\theta \mid x+\frac{1}{2}, n-x+\frac{1}{2})$ *(percentage errors in parentheses)*

True value	*Laplace approximations*		*Pearson approximation*
	$\hat{E}[\theta \mid x]$	$\tilde{E}[\theta \mid x]$	$E^*[\theta \mid x]$
0.583	0.563 (3.6%)	0.580 (0.6%)	0.585 (0.3%)

We see from Table 5.1 that the Pearson approximation, which is, in some sense, preselected to be best, does, in fact, outperform the others. However, it is striking that the performance of the Laplace approximation under reparametrisation leads to such a considerable improvement over that based on the original parametrisation, and is a very satisfactory alternative to the "optimal" Pearson form. Further examples are given in Achcar and Smith (1989).

In general, it would appear that, in cases involving a relatively small number of parameters, the Laplace approach, in combination with judicious reparametrisation, can provide excellent approximations to general Bayesian inference summaries, whether in the form of posterior moments or marginal posterior densities. However, in multiparameter contexts there may be numerical problems with the evaluation of local derivatives in cases where analytic forms are unobtainable or too tedious to identify explicitly. In addition, there are awkward complications if the integrands are multimodal. At the time of writing, this area of approximation theory is very much still an active research field and the full potential of this and related methods (see, also, Lindley, 1980b, Leonard *et al.*, 1989) has yet to be clarified.

5.5.2 Iterative Quadrature

It is well known that univariate integrals of the type

$$\int_{-\infty}^{\infty} \mathrm{e}^{-t^2} f(t)dt$$

are often well approximated by Gauss-Hermite quadrature rules of the form

$$\sum_{i=1}^{n} w_i f(t_i),$$

where t_i is the ith zero of the Hermite polynomial $H_n(t)$. In particular, if $f(t)$ is a polynomial of degree at most $2n-1$, then the quadrature rule approximates the integral without error. This implies, for example, that, if $h(t)$ is a suitably well behaved function and

$$g(t) = h(t)\,(2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2 \right\},$$

then

$$\int_{-\infty}^{\infty} g(t)dt \approx \sum_{i=1}^{n} m_i g(z_i),$$

where

$$m_i = w_i \exp(t_i^2)\sqrt{2\sigma}, \quad z_i = \mu + \sqrt{2\sigma}\, t_i$$

(see, for example, Naylor and Smith, 1982).

It follows that Gauss-Hermite rules are likely to prove very efficient for functions which, expressed in informal terms, closely resemble "polynomial $\times$ normal" forms. In fact, this is a rather rich class which, even for moderate n (less than 12, say), covers many of the likelihood $\times$ prior shapes we typically encounter for parameters defined on $(-\infty, \infty)$. Moreover, the applicability of this approximation is vastly extended by working with suitable transformations of parameters defined on other ranges such as $(0, \infty)$ or (a, b), using, for example, $\log(t)$ or $\log(t-a) - \log(b-t)$, respectively. Of course, to use the above form we must specify μ and σ in the normal component. It turns out that, given reasonable starting values (from any convenient source, prior information, maximum likelihood estimates, etc.), we typically can successfully iterate the quadrature rule, substituting estimates of the posterior mean and variance obtained using previous values of m_i and z_i. Moreover, we note that if the posterior density is well-approximated by the product of a normal and a polynomial of degree at most $2n-3$, then an n-point Gauss-Hermite rule will prove effective for *simultaneously evaluating the normalising constant and the first and second moments*, using the same (iterated)

set of m_i and z_i. In practice, it is efficient to begin with a small grid size ($n = 4$ or $n = 5$) and then to gradually increase the grid size until stable answers are obtained both within and between the last two grid sizes used.

Our discussion so far has been for the one-dimensional case. Clearly, however, the need for an efficient strategy is most acute in higher dimensions. The "obvious" extension of the above ideas is to use a cartesian product rule giving the approximation

$$\int \dots \int f(t_1, \dots, t_k) dt_1 \dots dt_k \approx \sum_{i_k} m_{i_k}^{(k)} g(z_{i_1}^{(k)}, \dots, z_{i_k}^{(k)}),$$

where the grid points and the weights are found by substituting the appropriate iterated estimates of μ and σ^2 corresponding to the marginal component t_j.

The problem with this "obvious" strategy is that the product form is only efficient if we are able to make an (at least approximate) assumption of posterior independence among the individual components. In this case, the lattice of integration points formed from the product of the two one-dimensional grids will efficiently cover the bulk of the posterior density. However, if high posterior correlations exist, these will lead to many of the lattice points falling in areas of negligible posterior density, thus causing the cartesian product rule to provide poor estimates of the normalising constant and moments.

To overcome this problem, we could first apply individual parameter transformations of the type discussed above and then attempt to transform the resulting parameters, via an appropriate linear transformation, to a new, approximately orthogonal, set of parameters. At the first step, this linear transformation derives from an initial guess or estimate of the posterior covariance matrix (for example, based on the observed information matrix from a maximum likelihood analysis). Successive transformations are then based on the estimated covariance matrix from the previous iteration.

The following general strategy has proved highly effective for problems involving up to six parameters (see, for example, Naylor and Smith, 1982, Smith *et al.*, 1985, 1987, Naylor and Smith, 1988).

(1) Reparametrise individual parameters so that the resulting working parameters all take values on the real line.

(2) Using initial estimates of the joint posterior mean vector and covariance matrix for the working parameters, transform further to a centred, scaled, more "orthogonal" set of parameters.

(3) Using the derived initial location and scale estimates for these "orthogonal" parameters, carry out, on suitably dimensioned grids, cartesian product integration of functions of interest.

(4) Iterate, successively updating the mean and covariance estimates, until stable results are obtained both within and between grids of specified dimension.

For problems involving larger numbers of parameters, say between six and twenty, cartesian product approaches become computationally prohibitive and alternative approaches to numerical integration are required.

One possibility is the use of spherical quadrature rules (Stroud, 1971, Sections 2.6, and 2.7), derived by transforming from cartesian to spherical polar coordinates and constructing optimal integration formulae based on symmetric configurations over concentric spheres. Full details of this approach will be given in the volume *Bayesian Computation*. For a brief introduction, see Smith (1991). Other relevant references on numerical quadrature include Shaw (1988b), Flournoy and Tsutakawa (1991), O'Hagan (1991) and Dellaportas and Wright (1992).

The efficiency of numerical quadrature methods is often very dependent on the particular parametrisation used. For further information on this topic, see Marriott (1988), Hills and Smith (1992, 1993) and Marriott and Smith (1992). For related discussion, see Kass and Slate (1992).

The ideas outlined above relate to the use of numerical quadrature formulae to implement Bayesian statistical methods. It is amusing to note that the roles can be reversed and Bayesian statistical methods used to derive optimal numerical quadrature formulae! See, for example, Diaconis (1988b) and O'Hagan (1992).

5.5.3 Importance Sampling

The importance sampling approach to numerical integration is based on the observation that, if f is a function and g is a probability density function

$$\begin{aligned}\int f(x)dx &= \int \left[\frac{f(x)}{g(x)}\right] g(x)dx \\ &= \int \left[\frac{f(x)}{g(x)}\right] dG(x) \\ &= E_G\left[\frac{f(x)}{g(x)}\right] ,\end{aligned}$$

which suggest the "statistical" approach of generating a sample from the distribution function G—referred to in this context as the *importance sampling* distribution—and using the average of the values of the ratio f/g as an unbiased estimator of $\int f(x)dx$. However, the variance of such an estimator clearly depends critically on the choice of G, it being desirable to choose g to be "similar" to the shape of f.

In multiparameter Bayesian contexts, exploitation of this idea requires designing importance sampling distributions which are efficient for the kinds of integrands arising in typical Bayesian applications. A considerable amount of work has focused on the use of multivariate normal or Student forms, or modifications thereof,

much of this work motivated by econometric applications. We note, in particular, the contributions of Kloek and van Dijk (1978), van Dijk and Kloek (1983, 1985), van Dijk *et al.* (1987) and Geweke (1988, 1989).

An alternative line of development (Shaw, 1988a) proceeds as follows. In the univariate case, if we choose g to be heavier-tailed than f, and if we work with $y = G(x)$, the required integral is the expected value of $f[G^{-1}(x)]/g[G^{-1}(x)]$ with respect to a uniform distribution on the interval $(0, 1)$. Owing to the periodic nature of the ratio function over this interval, we are likely to get a reasonable approximation to the integral by simply taking some equally spaced set of points on $(0, 1)$, rather than actually generating "uniformly distributed" random numbers. If f is a function of more than one argument (k, say), an exactly parallel argument suggess that the choice of a suitable g followed by the use of a suitably selected "uniform" configuration of points in the k-dimensional unit hypercube will provide an efficient multidimensional integration procedure.

However, the effectiveness of all this depends on choosing a suitable G, bearing in mind that we need to have available a flexible set of possible distributional shapes, for which G^{-1} is available explicitly. In the univariate case, one such family defined on $\Re$ is provided by considering the random variable

$$x_a = a\, h(u) - (1 - a)\, h(1 - u),$$

where u is uniformly distributed on $(0, 1)$, $h : (0, 1) \to \Re$ is a monotone increasing function such that

$$\lim_{u \to 0} h(u) = -\infty$$

and $0 \leq a \leq 1$ is a constant. The choice $a = 0.5$ leads to symmetric distributions; as $a \to 0$ or $a \to 1$ we obtain increasingly skew distributions (to the left or right). The tail-behaviour of the distributions is governed by the choice of the function h. Thus, for example, $h(u) = \log(u)$ leads to a family whose symmetric member is the logistic distribution; $h(u) = -\tan[\pi(1 - u)/2]$ leads to a family whose symmetric member is the Cauchy distribution. Moreover, the moments of the distributions of the x_a are polynomials in a (of corresponding order), the median is linear in a, etc., so that sample information about such quantities provides (for any given choice of h) operational guidance on the appropriate choice of a. To use this family in the multiparameter case, we again employ individual parameter transformations, so that all parameters belong to $\Re$, together with "orthogonalising" transformations, so that parameters can be treated "independently". In the transformed setting, it is natural to consider an iterative importance sampling strategy which attempts to learn about an appropriate choice of G for each parameter.

As we remarked earlier, part of this strategy requires the specification of "uniform" configurations of points in the k-dimensional unit hypercube. This problem has been extensively studied by number theorists and systematic experimentation

with various suggested forms of "quasi-random" sequences has identified effective forms of configuration for importance sampling purposes: for details, see Shaw (1988a). The general strategy is then the following.

(1) Reparametrise individual parameters so that resulting working parameters all take values on the real line.

(2) Using initial estimates of the posterior mean vector and covariance matrix for the working parameters, transform to a centred, scaled, more "orthogonal" set of parameters.

(3) In terms of these transformed parameters, set

$$g(\boldsymbol{x}) = \prod_{j=1}^{k} g_j(x_j),$$

for "suitable" choices of g_j, $j = 1, \ldots, k$.

(4) Use the inverse distribution function transformation to reduce the problem to that of calculating an average over a "suitable" uniform configuration in the k-dimensional hypercube.

(5) Use information from this "sample" to learn about skewness, tailweight, etc. for each g_j, and hence choose "better" g_j, $j = 1, \ldots, k$, and revise estimates of the mean vector and covariance matrix.

(6) Iterate until the sample variance of replicate estimates of the integral value is sufficiently small.

Teichroew (1965) provides a historical perspective on simulation techniques. For further advocacy and illustration of the use of (non-Markov-chain) Monte Carlo methods in Bayesian Statistics, see Stewart (1979, 1983, 1985, 1987), Stewart and Davis (1986), Shao (1989, 1990) and Wolpert (1991).

5.5.4 Sampling-importance-resampling

Instead of just using importance sampling to estimate integrals—and hence calculate posterior normalising constants and moments—we can also exploit the idea in order to produce simulated samples from posterior or predictive distributions. This technique is referred to by Rubin (1988) as *sampling-importance-resampling* (SIR).

We begin by taking a fresh look at Bayes' theorem from this sampling-importance-resampling perspective, shifting the focus in Bayes' theorem from densities to samples. Our account is based on Smith and Gelfand (1992).

As a first step, we note the essential duality between a sample and the distribution from which it is generated: clearly, the distribution can generate the sample; conversely, given a sample we can re-create, at least approximately, the distribution

(as a histogram, an empirical distribution function, a kernel density estimate, or whatever). In terms of densities, Bayes' theorem defines the inference process as the modification of the prior density $p(\boldsymbol{\theta})$ to form the posterior density $p(\boldsymbol{\theta} \mid \boldsymbol{x})$, through the medium of the likelihood function $p(\boldsymbol{x} \mid \boldsymbol{\theta})$. Shifting to a sampling perspective, this corresponds to the modification of a sample from $p(\boldsymbol{\theta})$ to form a sample from $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ through the medium of the likelihood function $p(\boldsymbol{x} \mid \boldsymbol{\theta})$.

To gain insight into the general problem of how a sample from one density may be modified to form a sample from a different density, consider the following. Suppose that a sample of random quantities has been generated from a density $g(\boldsymbol{\theta})$, but that what it is required is a sample from the density

$$h(\boldsymbol{\theta}) = \frac{f(\boldsymbol{\theta})}{\int f(\boldsymbol{\theta})\, d\boldsymbol{\theta}} ,$$

where only the functional form of $f(\boldsymbol{\theta})$ is specified. Given $f(\boldsymbol{\theta})$ and the sample from $g(\boldsymbol{\theta})$, how can we derive a sample from $h(\boldsymbol{\theta})$?

In cases where there exists an identifiable constant $M > 0$ such that

$$f(\boldsymbol{\theta})/g(\boldsymbol{\theta}) \leq M, \quad \text{for all } \boldsymbol{\theta} ,$$

an exact sampling procedure follows immediately from the well known rejection method for generating random quantities (see, for example, Ripley, 1987, p. 60):

(i) consider a $\boldsymbol{\theta}$ generated from $g(\boldsymbol{\theta})$;

(ii) generate u from $\text{Un}(u \mid 0, 1)$;

(iii) if $u \leq f(\boldsymbol{\theta})/Mg(\boldsymbol{\theta})$ accept $\boldsymbol{\theta}$; otherwise repeat (i)–(iii).

Any accepted $\boldsymbol{\theta}$ is then a random quantity from $h(\boldsymbol{\theta})$. Given a sample of size N for $g(\boldsymbol{\theta})$, it is immediately verified that the expected sample size from $h(\boldsymbol{\theta})$ is $M^{-1}N \int f(x)dx$.

In cases where the bound M in the above is not readily available, we can approximate samples from $h(\boldsymbol{\theta})$ as follows. Given $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ from $g(\boldsymbol{\theta})$, calculate

$$q_i = \frac{w_i}{\sum_{i=1}^N w_i} , \quad \text{where} \quad w_i = \frac{f(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)} .$$

If we now draw $\boldsymbol{\theta}^*$ from the discrete distribution $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$ having mass q_i on $\boldsymbol{\theta}_i$, then $\boldsymbol{\theta}^*$ is approximately distributed as a random quantity from $h(\boldsymbol{\theta})$. To see this, consider, for mathematical convenience, the univariate case. Then, under appropriate regularity conditions, if P describes the actual distribution of θ^*,

$$\begin{aligned} P(\theta^* \leq a) &= \sum_{i=1}^N q_i \mathbf{1}_{(-\infty,a]}(\theta_i) \\ &= \frac{n^{-1} \sum_{i=1}^N w_i \mathbf{1}_{(-\infty,a]}(\theta_i)}{n^{-1} \sum_{i=1}^N w_i} , \end{aligned}$$

so that

$$\lim_{n\to\infty} P(\theta^* \le a) = \frac{E_g\left\{\dfrac{f(\theta)}{g(\theta)}\mathbf{1}_{(-\infty,a]}(\theta)\right\}}{E_g\left\{\dfrac{f(\theta)}{g(\theta)}\right\}}$$
$$= \frac{\displaystyle\int_{-\infty}^{a} f(\theta)\,d\theta}{\displaystyle\int_{-\infty}^{\infty} f(\theta)\,d\theta} = \int_{-\infty}^{a} h(\theta)\,d\theta\ .$$

Since sampling with replacement is not ruled out, the sample size generated in this case can be as large as desired. Clearly, however, the less $h(\theta)$ resembles $g(\theta)$ the larger N will need to be if the distribution of θ^* is to be a reasonable approximation to $h(\theta)$.

With this sampling-importance-resampling procedure in mind, let us return to the prior to posterior sample process defined by Bayes' theorem. For fixed $\boldsymbol{x}$, define $f_{\boldsymbol{x}}(\boldsymbol{\theta}) = p(\boldsymbol{x}\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta})$. Then, if $\hat{\boldsymbol{\theta}}$ maximising $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$ is available, the rejection procedure given above can be applied to a sample for $p(\boldsymbol{\theta})$ to obtain a sample from $p(\boldsymbol{\theta}\,|\,\boldsymbol{x})$ by taking $g(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$, $f(\boldsymbol{\theta}) = f_{\boldsymbol{x}}(\boldsymbol{\theta})$ and $M = p(\boldsymbol{x}\,|\,\hat{\boldsymbol{\theta}})$. Bayes' theorem then takes the simple form:

For each $\boldsymbol{\theta}$ in the prior sample, accept $\boldsymbol{\theta}$ into the posterior sample with probability

$$\frac{f_{\boldsymbol{x}}(\boldsymbol{\theta})}{Mp(\boldsymbol{\theta})} = \frac{p(\boldsymbol{x}\,|\,\boldsymbol{\theta})}{p(\boldsymbol{x}\,|\,\hat{\boldsymbol{\theta}})}\ .$$

The likelihood therefore acts in an intuitive way to define the resampling probability: those $\boldsymbol{\theta}$ with high likelihoods are more likely to be represented in the posterior sample. Alternatively, if M is not readily available, we can use the approximate resampling method, which selects $\boldsymbol{\theta}_i$ into the posterior sample with probability

$$q_i = \frac{p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_i)}{\sum_{j=1}^{N} p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_j)}\ .$$

Again we note that this is proportional to the likelihood, so that the inference process via sampling proceeds in an intuitive way.

The sampling-resampling perspective outlined above opens up the possibility of novel applications of exploratory data analytic and computer graphical techniques in Bayesian statistics. We shall not pursue these ideas further here, since the topic is more properly dealt with in the subsequent volume *Bayesian Computation*. For an illustration of the method in the context of sensitivity analysis and intractable reference analysis, see Stephens and Smith (1992); for pedagogical illustration, see Albert (1993).

5.5.5 Markov Chain Monte Carlo

The key idea is very simple. Suppose that we wish to generate a sample from a posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ for $\boldsymbol{\theta} \in \Theta \subset \Re^k$ but cannot do this directly. However, suppose that we can construct a Markov chain with state space Θ, which is straightforward to simulate from, and whose equilibrium distribution is $p(\boldsymbol{\theta}|\boldsymbol{x})$. If we then run the chain for a long time, simulated values of the chain can be used as a basis for summarising features of the posterior $p(\boldsymbol{\theta}|\boldsymbol{x})$ of interest. To implement this strategy, we simply need algorithms for constructing chains with specified equilibrium distributions. For recent accounts and discussion, see, for example, Gelfand and Smith (1990), Casella and George (1992), Gelman and Rubin (1992a, 1992b), Geyer (1992), Raftery and Lewis (1992), Ritter and Tanner (1992), Roberts (1992), Tierney (1992), Besag and Green (1993), Chan (1993), Gilks *et al.* (1993) and Smith and Roberts (1993); see, also, Tanner and Wong (1987) and Tanner (1991).

Under suitable regularity conditions, asymptotic results exist which clarify the sense in which the sample output from a chain with equilibrium distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ can be used to mimic a random sample from $p(\boldsymbol{\theta}|\boldsymbol{x})$ or to estimate the expected value, with respect to $p(\boldsymbol{\theta}|\boldsymbol{x})$, of a function $g(\boldsymbol{\theta})$ of interest.

If $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^t, \ldots$ is a realisation from an appropriate chain, typically available asymptotic results as $t \to \infty$ include

$$\boldsymbol{\theta}^t \to \boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\boldsymbol{x}), \text{ in distribution}$$

and

$$\frac{1}{t}\sum_{i=1}^{t} g(\boldsymbol{\theta}^i) \to E_{\boldsymbol{\theta}|\boldsymbol{x}}\{g(\boldsymbol{\theta})\} \quad \text{almost surely.}$$

Clearly, successive $\boldsymbol{\theta}^t$ will be correlated, so that, if the first of these asymptotic results is to be exploited to mimic a random sample from $p(\boldsymbol{\theta}|\boldsymbol{x})$, suitable spacings will be required between realisations used to form the sample, or parallel independent runs of the chain might be considered. The second of the asymptotic results implies that *ergodic averaging* of a function of interest over realisations from a single run of the chain provides a consistent estimator of its expectation.

In what follows, we outline two particular forms of Markov chain scheme, which have proved particularly convenient for a range of applications in Bayesian statistics.

The Gibbs Sampling Algorithm

Suppose that $\boldsymbol{\theta}$, the vector of unknown quantities appearing in Bayes' theorem, has components $\theta_1, \ldots, \theta_k$, and that our objective is to obtain summary inferences from the joint posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x}) = p(\theta_1, \ldots, \theta_k|\boldsymbol{x})$. As we have already observed in this section, except in simple, stylised cases, this will typically lead, unavoidably, to challenging problems of numerical integration.

In fact, this apparent need for sophisticated numerical integration technology can often be avoided by recasting the problem as one of iterative sampling of random quantities from appropriate distributions to produce an appropriate Markov chain. To this end, we note that

$$p(\theta_i \mid \boldsymbol{x}, \theta_j, j \neq i), \quad i = 1, \ldots, k \ ,$$

the so-called *full conditional* densities for the individual components, given the data and specified values of all the other components of $\boldsymbol{\theta}$, are typically easily identified, as functions of θ_i, by inspection of the form of $p(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$ in any given application. Suppose then, that given an arbitrary set of starting values,

$$\theta_1^{(0)}, \ldots, \theta_k^{(0)}$$

for the unknown quantities, we implement the following iterative procedure:

draw $\theta_1^{(1)}$ from $p(\theta_1 \mid \boldsymbol{x}, \theta_2^{(0)}, \ldots, \theta_k^{(0)})$,
draw $\theta_2^{(1)}$ from $p(\theta_2 \mid \boldsymbol{x}, \theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)})$,
draw $\theta_3^{(1)}$ from $p(\theta_3 \mid \boldsymbol{x}, \theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)}, \ldots, \theta_k^{(0)})$,
$\vdots$
draw $\theta_k^{(1)}$ from $p(\theta_k \mid \boldsymbol{x}, \theta_1^{(1)}, \ldots, \theta_{k-1}^{(1)})$,
draw $\theta_1^{(2)}$ from $p(\theta_1 \mid \boldsymbol{x}, \theta_2^{(1)}, \ldots, \theta_k^{(1)})$,
$\vdots$

and so on.

Now suppose that the above procedure is continued through t iterations and is independently replicated m times so that from the current iteration we have m replicates of the sampled vector $\boldsymbol{\theta}^t = (\theta_1^{(t)}, \ldots, \theta_k^{(t)})$, where $\boldsymbol{\theta}^t$ is a realisation of a Markov chain with transition probabilities given by

$$\pi(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t+1}) = \prod_{l=1}^{k} p(\theta_l^{t+1} \mid \theta_j^t, j > l, \ \theta_j^{t+1}, j < l, \ \boldsymbol{x}).$$

Then (see, for example, Geman and Geman, 1984, Roberts and Smith, 1994), as $t \to \infty$, $(\theta_1^{(t)}, \ldots, \theta_k^{(t)})$ tends in distribution to a random vector whose joint density is $p(\boldsymbol{\theta} \mid \boldsymbol{x})$. In particular, $\theta_i^{(t)}$ tends in distribution to a random quantity whose density is $p(\theta_i \mid \boldsymbol{x})$. Thus, for large t, the replicates $(\theta_{i1}^{(t)}, \ldots, \theta_{im}^{(t)})$ are approximately a random sample from $p(\theta_i \mid \boldsymbol{x})$. It follows, by making m suitably

large, that an estimate $\hat{p}(\theta_i \mid \boldsymbol{x})$ for $p(\theta_i \mid \boldsymbol{x})$ is easily obtained, either as a kernel density estimate derived from $(\theta_{i1}^{(t)}, \ldots, \theta_{im}^{(t)})$, or from

$$\hat{p}(\theta_i \mid \boldsymbol{x}) = \frac{1}{m} \sum_{l=1}^{m} p(\theta_i \mid \boldsymbol{x},\ \theta_{jl}^{(t)}, j \neq i).$$

So far as sampling from the $p(\theta_i \mid \boldsymbol{x},\ \theta_{jl}^{(t)}, j \neq i)$ is concerned, $i = 1, \ldots, k$, either the full conditionals assume familiar forms, in which case computer routines are typically already available, or they are simple arbitrary mathematical forms, in which case general stochastic simulation techniques are available—such as envelope rejection and ratio of uniforms—which can be adapted to the specific forms (see, for example, Devroye, 1986, Ripley, 1987, Wakefield *et al.*, 1991, Gilks, 1992, Gilks and Wild, 1992, and Dellaportas and Smith, 1993). See, also, Carlin and Gelfand (1991).

The potential of this iterative scheme for routine implementation of Bayesian analysis has been demonstrated in detail for a wide variety of problems: see, for example, Gelfand and Smith (1990), Gelfand *et al.* (1990) and Gilks *et al.* (1993). We shall not provide a more extensive discussion here, since illustration of the technique in complex situations more properly belongs to the second volume of this work. We note, however, that simulation approaches are ideally suited to providing *summary inferences* (we simply report an appropriate summary of the sample), *inferences for arbitrary functions* of $\theta_1, \ldots, \theta_k$ (we simply form a sample of the appropriate function from the samples of the θ_i's) or *predictions* (for example, in an obvious notation, $p(\boldsymbol{y} \mid \boldsymbol{x}) = m^{-1} \sum_{i=1}^{m} p(\boldsymbol{y} \mid \boldsymbol{\theta}_i^{(t)})$, the average being over the $\boldsymbol{\theta}_i^{(t)}$, which have an approximate $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ distribution for large t).

The Metropolis-Hastings algorithm

This algorithm constructs a Markov chain $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^t, \ldots$ with state space Θ and equilibrium distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ by defining the transition probability from $\boldsymbol{\theta}^t = \boldsymbol{\theta}$ to the next realised state $\boldsymbol{\theta}^{t+1}$ as follows.

Let $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ denote a (for the moment arbitrary) transition probability function, such that, if $\boldsymbol{\theta}^t = \boldsymbol{\theta}$, the vector $\boldsymbol{\theta}'$ drawn from $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is considered as a proposed possible value for $\boldsymbol{\theta}^{t+1}$. However, a further randomisation now takes place. With some probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}')$, we actually accept $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}'$; otherwise, we reject the value generated from $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and set $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}$. This construction defines a Markov chain with transition probabilities given by

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\theta}') &= q(\boldsymbol{\theta}, \boldsymbol{\theta}')\, \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') \\ &\quad + I(\boldsymbol{\theta} = \boldsymbol{\theta}') \left[1 - \int q(\boldsymbol{\theta}, \boldsymbol{\theta}'')\, \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'')\, d\boldsymbol{\theta}'' \right], \end{aligned}$$

where $I(.)$ is the indicator function. If now we set

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min \left\{ 1, \quad \frac{p(\boldsymbol{\theta}' \mid \boldsymbol{x}) q(\boldsymbol{\theta}', \boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{x}) q(\boldsymbol{\theta}, \boldsymbol{\theta}')} \right\}$$

it is easy to check that $p(\boldsymbol{\theta}|\boldsymbol{x})p(\boldsymbol{\theta}, \boldsymbol{\theta}') = p(\boldsymbol{\theta}|\boldsymbol{x})p(\boldsymbol{\theta}', \boldsymbol{\theta})$, which, provided that the thus far arbitrary $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is chosen to be irreducible and aperiodic on a suitable state space, is a sufficient condition for $p(\boldsymbol{\theta}|\boldsymbol{x})$ to be the equilibrium distribution of the constructed chain.

This general algorithm is due to Hastings (1970); see, also, Metropolis *et al.* (1953), Peskun (1973), Tierney (1992), Besag and Green (1993), Roberts and Smith (1994) and Smith and Roberts (1993). It is important to note that the (equilibrium) distribution of interest, $p(\boldsymbol{\theta}|\boldsymbol{x})$, only enters $p(\boldsymbol{\theta}, \boldsymbol{\theta}')$ through the ratio $p(\boldsymbol{\theta}'|\boldsymbol{x})/p(\boldsymbol{\theta}|\boldsymbol{x})$. This is quite crucial since it means that knowledge of the distribution up to proportionality (given by the likelihood multiplied by the prior) is sufficient for implementation.

5.6 DISCUSSION AND FURTHER REFERENCES

5.6.1 An Historical Footnote

Blackwell (1988) gave a very elegant demonstration of the way in which a simple finite additivity argument can be used to give powerful insight into the relation between frequency and belief probability. The calculation involved has added interest in that—according to Stigler (1982)—it might very well have been made by Bayes himself.

The argument goes as follows. Suppose that 0–1 observables $x_1, \ldots, x_{n+1}$ are finitely exchangeable. We observe $\boldsymbol{x} = (x_1, \ldots, x_n)$ and wish to evaluate

$$\frac{P(x_{n+1} = 1|\boldsymbol{x})}{P(x_{n+1} = 0|\boldsymbol{x})} .$$

Writing $s = x_1 + \cdots + x_n$, $p(t) = P(x_1 + \cdots + x_{n+1} = t)$, this ratio, by virtue of exchangeability, is easily seen to be equal to

$$\frac{p(s+1) \Big/ \binom{n+1}{s+1}}{p(s) \Big/ \binom{n+1}{s}} = \frac{p(s+1)}{p(s)} \cdot \frac{s+1}{n-s+1} \approx \frac{s}{n-s} ,$$

if $p(s) \simeq p(s+1)$ and s and $n - s$ are not too small.

This can be interpreted as follows. If, before observing $\boldsymbol{x}$, we considered s and $s+1$ to be about equally plausible as values for $x_1 + \cdots + x_{n+1}$, the resulting posterior odds for $x_{n+1} = 1$ will be essentially the frequency odds based on the first n trials.

Inverting the argument, we see that if one wants to have this "convergence" of beliefs and frequencies it is necessary that $p(s) \approx p(s+1)$. But what does this entail?

Reverting to an infinite exchangeability assumption, and hence the familiar binomial framework, suppose we require that $p(\theta)$ be chosen such that

$$p(s) = \int_0^1 \binom{n}{s} \theta^s (1-\theta)^{n-s} p(\theta) d\theta$$

does not depend on s. An easy calculation shows that this is satisfied if $p(\theta)$ is taken to be uniform on $(0,1)$—the so-called *Bayes (or Bayes-Laplace) Postulate*.

Stigler (1982) has argued that an argument like the above could have been Bayes' motivation for the adoption of this uniform prior.

5.6.2 Prior Ignorance

To many attracted to the formalism of the Bayesian inferential paradigm, the idea of a *non-informative* prior distribution, representing "ignorance" and "letting the data speak for themselves" has proved extremely seductive, often being regarded as synonymous with providing *objective* inferences. It will be clear from the general subjective perspective we have maintained throughout this volume, that we regard this search for "objectivity" to be misguided. However, it will also be clear from our detailed development in Section 5.4 that we recognise the rather special nature and role of the concept of a "minimally informative" prior specification —appropriately defined! In any case, the considerable body of conceptual and theoretical literature devoted to identifying "appropriate" procedures for formulating prior representations of "ignorance" constitutes a fascinating chapter in the history of Bayesian Statistics. In this section we shall provide an overview of some of the main directions followed in this search for a Bayesian "Holy Grail".

In the early works by Bayes (1763) and Laplace (1814/1952), the definition of a non-informative prior is based on what has now become known as the principle of *insufficient reason*, or the Bayes-Laplace postulate (see Section 5.6.1). According to this principle, in the absence of evidence to the contrary, all possibilities should have the same initial probability. This is closely related to the so-called Laplace-Bertrand paradox; see Jaynes (1971) for an interesting Bayesian resolution.

In particular, if an unknown quantity, ϕ, say, can only take a finite number of values, M, say, the non-informative prior suggested by the principle is the discrete uniform distribution $p(\phi) = \{1/M, \ldots, 1/M\}$. This may, at first sight, seem

intuitively reasonable, but Example 5.16 showed that even in simple, finite, discrete cases care can be required in appropriately defining the unknown *quantity of interest*. Moreover, in countably infinite, discrete cases the uniform (now *improper*) prior is known to produce unappealing results. Jeffreys (1939/1961, p. 238) suggested, for the case of the integers, the prior

$$\pi(n) \propto n^{-1}, \quad n = 1, 2, \ldots .$$

More recently, Rissanen (1983) used a coding theory argument to motivate the prior

$$\pi(n) \propto \frac{1}{n} \times \frac{1}{\log n} \times \frac{1}{\log \log n} \times \ldots, \quad n = 1, 2, \ldots .$$

However, as indicated in Example 5.23, embedding the discrete problem within a continuous framework and subsequently discretising the resulting reference prior for the continuous case may produce better results.

If the space, Φ, of ϕ values is a continuum (say, the real line) the principle of insufficient reason has been interpreted as requiring a uniform distribution over Φ. However, a uniform distribution for ϕ implies a non-uniform distribution for any non-linear monotone transformation of ϕ and thus the Bayes-Laplace postulate is inconsistent in the sense that, intuitively, "ignorance about ϕ" should surely imply "equal ignorance" about a one-to-one transformation of ϕ. Specifically, if some procedure yields $p(\phi)$ as a non-informative prior for ϕ and the same procedure yields $p(\zeta)$ as a non-informative prior for a one-to-one transformation $\zeta = \zeta(\phi)$ of ϕ, consistency would seem to demand that $p(\zeta)d\zeta = p(\phi)d\phi$; thus, a procedure for obtaining the "ignorance" prior should presumably be invariant under one-to-one reparametrisation.

Based on these invariance considerations, Jeffreys (1946) proposed as a non-informative prior, with respect to an experiment $e = \{X, \phi, p(x \mid \phi)\}$, involving a parametric model which depends on a single parameter ϕ, the (often improper) density

$$\pi(\phi) \propto h(\phi)^{1/2},$$

where

$$h(\phi) = -\int_X p(x \mid \phi) \frac{\partial^2}{\partial \phi^2} \log p(x \mid \phi) \, dx \, .$$

In effect, Jeffreys noted that the logarithmic divergence locally behaves like the square of a distance, determined by a Riemannian metric, whose natural length element is $h(\phi)^{1/2}$, and that natural length elements of Riemannian metrics are invariant to reparametrisation. In an illuminating paper, Kass (1989) elaborated on this *geometrical* interpretation by arguing that, more generally, natural volume elements generate "uniform" measures on manifolds, in the sense that equal mass

is assigned to regions of equal volume, the essential property that makes Lebesgue measure intuitively appealing.

In his work, Jeffreys explored the implications of such a non-informative prior for a large number of inference problems. He found that his *rule* (by definition restricted to a continuous parameter) works well in the one-dimensional case, but can lead to unappealing results (Jeffreys, 1939/1961, p. 182) when one tries to extend it to multiparameter situations.

The procedure proposed by Jeffreys' preferred rule was rather *ad hoc*, in that there are many other procedures (some of which he described) which exhibit the required type of invariance. His intuition as to what is required, however, was rather good. Jeffreys' solution for the one-dimensional continuous case has been widely adopted, and a number of alternative justifications of the procedure have been provided.

Perks (1947) used an argument based on the asymptotic size of confidence regions to propose a non-informative prior of the form

$$\pi(\phi) \propto s(\phi)^{-1}$$

where $s(\phi)$ is the asymptotic standard deviation of the maximum likelihood estimate of ϕ. Under regularity conditions which imply asymptotic normality, this turns out to be equivalent to Jeffreys' rule.

Lindley (1961b) argued that, in practice, one can always replace a continuous range of ϕ by discrete values over a grid whose mesh size, $\delta(\phi)$, say, describes the precision of the measuring process, and that a possible operational interpretation of "ignorance" is a probability distribution which assigns equal probability to all points of this grid. In the continuous case, this implies a prior proportional to $\delta(\phi)^{-1}$. To determine $\delta(\phi)$ in the context of an experiment $e = \{X, \phi, p(x \mid \phi)\}$, Lindley showed that if the quantity can only take the values ϕ or $\phi + \delta(\phi)$, the amount of information that e may be expected to provide about ϕ, if $p(\phi) = p(\phi + \delta(\phi)) = \frac{1}{2}$, is $2\delta^2(\phi)h(\phi)$. This expected information will be independent of ϕ if $\delta(\phi) \propto h(\phi)^{-1/2}$, thus defining an appropriate mesh; arguing as before, this suggests Jeffreys' prior $\pi(\phi) \propto h(\theta)^{1/2}$. Akaike (1978a) used a related argument to justify Jeffreys' prior as "locally impartial".

Welch and Peers (1963) and Welch (1965) discussed conditions under which there is formal mathematical equivalence between one-dimensional Bayesian credible regions and corresponding frequentist confidence intervals. They showed that, under suitable regularity assumptions, one-sided intervals asymptotically coincide if the prior used for the Bayesian analysis is Jeffreys' prior. Peers (1965) later showed that the argument does not extend to several dimensions. Hartigan (1966b) and Peers (1968) discuss two-sided intervals. Tibshirani (1989), Mukerjee and Dey (1993) and Nicolau (1993) extend the analysis to the case where there are nuisance parameters.

Hartigan (1965) reported that the prior density which minimises the bias of the estimator d of ϕ associated with the loss function $l(d, \phi)$ is

$$\pi(\phi) = h(\phi) \left[\frac{\partial^2}{\partial d^2} l(d, \phi) \right]^{-1/2} \Bigg|_{d=\phi} .$$

If, in particular, one uses the discrepancy measure

$$l(d, \phi) = \int p(x \mid \phi) \log \frac{p(x \mid \phi)}{p(x \mid d)} \, dx$$

as a natural loss function (see Definition 3.15), this implies that $\pi(\phi) = h(\phi)^{1/2}$, which is, again, Jeffreys' prior.

Good (1969) derived Jeffreys' prior as the "least favourable" initial distribution with respect to a logarithmic scoring rule, in the sense that it minimises the expected score from reporting the true distribution. Since the logarithmic score is proper, and hence is maximised by reporting the true distribution, Jeffreys' prior may technically be described, under suitable regularity conditions, as a minimax solution to the problem of scientific reporting when the utility function is the logarithmic score function. Kashyap (1971) provided a similar, more detailed argument; an axiom system is used to justify the use of an information measure as a payoff function and Jeffreys' prior is shown to be a minimax solution in a —two person— zero sum game, where the statistician chooses the "non-informative" prior and nature chooses the "true" prior.

Hartigan (1971, 1983, Chapter 5) defines a similarity measure for events E, F to be $P(E \cap F)/P(E)P(F)$ and shows that Jeffreys' prior ensures, asymptotically, constant similarily for current and future observations.

Following Jeffreys (1955), Box and Tiao (1973, Section 1.3) argued for selecting a prior by convention to be used as a *standard of reference*. They suggested that the principle of insufficient reason may be sensible in location problems, and proposed as a conventional prior $\pi(\phi)$ for a model parameter ϕ that $\pi(\phi)$ which implies a uniform prior

$$\pi(\zeta) = \pi(\phi) \left| \frac{\partial \zeta}{d\phi} \right|^{-1} \propto c$$

for a function $\zeta = \zeta(\phi)$ such that $p(x \mid \zeta)$ is, at least approximately, a location parameter family; that is, such that, for some functions g and f,

$$p(x \mid \phi) \sim g \left[\zeta(\phi) - f(x) \right] .$$

Using standard asymptotic theory, they showed that, under suitable regularity conditions and for large samples, this will happen if

$$\zeta(\phi) = \int^{\phi} h(\phi)^{1/2} d\phi \, ,$$

i.e., if the non-informative prior is Jeffreys' prior. For a recent reconsideration and elaboration of these ideas, see Kass (1990), who extends the analysis by conditioning on an ancillary statistic.

Unfortunately, although many of the arguments summarised above generalise to the multiparameter continuous case, leading to the so-called multivariate Jeffreys' rule

$$\pi(\boldsymbol{\theta}) \propto |\, \boldsymbol{H}(\boldsymbol{\theta}) \,|^{1/2},$$

where

$$[\boldsymbol{H}(\boldsymbol{\theta})]_{ij} = -\int p(x \,|\, \boldsymbol{\theta}) \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(x \,|\, \boldsymbol{\theta})\, dx$$

is Fisher's *information matrix*, the results thus obtained typically have intuitively unappealing implications. An example of this, pointed out by Jeffreys himself (Jeffreys, 1939/1961 p. 182) is provided by the simple location-scale problem, where the multivariate rule leads to $\pi(\theta, \sigma) \propto \sigma^{-2}$, where θ is the location and σ the scale parameter. See, also, Stein (1962).

Example 5.25. ***(Univariate normal model)***. Let $\{x_1, \ldots, x_n\}$ be a random sample from $N(x \,|\, \mu, \lambda)$, and consider $\sigma = \lambda^{-1/2}$, the (unknown) standard deviation. In the case of known mean, $\mu = 0$, say, the appropriate (univariate) Jeffreys' prior is $\pi(\sigma) \propto \sigma^{-1}$ and the posterior distribution of σ would be such that $[\Sigma_{i=1}^n x_i^2]/\sigma^2$ is χ_n^2. In the case of unknown mean, if we used the multivariate Jeffreys' prior $\pi(\mu, \sigma) \propto \sigma^{-2}$ the posterior distribution of σ would be such that $[\Sigma_{i=1}^n (x_i - \overline{x})^2]/\sigma^2$ is, again, χ_n^2. This is widely recognised as unacceptable, in that one does not lose any degrees of freedom even though one has lost the knowledge that $\mu = 0$, and conflicts with the use of the widely adopted reference prior $\pi(\mu, \sigma) = \sigma^{-1}$ (see Example 5.17 in Section 5.4), which implies that $[\Sigma_{i=1}^n (x_i - \overline{x})^2]/\sigma^2$ is χ_{n-1}^2.

The kind of problem exemplified above led Jeffreys to the *ad hoc* recommendation, widely adopted in the literature, of independent a priori treatment of location and scale parameters, applying his rule separately to each of the two subgroups of parameters, and then multiplying the resulting forms together to arrive at the overall prior specification. For an illustration of this, see Geisser and Cornfield (1963): for an elaboration of the idea, see Zellner (1986a).

At this point, one may wonder just what has become of the intuition motivating the arguments outlined above. Unfortunately, although the implied information limits are mathematically well-defined in one dimension, in higher dimensions the forms obtained may depend on the path followed to obtain the limit. Similar problems arise with other intuitively appealing desiderata. For example, the Box and Tiao suggestion of a uniform prior following transformation to a parametrisation ensuring data translation generalises, in the multiparameter setting, to the requirement of uniformity following a transformation which ensures that credible regions

are of the same size. The problem, of course, is that, in several dimensions, such regions can be of the same size but very different in form.

Jeffreys' original requirement of invariance under reparametrisation remains perhaps the most intuitively convincing. If this is conceded, it follows that, whatever their apparent motivating intuition, approaches which do not have this property should be regarded as unsatisfactory. Such approaches include the use of limiting forms of conjugate priors, as in Haldane (1948), Novick and Hall (1965), Novick (1969), DeGroot (1970, Chapter 10) and Piccinato (1973, 1977), a predictivistic version of the principle of insufficient reason, Geisser (1984), and different forms of information-theoretical arguments, such as those put forward by Zellner (1977, 1991), Geisser (1979) and Torgesen (1981).

Maximising the expected information (as opposed to maximising the expected *missing* information) gives invariant, but unappealing results, producing priors that can have finite support (Berger *et al.*, 1989). Other information-based suggestions are those of Eaton (1982), Spall and Hill (1990) and Rodríguez (1991).

Partially satisfactory results have nevertheless been obtained in multiparameter problems where the parameter space can be considered as a group of transformations of the sample space. Invariance considerations within such a group suggest the use of *relatively invariant* (Hartigan, 1964) priors like the Haar measures. This idea was pioneered by Barnard (1952). Stone (1965) recognised that, in an appropriate sense, it should be possible to approximate the results obtained using a non-informative prior by those obtained using a convenient sequence of proper priors. He went on to show that, if a group structure is present, the corresponding *right* Haar measure is the only prior for which such a desirable convergence is obtained. It is reassuring that, in those one-dimensional problems for which a group of transformations does exist, the right Haar measures coincides with the relevant Jeffreys' prior. For some undesirable consequences of the *left* Haar measure see Bernardo (1978b). Further developments involving Haar measures are provided by Zidek (1969), Villegas (1969, 1971, 1977a, 1977b, 1981), Stone (1970), Florens (1978, 1982), Chang and Villegas (1986) and Chang and Eaves (1990). Dawid (1983b) provides an excellent review of work up to the early 1980's. However, a large group of interesting models do not have any group structure, so that these arguments cannot produce general solutions.

Even when the parameter space may be considered as a group of transformations there is no definitive answer. In such situations, the right Haar measures are the obvious choices and yet even these are open to criticism.

Example 5.26. ***(Standardised mean)***. Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from a normal distribution $N(x \mid \mu, \lambda)$. The standard prior recommended by group invariance arguments is $\pi(\mu, \sigma) = \sigma^{-1}$ where $\lambda = \sigma^{-2}$. Although this gives adequate results if one wants to make inferences about either μ or σ, it is quite unsatisfactory if inferences about the standardised mean $\phi = \mu/\sigma$ are required. Stone and Dawid (1972) show that the posterior

distribution of ϕ obtained from such a prior depends on the data through the statistic

$$t = \frac{\sum_{i=1}^{n} x_i}{(\sum_{i=1}^{n} x_i^2)^{1/2}} ,$$

whose sampling distribution,

$$\begin{aligned} p(t \,|\, \mu, \sigma) &= p(t \,|\, \phi) \\ &= e^{-n\phi^2/2} \left\{ 1 - \frac{t^2}{n} \right\}^{(n-3)/2} \int_0^\infty \omega^{n-1} \exp\left\{ -\frac{\omega^2}{2} + t\phi\omega \right\} d\omega, \end{aligned}$$

only depends on ϕ. One would, therefore, expect to be able to "match" the original inferences about ϕ by the use of $p(t \,|\, \phi)$ together with some appropriate prior for ϕ. However, no such prior exists.

On the other hand, the reference prior relative to the ordered partition (ϕ, σ) is (see Example 5.18)

$$\pi(\phi, \sigma) = (2 + \phi^2)^{-1/2} \sigma^{-1}$$

and the corresponding posterior distribution for ϕ is

$$\pi(\phi \,|\, \boldsymbol{x}) \propto (2 + \phi^2)^{-1/2} \left[e^{-n\phi^2/2} \int_0^\infty \omega^{n-1} \exp\Big\{ -\frac{\omega^2}{2} + t\phi\omega \Big\} d\omega \right] .$$

We observe that the factor in square brackets is proportional to $p(t \,|\, \phi)$ and thus the inconsistency disappears.

This type of *marginalisation paradox*, further explored by Dawid, Stone and Zidek (1973), appears in a large number of multivariate problems and makes it difficult to believe that, for any given model, a *single* prior may be usefully regarded as "universally" non-informative. Jaynes (1980) disagrees.

An acceptable general theory for non-informative priors should be able to provide consistent answers to the same inference problem whenever this is posed in different, but equivalent forms. Although this idea has failed to produce a constructive procedure for deriving priors, it may be used to discard those methods which fail to satisfy this rather intuitive requirement.

Example 5.27. ***(Correlation coefficient)***. Let $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a random sample from a bivariate normal distribution, and suppose that inferences about the correlation coefficient ρ are required. It may be shown that if the prior is of the form

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \pi(\rho) \sigma_1^{-a} \sigma_2^{-a},$$

which includes all proposed "non-informative" priors for this model that we are aware of, then the posterior distribution of ρ is given by

$$\begin{aligned} \pi\left(\rho \,|\, \boldsymbol{x}, \boldsymbol{y}\right) &= \pi(\rho \,|\, r) \\ &= \frac{\pi(\rho)(1-\rho^2)^{(n+2a-3)/2}}{(1-\rho r)^{n+a-(5/2)}} \; F\left(\tfrac{1}{2}, \tfrac{1}{2}, n + a - \tfrac{3}{2}, \frac{1+\rho r}{2} \right) , \end{aligned}$$

where

$$r = \frac{\sum_i x_i y_i - n\overline{x}\,\overline{y}}{[\sum_i (x_i - \overline{x})^2]^{1/2}[\sum_i (y_i - \overline{y})^2]^{1/2}}$$

is the sample correlation coefficient, and F is the hypergeometric function. This posterior distribution only depends on the data through the sample correlation coefficient r; thus, with this form of prior, r is sufficient. On the other hand, the sampling distribution of r is

$$\begin{aligned} p(r \,|\, \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) &= p(r \,|\, \rho) \\ &= \frac{(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}}{(1-\rho r)^{n-3/2}} F\left(\tfrac{1}{2}, \tfrac{1}{2}, n - \tfrac{1}{2}, \frac{1+\rho r}{2}\right). \end{aligned}$$

Moreover, using the transformations $\delta = \tanh^{-1}\rho$ and $t = \tanh^{-1} r$, Jeffreys' prior for this univariate model is found to be $\pi(\rho) \propto (1-\rho^2)^{-1}$ (see Lindley, 1965, pp. 215–219).

Hence one would expect to be able to match, using this reduced model, the posterior distribution $\pi(\rho \,|\, r)$ given previously, so that

$$\pi(\rho \,|\, r) \propto p(r \,|\, \rho)(1-\rho^2)^{-1}.$$

Comparison between $\pi(\rho \,|\, r)$ and $p(r \,|\, \rho)$ shows that this is possible if and only if $a = 1$, and $\pi(\rho) = (1-\rho^2)^{-1}$. Hence, to avoid inconsistency the joint reference prior must be of the form

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = (1-\rho^2)^{-1}\sigma_1^{-1}\sigma_2^{-1},$$

which is precisely (see Example 5.22, p. 337) the reference prior relative to the natural order, $\{\rho, \mu_1, \mu_2, \sigma_1, \sigma_2\}$.

However, it is easily checked that Jeffreys' multivariate prior is

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = (1-\rho^2)^{-3/2}\sigma_1^{-2}\sigma_2^{-2}$$

and that the "two-step" Jeffreys' multivariate prior which separates the location and scale parameters is

$$\pi(\mu, \mu_2)\pi(\sigma_1, \sigma_2, \rho) = (1-\rho^2)^{-3/2}\sigma_1^{-1}\sigma_2^{-1}.$$

For further detailed discussion of this example, see Bayarri (1981).

Once again, this example suggests that different non-informative priors may be appropriate *depending on the particular function of interest* or, more generally, on the ordering of the parameters.

Although marginalisation paradoxes disappear when one uses proper priors, to use proper approximations to non-informative priors as an approximate description of "ignorance" does not solve the problem either.

Example 5.28. ***(Stein's paradox)***. Let $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a random sample from a multivariate normal distribution $N_k(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{I}_k\}$. Let $\overline{x}_i$ be the mean of the n observations from coordinate i and let $t = \sum_i \overline{x}_i^2$. The universally recommended "non-informative" prior for this model is $\pi(\mu_1, \ldots, \mu_k) = 1$, which may be approximated by the proper density

$$\pi(\mu_1, \ldots, \mu_k) = \prod_{i=1}^{m} N(\mu_i \mid 0, \lambda),$$

where λ is very small. However, if inferences about $\phi = \sum_i \mu_i^2$ are desired, the use of this prior overwhelms, for large k, what the data have to say about ϕ. Indeed, with such a prior the posterior distribution of $n\phi$ is a non-central χ^2 distribution with k degrees of freedom and non-centrality parameter nt, so that

$$E[\phi \mid \boldsymbol{x}] = t + \frac{k}{n}, \quad V[\phi \mid x] = \frac{2}{n}\left[2t + \frac{k}{n}\right] ,$$

while the sampling distribution of nt is a non-central χ^2 distribution with k degrees of freedom and parameter $n\theta$ so that $E[t \mid \phi] = \phi + k/n$. Thus, with, say, $k = 100$, $n = 1$ and $t = 200$, we have $E[\phi \mid \boldsymbol{x}] \approx 300$, $V[\phi \mid x] \approx 32^2$, whereas the unbiased estimator based on the sampling distribution gives $\hat{\phi} = t - k \approx 100$.

However, the asymptotic posterior distribution of ϕ is $N(\phi \mid \hat{\phi}, (4\hat{\phi})^{-1})$ and hence, by Proposition 5.2, the reference posterior for ϕ relative to $p(t \mid \phi)$ is

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

whose mode is close to $\hat{\phi}$. It may be shown that this is also the posterior distribution of ϕ derived from the reference prior relative to the ordered partition $\{\phi, \omega_1, \ldots, \omega_{k-1}\}$, obtained by reparametrising to polar coordinates in the full model. For further details, see Stein (1959), Efron (1973), Bernardo (1979b) and Ferrándiz (1982).

Naïve use of apparently "non-informative" prior distributions can lead to posterior distributions whose corresponding credible regions have untenable coverage probabilities, in the sense that, for some region C, the corresponding posterior probabilities $P(C \mid \boldsymbol{z})$ may be completely different from the conditional values $P(C \mid \theta)$ for almost all θ values.

Such a phenomenon is often referred to as *strong inconsistency* (see, for example, Stone, 1976). However, by carefully distinguishing between parameters of interest and nuisance parameters, reference analysis avoids this type of inconsistency. An illuminating example is provided by the reanalysis by Bernardo (1979b, reply to the discussion) of Stone's (1976) *Flatland* example. For further discussion of strong inconsistency and related topics, see Appendix B, Section 3.2.

Jaynes (1968) introduced a more general formulation of the problem. He allowed for the existence of a certain amount of initial "objective" information and then tried to determine a prior which reflected this initial information, but nothing

else (see, also, Csiszár, 1985). Jaynes considered the entropy of a distribution to be the appropriate measure of uncertainty subject to any "objective" information one might have. If no such information exists and ϕ can only take a finite number of values, Jaynes' *maximum entropy* solution reduces to the Bayes-Laplace postulate. His arguments are quite convincing in the finite case; however, if ϕ is continuous, the non-invariant entropy functional, $H\{p(\phi)\} = -\int p(\phi)\log p(\phi)d\phi$, no longer has a sensible interpretation in terms of uncertainty. Jaynes' solution is to introduce a "reference" density $\pi(\phi)$ in order to define an "invariantised" entropy,

$$-\int p(\phi)\log\frac{p(\phi)}{\pi(\phi)}d\phi,$$

and to use the prior which maximises this expression, subject, again, to any initial "objective" information one might have. Unfortunately, $\pi(\phi)$ must itself be a representation of ignorance about ϕ so that no progress has been made. If a convenient group of transformations is present, Jaynes suggests invariance arguments to select the reference density. However, no general procedure is proposed.

Context-specific "non-informative" Bayesian analyses have been produced for specific classes of problems, with no attempt to provide a general theory. These include dynamic models (Pole and West, 1989) and finite population survey sampling (Meeden and Vardeman, 1991).

The quest for non-informative priors could be summarised as follows.

(i) In the finite case, Jaynes' principle of maximising the entropy is convincing, but cannot be extended to the continuous case.

(ii) In one-dimensional continuous regular problems, Jeffreys' prior is appropriate.

(iii) The infinite discrete case can often be handled by suitably embedding the problem within a continuous framework.

(iv) In continuous multiparameter situations there is no hope for a single, unique, "non-informative prior", appropriate for all the inference problems within a given model. To avoid having the prior dominating the posterior for *some* function ϕ of interest, the prior has to depend not only on the model but also on the parameter of interest or, more generally, on some notion of the order of importance of the parameters.

The reference prior theory introduced in Bernardo (1979b) and developed in detail in Section 5.4 avoids most of the problems encountered with other proposals. It reduces to Jaynes' form in the finite case and to Jeffreys' form in one-dimensional regular continuous problems, avoiding marginalisation paradoxes by insisting that the reference prior be tailored to the parameter of interest. However, subsequent work by Berger and Bernardo (1989) has shown that the heuristic arguments in Bernardo (1979b) can be misleading in complicated situations, thus necessitating more precise definitions. Moreover, Berger and Bernardo (1992a, 1992b, 1992c)

showed that the partition into parameters of interest and nuisance parameter may not go far enough and that reference priors should be viewed relative to a given ordering—or, more generally, a given ordered grouping—of the parameters. This approach was described in detail in Section 5.4. Ye (1993) derives reference priors for sequential experiments.

A completely different objection to such approaches to non-informative priors lies in the fact that, for continuous parameters, they depend on the likelihood function. This is recognised to be potentially inconsistent with a personal interpretation of probability. For many subjectivists, the initial density $p(\phi)$ is a description of the opinions held about ϕ, independent of the experiment performed;

> why should one's knowledge, or ignorance, of a quantity depend on the experiment being used to determine it? Lindley (1972, p. 71).

In many situations, we would accept this argument. However, as we argued earlier, priors which reflect knowledge of the experiment can sometimes be genuinely appropriate in Bayesian inference, and may also have a useful role to play (see, for example, the discussion of stopping rules in Section 5.1.4) as technical devices to produce *reference* posteriors. Posteriors obtained from actual prior opinions could then be compared with those derived from a reference analysis in order to assess the relative importance of the initial opinions on the final inference.

> In general we feel that it is sensible to choose a non-informative prior which expresses ignorance *relative* to information which can be supplied by a particular experiment. If the experiment is changed, then the expression of relative ignorance can be expected to change correspondingly. (Box and Tiao, 1973, p. 46).

Finally, "non-informative" distributions have sometimes been criticised on the grounds that they are typically improper and may lead, for instance, to inadmissible estimates (see, e.g. Stein, 1956). However, sensible "non-informative" priors may be seen to be, in an appropriate sense, limits of proper priors (Stone, 1963, 1965, 1970; Stein, 1965; Akaike, 1980a). Regarded as a "baseline" for admissible inferences, posterior distributions derived from "non-informative" priors need not be themselves admissible, but only arbitrarily close to admissible posteriors.

However, there can be no final word on this topic! For example, recent work by Eaton (1992), Clarke and Wasserman (1993), George and McCulloch (1993b) and Ye (1993) seems to open up new perspectives and directions.

5.6.3 Robustness

In Section 4.8.3, we noted that some aspects of model specification, either for the parametric model or the prior distribution components, can seem arbitrary, and cited

as an example the case of the choice between normal and Student-t distributions as a parametric model component to represent departures of observables from their conditional expected values. In this section, we shall provide some discussion of how insight and guidance into appropriate choices might be obtained.

We begin our discussion with a simple, direct approach to examining the ways in which a posterior density for a parameter depends on the choices of parametric model or prior distribution components. Consider, for simplicity, a single observable $x \in \Re$ having a parametric density $p(x|\theta)$, with $\theta \in \Re$ having prior density $p(\theta)$. The mechanism of Bayes' theorem,

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} ,$$

involves multiplication of the two model components, $p(x|\theta)$, $p(\theta)$, followed by normalisation, a somewhat "opaque" operation from the point of view of comparing specifications of $p(x|\theta)$ or $p(\theta)$ on a "what if?" basis.

However, suppose we take logarithms in Bayes' theorem and subsequently differentiate with respect to θ. This now results in a *linear* form

$$\frac{\partial}{\partial\theta}\log p(\theta|x) = \frac{\partial}{\partial\theta}\log p(x|\theta) + \frac{\partial}{\partial\theta}\log p(\theta).$$

The first term on the right-hand side is (apart from a sign change) a quantity known in classical statistics as the *efficient score function* (see, for example, Cox and Hinkley, 1974). On the linear scale, this is the quantity which transforms the prior into the posterior and hence opens the way, perhaps, to insight into the effect of a particular choice of $p(x|\theta)$ given the form of $p(\theta)$. See, for example, Ramsey and Novick (1980) and Smith (1983). Conversely, examination of the second term on the right-hand side for given $p(x|\theta)$ may provide insight into the implications of the mathematical specification of the prior.

For convenience of exposition—and perhaps because the prior component is often felt to be the less secure element in the model specification—we shall focus the following discussion on the sensitivity of characteristics of $p(\theta \mid x)$ to the choice of $p(\theta)$. Similar ideas apply to the choice of $p(x \mid \theta)$.

With x denoting the mean of n independent observables from a normal distribution with mean θ and precision λ, we shall illustrate these ideas by considering the form of the posterior mean for θ when $p(x|\theta) = \mathrm{N}(x|\theta, n\lambda)$ and $p(\theta)$ is of "arbitrary" form.

Defining

$$p(x) = \int p(x|\theta)p(\theta)d\theta,$$
$$s(x) = \frac{\partial \log p(x)}{\partial x} ,$$

it can be shown (see, for example, Pericchi and Smith, 1992) that

$$E(\theta|x) = x - n^{-1}\lambda^{-1}s(x).$$

Suppose we carry out a "what if?" analysis by asking how the behaviour of the posterior mean depends on the mathematical form adopted for $p(\theta)$.

What if we take $p(\theta)$ to be *normal*? With $p(\theta) = \mathrm{N}(\theta|\mu, \lambda_0)$, the reader can easily verify that in this case $p(x)$ will be normal, and hence $s(x)$ will be a linear combination of x and the prior mean. The formula given for $E(\theta|x)$ therefore reproduces the weighted average of sample and prior means that we obtained in Section 5.2, so that

$$E(\theta|x) = (n\lambda + \lambda_0)^{-1}(n\lambda x + \lambda_0\mu).$$

What if we take $p(\theta)$ to be *Student-t*? With $p(\theta) = \mathrm{St}(\theta|\mu, \lambda_0, \alpha)$ the exact treatment of $p(x)$ and $s(x)$ becomes intractable. However, detailed analysis (Pericchi and Smith, 1992) provides the approximation

$$E(\theta|x) = x - \frac{(\alpha+1)(x-\mu)}{n\lambda[\alpha\lambda_0^{-1} + (x-\mu)^2]}\,.$$

What if we take $p(\theta)$ to be *double-exponential*? In this case,

$$p(\theta) = \frac{1}{\nu\sqrt{2}} \exp\left(-\frac{\sqrt{2}}{\nu}|\theta - \mu|\right),$$

for some $\nu > 0$, $\mu \in \Re$ and the evaluation of $p(x)$ and $s(x)$ is possible, but tedious. After some algebra—see Pericchi and Smith (1992)—it can be shown that, if $b = n^{-1}\nu^{-1}\lambda^{-1}\sqrt{2}$,

$$E(\theta|x) = w(x)(x+b) + [1 - w(x)](x-b),$$

where $w(x)$ is a weight function, $0 \le w(x) \le 1$, so that

$$x - b \le E(\theta|x) \le x + b.$$

Examination of the three forms for $E(\theta|x)$ reveals striking qualitative differences. In the case of the normal, the posterior mean is unbounded in $x - \mu$, the departure of the observed mean from the prior mean. In the case of the Student-t, we see that for very small $x - \mu$ the posterior mean is approximately linear in $x - \mu$, like the normal, whereas for $x - \mu$ very large the posterior mean approaches x. In the case of the double-exponential, the posterior mean is bounded, with limits equal to x plus or minus a constant.

Consideration of these qualitative differences might provide guidance regarding an otherwise arbitrary choice if, for example, one knew how one would like the Bayesian learning mechanism to react to an "outlying" x, which was far from μ. See Smith (1983) and Pericchi *et al.* (1993) for further discussion and elaboration. See Jeffreys (1939/1961) for seminal ideas relating to the effect of the tail-weight of the distribution of the parametric model on posterior inferences. Other relevant references include Masreliez (1975), O'Hagan (1979, 1981, 1988b), West (1981), Maín (1988), Polson (1991), Gordon and Smith (1993) and O'Hagan and Le (1994).

The approach illustrated above is well-suited to probing qualitative differences in the posterior by considering, individually, the effects of a small number of potential alternative choices of model component (parametric model or prior distribution).

Suppose, instead, that someone has in mind a specific candidate component specification, p_0, say, but is all too aware that aspects of the specification have involved somewhat arbitrary choices. It is then natural to be concerned about whether posterior conclusions might be highly sensitive to the particular specification p_0, viewed in the context of alternative choices in an appropriately defined *neighbourhood* of p_0.

In the case of specifying a parametric component p_0—for example an "error" model for differences between observables and their (conditional) expected values—such concern might be motivated by definite knowledge of symmetry and unimodality, but an awareness of the arbitrariness of choosing a conventional distributional form such as normality. Here, a suitable neighbourhood might be formed by taking p_0 to be normal and forming a class of distributions whose tail-weights deviate (lighter and heavier) from normal: see, for example, the seminal papers of Box and Tiao (1962, 1964).

In the case of specifying a prior component p_0, such concern might be motivated by the fact that elicitation of prior opinion has only partly determined the specification (for example, by identifying a few quantiles), with considerable remaining arbitrariness in "filling out" the rest of the distribution. Here, a suitable neighbourhood of p_0 might consist of a class of priors all having the specified quantiles but with other characteristics varying: see, for example, O'Hagan and Berger (1988).

From a mathematical perspective, this formulation of the robustness problem presents some intriguing challenges. How to formulate interesting neighbourhood classes of distributions? How to calculate, with respect to such prior classes, bounds on posterior quantities of interest such as expectations or probabilities?

At the time of writing, this is an area of intensive research. For example, should neighbourhoods be defined parametrically or non-parametrically? And, if nonparametrically, what measures of distance should be used to define a neighbourhood "close" to p_0? Should the elements, p, of the neighbourhood be those such that the density ratio p/p_0 is bounded in some sense? Or such that the maximum

difference in the probability assigned to any event under p and p_0 is bounded? Or such that p can be written as a "contamination" of p_0, $p = (1-\varepsilon)p_0 + \varepsilon q$, for small ε and q belonging to a suitable class?

As yet, few issues seem to be resolved and we shall not, therefore, attept a detailed overview. Relevant references include; Edwards *et al.* (1963), Dawid (1973), Dempster (1975), Hill (1975), Meeden and Isaacson (1977), Rubin (1977, 1988a, 1988b), Kadane and Chuang (1978), Berger (1980, 1982, 1985a), DeRobertis and Hartigan (1981), Hartigan (1983), Kadane (1984), Berger and Berliner (1986), Kempthorne (1986), Berger and O'Hagan (1988), Cuevas and Sanz (1988), Pericchi and Nazaret (1988), Polasek and Pötzelberger (1988, 1994), Carlin and Dempster (1989), Delampady (1989), Sivaganesan and Berger (1989, 1993), Wasserman (1989, 1992a, 1992b), Berliner and Goel (1990), Delampady and Berger (1990), Doksum and Lo (1990), Wasserman and Kadane (1990, 1992a, 1992b), Ríos (1990, 1992), Angers and Berger (1991), Berger and Fan (1991), Berger and Mortera (1991b, 1994), Lavine (1991a, 1991b, 1992a, 1992b, 1994), Lavine *et al.* (1991, 1993), Moreno and Cano (1991), Pericchi and Walley (1991), Pötzelberger and Polasek (1991), Sivaganesan (1991), Walley (1991), Berger and Jefferys (1992), Gilio (1992b), Gómez-Villegas and Maín (1992), Moreno and Pericchi (1992, 1993), Nau (1992), Sansó and Pericchi (1992), Liseo *et al.* (1993), Osiewalski and Steel (1993), Bayarri and Berger (1994), de la Horra and Fernández (1994), Delampady and Dey (1994), O'Hagan (1994b), Pericchi and Pérez (1994), Ríos and Martín (1994), Salinetti (1994). There are excellent reviews by Berger (1984a, 1985a, 1990, 1994) and Wasserman (1992a), which together provide a wealth of further references.

Finally, in the case of a large data sample, one might wonder whether the data themselves could be used to suggest a suitable form of parametric model component, thus removing the need for detailed specification and hence the arbitrariness of the choice. The so-called *Bayesian bootstrap* provides such a possible approach; see, for instance, Rubin (1981) and Lo (1987, 1993). However, since it is a heavily computationally based method we shall defer discussion to the volume *Bayesian Computation*.

> The term *Bootstrap* is more familiar to most statisticians as a computationally intensive *frequentist* data-based simulation method for statistical inference; in particular, as a computer-based method for assigning frequentist measures of accuracy to point estimates. For an introduction to the method–and to the related technique of *jackknifing*—see Efron (1982). For a recent textbook treatment, see Efron and Tibshirani (1993). See, also, Hartigan (1969, 1975).

5.6.4 Hierarchical and Empirical Bayes

In Section 4.6.5, we motivated and discussed model structures which take the form of an hierarchy. Expressed in terms of generic densities, a simple version of such

an hierarchical model has the form

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) = \prod_{i=1}^{k} p(\boldsymbol{x}_i|\boldsymbol{\theta}_i),$$

$$p(\boldsymbol{\theta}|\boldsymbol{\phi}) = p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k|\boldsymbol{\phi}) = \prod_{i=1}^{k} p(\boldsymbol{\theta}_i|\boldsymbol{\phi}),$$

$$p(\boldsymbol{\phi}).$$

The basic interpretation is as follows. Observables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are available from k different, but related, sources: for example, k individuals in a homogeneous population, or k clinical trial centres involved in the same study. The first stage of the hierarchy specifies parametric model components for each of the k observables. But because of the "relatedness" of the k observables, the parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ are themselves judged to be exchangeable. The second and third stages of the hierarchy thus provide a prior for $\boldsymbol{\theta}$ of the familiar mixture representation form

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) = \int \prod_{i=1}^{k} p(\boldsymbol{\theta}_i|\boldsymbol{\phi}) p(\boldsymbol{\phi}) d\boldsymbol{\phi}.$$

Here, the "hyperparameter" $\boldsymbol{\phi}$ typically has an interpretation in terms of characteristics—for example, mean and covariance—of the population (of individuals, trial centres) from which the k units are drawn.

In many applications, it may be of interest to make inferences both about the unit characteristics, the $\boldsymbol{\theta}_i$'s, and the population characteristics, $\boldsymbol{\phi}$. In either case, straightforward probability manipulations involving Bayes' theorem provide the required posterior inferences as follows:

$$p(\boldsymbol{\theta}_i|\boldsymbol{x}) = \int p(\boldsymbol{\theta}_i|\boldsymbol{\phi}, \boldsymbol{x}) p(\boldsymbol{\phi}|\boldsymbol{x}) d\boldsymbol{\phi},$$

where

$$p(\boldsymbol{\theta}_i|\boldsymbol{\phi}, \boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i|\boldsymbol{\phi})$$

$$p(\boldsymbol{\phi}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\phi}) p(\boldsymbol{\phi}),$$

and

$$p(\boldsymbol{x}|\boldsymbol{\phi}) = \int p(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\phi}) d\boldsymbol{\theta}.$$

Of course, actual implementation requires the evaluation of the appropriate integrals and this may be non-trivial in many cases. However, as we shall see in the volumes *Bayesian Computation* and *Bayesian Methods*, such models can be implemented in a fully Bayesian way using appropriate computational techniques.

A detailed analysis of hierarchical models will be provided in those volumes; some key references are Good (1965, 1980b), Ericson (1969a, 1969b), Hill (1969, 1974), Lindley (1971), Lindley and Smith (1972), Smith (1973a, 1973b), Goldstein and Smith (1974), Leonard (1975), Mouchart and Simar (1980), Goel and DeGroot (1981), Goel (1983), Dawid (1988b), Berger and Robert (1990), Pérez and Pericchi (1992), Schervish *et al.* (1992), van der Merwe and van der Merwe (1992), Wolpert and Warren-Hicks (1992) and George *et al.* (1993, 1994).

A tempting approximation is suggested by the first line of the analysis above. We note that if $p(\boldsymbol{\phi}|\boldsymbol{x})$ were fairly sharply peaked around its mode, $\boldsymbol{\phi}^*$, say, we would have

$$p(\boldsymbol{\theta}_i|\boldsymbol{x}) \approx p(\boldsymbol{\theta}_i \mid \boldsymbol{\phi}^*, \boldsymbol{x}).$$

The form that results can be thought of *as if* we first use the data to estimate $\boldsymbol{\phi}$ and then, with $\boldsymbol{\phi}^*$ as a "plug-in" value, use Bayes' theorem for the first two stages of the hierarchy. The analysis thus has the flavour of a Bayesian analysis, but with an "empirical" prior based on the data.

Such short-cut approximations to a fully Bayesian analysis of hierarchical models have become known as *Empirical Bayes* methods. This is actually slightly confusing, since the term was originally used to describe frequentist estimation of the second-stage distribution: see Robbins (1955, 1964, 1983). However, more recently, following the line of development of Efron and Morris (1972, 1975) and Morris (1983), the term has come to refer mainly to work aimed at approximating (aspects of) posterior distributions arising from hierarchical models.

The naïve approximation outlined above is clearly deficient in that it ignores uncertainty in $\boldsymbol{\phi}$. Much of the development following Morris (1983) has been directed to finding more defensible approximations. For more whole-hearted Bayesian approaches, see Deely and Lindley (1981), Gilliland *et al.* (1982), Kass and Steffey (1989) and Ghosh (1992a). An eclectic account of empirical Bayes methods is given by Maritz and Lwin (1989).

5.6.5 Further Methodological Developments

The distinction between *theory* and *methods* is not always clear-cut and the extensive Bayesian literature on specific methodological topics obviously includes a wealth of material relating to Bayesian concepts and theory. We shall review this material in the volume *Bayesian Methods* and confine ourselves here to simply providing a few references.

Among the areas which have stimulated the development of Bayesian theory, we note the following: *Actuarial Science and Insurance* (Jewell, 1974, 1988; Singpurwalla and Wilson, 1992), *Calibration* (Dunsmore, 1968; Hoadley, 1970; Brown and Mäkeläinen, 1992), *Classification and Discrimination* (Geisser, 1964, 1966; Binder, 1978; Bernardo, 1988, 1994; Bernardo and Girón, 1989; Dawid

and Fang, 1992), *Contingency Tables* (Lindley, 1964; Good, 1965, 1967; Leonard, 1975; Leonard and Hsu, 1994), *Control Theory* (Aoki, 1967; Sawagari *et al.*, 1967), *Econometrics* (Mills, 1992; Steel, 1992), *Finite Population Sampling* (Basu, 1969, 1971; Ericson, 1969b, 1988; Godambe, 1969, 1970; Smouse, 1984; Lo, 1986), *Image Analysis* (Geman and Geman, 1984; Besag, 1986, 1989; Geman, 1988; Mardia *et al.*, 1992; Grenander and Miller, 1994), *Law* (Dawid, 1994), *Meta-Analysis* (DuMouchel and Harris, 1992; Wolpert and Warren-Hicks, 1992), *Missing Data* (Little and Rubin, 1987; Rubin, 1987; Meng and Rubin, 1992), *Mixtures* (Titterington *et al.*, 1985; Berliner, 1987; Bernardo and Girón, 1988; Florens *et al.*, 1992; West, 1992b; Diebolt and Robert, 1994; Robert and Soubiran, 1993; West *et al.*, 1994), *Multivariate Analysis* (Brown *et al.*, 1994), *Quality Assurance* (Wetherill and Campling, 1966; Hald, 1968; Booth and Smith, 1976; Irony *et al.*, 1992; Singpurwalla and Soyer, 1992), *Splines* (Wahba, 1978, 1983, 1988; Gu, 1992; Ansley *et al.*, 1993; Cox, 1993), *Stochastic Approximation* (Makov, 1988) and *Time Series and Forecasting* (Meinhold and Singpurwalla, 1983; West and Migon, 1985; Mortera, 1986; Smith and Gathercole, 1986; West and Harrison, 1986, 1989; Harrison and West, 1987; Ameen, 1992; Carlin and Polson, 1992; Gamerman, 1992; Smith, 1992; Gamerman and Migon, 1993; McCulloch and Tsay, 1993; Pole *et al.*, 1994).

5.6.6 Critical Issues

We conclude this chapter on inference by briefly discussing some further issues under the headings: (i) *Model Conditioned Inference*, (ii) *Prior Elicitation*, (iii) *Sequential Methods* and (iv) *Comparative Inference*.

Model Conditioned Inference

We have remarked on several occasions that the Bayesian learning process is predicated on a more or less formal framework. In this chapter, this has translated into model conditioned inference, in the sense that all prior to posterior or predictive inferences have taken place within the closed world of an assumed model structure.

It has therefore to be frankly acknowledged and recognised that all such inference is conditional. *If* we accept the model, *then* the mechanics of Bayesian learning—derived ultimately from the requirements of quantitative coherence—provide the appropriate uncertainty accounting and dynamics.

But what if, as individuals, we acknowledge some insecurity about the model? Or need to communicate with other individuals whose own models differ?

Clearly, issues of model criticism, model comparison, and, ultimately, model choice, are as much a part of the general world of confronting uncertainty as model conditioned thinking. We shall therefore devote Chapter 6 to a systematic exploration of these issues.

Prior Elicitation

We have emphasised, over and over, that our interpretation of a model requires—in conventional parametric representations—both a likelihood *and* a prior.

In accounts of Bayesian Statistics from a theoretical perspective—like that of this volume—discussions of the prior component inevitably focus on stylised forms, such as conjugate or reference specifications, which are amenable to a mathematical treatment, thus enabling general results and insights to be developed.

However, there is a danger of losing sight of the fact that, in real applications, prior specifications should be encapsulations of actual beliefs rather than stylised forms. This, of course, leads to the problem of how to elicit and encode such beliefs, i.e., how to structure questions to an individual, and how to process the answers, in order to arrive at a formal representation.

Much has been written on this topic, which clearly goes beyond the boundaries of statistical formalism and has proved of interest and importance to researchers from a number of other disciplines, including psychology and economics. However, despite its importance, the topic has a focus and flavour substantially different from the main technical concerns of this volume, and will be better discussed in the volume *Bayesian Methods*.

We shall therefore not attempt here any kind of systematic review of the very extensive literature. Very briefly, from the perspective of applications the best known protocol seems to be that described by Stäel von Holstein and Matheson (1979), the use of which in a large number of case studies has been reviewed by Merkhofer (1987). General discussion in a text-book setting is provided, for example, by Morgan and Henrion (1990), and Goodwin and Wright (1991). Warnings about the problems and difficulties are given in Kahneman *et al.* (1982). Some key references are de Finetti (1967), Winkler (1967a, 1967b), Edwards *et al.* (1968), Hogarth (1975, 1980) Dickey (1980), French (1980), Kadane (1980), Lindley (1982d), Jaynes (1985), Garthwaite and Dickey (1992), Leonard and Hsu (1992) and West and Crosse (1992).

Sequential Methods

In Section 2.6 we gave a brief overview of sequential decision problems but for most of our developments, we assumed that data were treated globally. It is obvious, however, that data are often available in sequential form and, moreover, there are often computational advantages in processing data sequentially, even if they are all immediately available.

There is a large Bayesian literature on sequential analysis and on sequential computation, which we will review in the volumes *Bayesian Computation* and *Bayesian Methods*. Key references include the seminal monograph of Wald (1947), Jackson (1960), who provides a bibliography of early work, Wetherill (1961), and the classic texts of Wetherill (1966) and DeGroot (1970). Berger and

Berry (1988) discuss the relevance of *stopping rules* in statistical inference. Some other references, primarily dealing with the analysis of stopping rules, are Amster (1963), Barnard (1967), Bartholomew (1967), Roberts (1967), Basu (1975) and Irony (1993). Witmer (1986) reviews multistage decision problems.

Comparative Inference

In this and in other chapters, our main concern has been to provide a self-contained systematic development of Bayesian ideas. However, both for completeness, and for the very obvious reason that there are still some statisticians who do not currently subscribe to the position adopted here, it seems necessary to make some attempt to compare and contrast Bayesian and non-Bayesian approaches.

We shall therefore provide, in Appendix B, a condensed critical overview of mainstream non-Bayesian ideas and developments. Any reader for whom our treatment is too condensed, should consult Thatcher (1964), Pratt (1965), Bartholomew (1971), Press (1972/1982), Barnett (1973/1982), Cox and Hinkley (1974), Box (1983), Anderson (1984), Casella and Berger (1987, 1990), DeGroot (1987), Piccinato (1992) and Poirier (1993).