CHAPTER 9

Noninformative Priors

9.1. Introduction. Determination of noninformative priors obtained by matching posterior and frequentist probabilities depends on higher order asymptotics. The noninformative priors introduced by Bernardo (1979) and Berger and Bernardo (1989) are based on first order asymptotics, namely, the asymptotic normality of posteriors. We discuss both briefly in this chapter, partly for the sake of completeness and partly because for a small number of parameters, the choice of a prior is usually important only for a moderately large n, $n = 10$ or so; for large n, the asymptotic normality of posterior (under regularity conditions on p and smoothness of π) leads to practically no influence of the prior on the posterior. So the choice of such priors does seem to be within the domain of higher order asymptotics.

Bernardo (1979) and Berger and Bernardo (1989) have called their noninformative priors reference priors. This seems an appropriate terminology for all noninformative priors. They may be viewed as an origin or a reference point against which a given prior, incorporating subjective opinion, can be judged. Moreover the posterior based on such a prior provides a Bayesian reporting of data which is close to being "objective" or free from prior subjective belief as far as this is permitted in the Bayesian paradigm. For both these reasons they are likely to play an important role in Bayesian analysis. [See also Berger (1985).] An emerging terminology for them is default or automatic priors, which indicates their use in quick Bayesian data analysis in default of a fully subjective Bayesian treatment.

As discussed in Ghosh and Mukerjee (1992b), there are usually four notions associated with noninformative priors:

1. Maximizing entropy or minimizing information.
2. Matching what a frequentist might do (since, one may argue, that is how a noninformative Bayesian should act).
3. Invariance.
4. Minimaxity (in a weak form).

In the next two sections we will be discussing notions 1 and 2. Asymptotically, consideration of invariance or minimaxity leads one to the Jeffreys prior, which will also appear as an entropy minimizing prior in the next section. Invariance under a group often leads to choosing the right Haar measure as a noninformative prior. Such a prior will also emerge from considerations like notion 2; we have already seen this in Example 8.2. If the inference problem is one of estimating the density itself and the loss is the entropy loss, then the criterion of Bernardo (1979) may be interpreted as choosing an asymptotically least favorable prior. We will not discuss notions 3 and 4 any further, except to remark that there is a vast literature on 3 and almost none on 4; 4 deserves more attention from noninformative Bayesians.

9.2. Reference priors. For a r.v. or random vector, with density (or probability function) $p(z)$, let

$$I(p) = \text{entropy of } p = -E_p(\log p(Z)). \tag{9.1}$$

We will also write $I(Z)$ for $I(p)$. If Y, Z are two r.v.'s with joint density p,

$$I(p) = I(Y,Z) = I(Y) + I_Y(Z),$$

where

$$I_Y(Z) = E(I(Z|Y))$$

and

$$I(Z|Y) = -E\{\log p(Z|Y)|Y\}.$$

Let $X = (X_1, X_2, \dots, X_n)$ have density (or p.f.) $p(x|\theta)$ and let θ have prior density $p(\theta)$. We will always assume $X_1, \dots, X_n$ are i.i.d. and $p(x|\theta)$ satisfies the sort of regularity condition needed for asymptotic normality of the posterior.

Lindley's measure of information in X is

$$I(X,p(\theta)) \underset{\text{def}}{=} I(\theta) - I_X(\theta), \tag{9.2}$$

which may be written as a Kullback–Leibler divergence between prior and posterior or as an expected divergence between $p(x|\theta)$ and $p(X) = \int p(x|\theta)p(\theta)\,d\theta$. Thus,

$$I(X,p(\theta)) = E\left\{\log\frac{p(\theta|x)}{p(\theta)}\right\} = E\left\{\log\frac{p(x|\theta)}{p(X)}\right\}. \tag{9.3}$$

Bernardo (1979) points out that the more information there is in X, the less information there is in the prior. He therefore proposed maximizing I in some asymptotic sense as $n \to \infty$. Fix an increasing sequence of compact sets K_i whose union is the whole parameter space. In the following text, initially we fix K_i and let $n \to \infty$. Then, as shown in Clarke and Barron (1990), under regularity conditions,

$$\begin{aligned} I(X,p) = {} & \frac{d}{2}\log\frac{n}{2\pi e} + \int_{K_i} p(\theta)\log\{\det I_F(\theta)\}^{1/2}\,d\theta \\ & - \int_{K_i} p(\theta)\log p(\theta)\,d\theta + o(1), \end{aligned} \tag{9.4}$$

where d is the dimension of θ and $I_F(\theta)$ is the $d \times d$ Fisher information matrix. (The suffix F will be dropped when there is no fear of confusion.) While (9.4) can be guessed easily from the posterior normality of Chapter 5 or Chapter 8 (see Theorem 8.1), it does not follow from them because the entropy distance used here is finer than L_1 distance, being more influenced by the tails. Similar results have also been obtained by many others including Polson (1988) and Ibragimov and Has'minskii (1982). According to (9.1), I is the sum of a constant which does not depend on the prior and a term which converges to the functional

$$(9.5)\quad J(p, K_i) = \int_{K_i} p(\theta)\log\{\det I_F(\theta)\}^{1/2}\, d\theta - \int_{K_i} p(\theta)\log p(\theta)\, d\theta.$$

We may therefore maximize J with respect to all priors p over K_i. This gives the Jeffreys prior restricted to K_i and normalized to integrate to 1, that is,

$$(9.5a)\qquad \begin{aligned} p_{K_i}(\theta) &= \text{const.}\{\det I_F(\theta)\}^{1/2} \quad \text{if } \theta \in K_i \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

If we now let $i \to \infty$, we may regard p_{K_i}'s as converging to the Jeffreys improper prior

$$(9.6)\qquad p_J(\theta) = \det\{I_F(\theta)\}^{1/2} \qquad \forall\, \theta$$

in the sense that for any two Borel sets B_1, B_2 contained in a compact set K_{i_0},

$$(9.7)\qquad \lim_{i \to \infty} \frac{\int_{B_1} p_{K_i}(\theta)\, d\theta}{\int_{B_2} p_{k_i}(\theta)\, d\theta} = \frac{\int_{B_1} p_J(\theta)\, d\theta}{\int_{B_2} p_J(\theta)\, d\theta}.$$

The most interesting application of this idea occurs when we have nuisance parameters. Let $\theta = (\theta_1, \theta_2)$, where θ_1 is the d_1-dimensional parameter of interest and θ_2 is the d_2-dimensional nuisance parameter, $d_1 + d_2 = d$. Suppose $p_i(\theta_2|\theta_1)$ is given on K_i and our only object is to find $p_i(\theta_1)$ on K_i. Then $p_i(\theta_1)$ is determined as follows. Consider the functional

$$(9.8)\qquad \begin{aligned} I(X, p_i(\theta_1)) &\underset{\text{def}}{=} E\left\{\log\frac{p_i(\theta_1|X)}{p_i(\theta_1)}\right\} \\ &= I(X, p_i(\theta_1, \theta_2)) - \int_{K_{i1}} p_i(\theta_1) I(X, p_i(\theta_2|\theta_1))\, d\theta_1, \end{aligned}$$

where K_{i1} is the projection of K_i to θ_1 space, $I(X, p_i(\theta_2|\theta_1))$ is Lindley's measure (9.2) with θ_1 held fixed and $p_i(\theta_2|\theta_1)$ taking the role of the prior.

Assuming that integration with respect to θ_1 and limit as $n \to \infty$ can be interchanged, we get from (9.4) applied to $I(X, p(\theta_2|\theta_1))$,

$$(9.9)\qquad I(X, p_i(\theta_1)) = \frac{d_1}{2}\log\frac{n}{2\pi e} + \int_{K_{i1}} p_i(\theta_1)\left\{\log\frac{\psi_i(\theta_1)}{p_i(\theta_1)}\right\} d\theta,$$

where

$$(9.10)\qquad \psi_i(\theta_1) = \exp\left\{\int_{K_1(\theta_1)} p_i(\theta_2|\theta_1)\log\left(\frac{\det I_F(\theta)}{\det I_{22}(\theta)}\right)^{1/2} d\theta_2\right\},$$

$$(9.11)\qquad I_F(\theta) = \text{Fisher information matrix for } \theta,$$

$$(9.12)\qquad I_{22}(\theta) = \text{Fisher information matrix for } \theta_2, \text{ given } \theta_1 \text{ is fixed,}$$

$$(9.13)\qquad K_i(\theta_1) = \{\theta_2; (\theta_1, \theta_2) \in K_i\}.$$

Maximizing with respect to all marginals supported on K_{i1} gives

$$(9.14)\qquad p_i(\theta_1) = \text{const.}\ \psi_i(\theta_1) \quad \text{on } K_{i1}.$$

The limiting process described in Berger and Bernardo (1989) is as follows. Let the compact sets K_i be rectangles $K_{i1} \times K_{i2}$, so that $K_i(\theta_1) = K_{i2}$ for $\theta_1 \in K_{i1}$. Consider the prior

$$(9.15)\qquad \begin{aligned} p_i(\theta_1, \theta_2) &= p_i(\theta_1)p_i(\theta_2|\theta_1) \quad \text{on } K_i \\ &= 0 \quad \text{outside } K_i. \end{aligned}$$

Fix θ_0 and assume

$$(9.16)\qquad \lim_{i\to\infty} p_i(\theta_1, \theta_2)/p_i(\theta_{10}, \theta_{20}) = p(\theta_1, \theta_2)$$

exists for all θ. Then $p(\theta_1, \theta_2)$ is the reference prior when θ_1 is the parameter of interest and θ_2 is the nuisance parameter. If the convergence in (9.16) is uniform on compact sets of θ, then (9.7) will hold with p_i, replacing p_{K_i}.

Berger and Bernardo (1989) choose $p_i(\theta_2|\theta_1)$ as follows. For fixed θ_1, choose a reference prior for θ_2 over K_{i2}, that is, in view of (9.5),

$$(9.17)\qquad p_i(\theta_2|\theta_1) = \text{const.}\{\det I_{22}(\theta)\}^{1/2} \quad \text{on } K_{i2}.$$

EXAMPLE 8.2 (Revisited). Here X_i's are i.i.d. normal with mean θ_2 and variance θ_1; θ_1 is the parameter of importance. Here, dropping the suffix F,

$$(9.18)\qquad I(\theta) = \begin{bmatrix} \dfrac{1}{2\theta_1^2} & 0 \\ 0 & \dfrac{1}{\theta_1} \end{bmatrix},$$

$$(9.19)\qquad p_i(\theta_2|\theta_1) = \text{const. (i.e., free of } \theta_2) \cdot \theta_1^{-1} \quad \text{on } K_{i2},$$

that is,

$$(9.20)\qquad p_i(\theta_2|\theta_1) = \frac{1}{\text{measure of } K_{i2}} \equiv d_i \quad \text{on } K_{i2}.$$

Hence

$$\psi_i(\theta) = \exp\left[\int_{K_{i2}} d_i \log\frac{1}{\sqrt{2}\,\theta_1}\,d\theta_2\right] = \exp\left\{\log\frac{1}{\sqrt{2}\,\theta_1}\right\}. \tag{9.21}$$

So

$$p_i(\theta_1, \theta_2) = c_i(1/\theta_1) \tag{9.22}$$

and

$$p(\theta_1, \theta_2) = (\theta_{10})/\theta_1, \tag{9.23}$$

which is (induced by) the right invariant Haar measure and also a solution of the differential equation (8.51). Jeffreys himself thought that this was the right choice of the noninformative measure in this example. The Jeffreys prior $\{\det I(\theta)\}^{1/2}$ is proportional to $\theta_1^{-3/2}$, which is (induced by) the left Haar measure.

If we do not take $p_i(\theta_2|\theta_1)$ as given, but choose both $p_i(\theta_1)$ and $p_i(\theta_2|\theta_1)$ by maximizing (9.8), then, as pointed out in Ghosh and Mukerjee (1992b), the solution is usually unacceptable because $p_i(\theta_2|\theta_1)$ tends to be discrete, in the way that least favorable priors sometimes turn out to be discrete. It is suggested in Ghosh and Mukerjee (1992b) that we should maximize (9.8) after introducing a penalty term for deviation from the uniform. An inspection of (9.5) will show that the second term on the right in (9.5) is such a penalty term. The reason for introducing a penalty is that one then has a compromise between maximizing I without deviating too much from the uniform. If one maximizes

$$\int_{K_i}\int p_i(\theta_1\,\theta_2)\log\left(\frac{\det I}{\det I_{22}}\right)^{1/2} d\theta_1\,d\theta_2 - \lambda I(p_i(\theta_1, \theta_2)) \tag{9.24}$$

with $\lambda = 1$, one gets

$$p_i(\theta_1, \theta_2) = \text{const.}(\det I/\det I_{22})^{1/2} \quad \text{on } K_i, \tag{9.25}$$

which in Example 8.2 again leads to the Berger–Bernardo reference prior (9.23). Clarke and Wasserman (1992) have examined the consequence of introducing a penalty for deviating from the Jeffreys prior.

We end this section with a quick comment on why all noninformative priors should depend on the relative importance of the parameters. As in the case of tests of randomness for a finite sequence, one cannot expect that the criteria for being noninformative would be satisfied by the prior with respect to all parametric functions or even all the components of the parameter. One tries to satisfy most of the criteria only for the components considered most important.

9.3. Noninformative priors via matching posterior and frequentist probabilities. We proceed as in Section 8.4, but with more stress this time on one-sided confidence sets. We assume there is a one-dimensional parameter $\theta = \theta_1$, or θ is two dimensional, $\theta = (\theta_1, \theta_2)$, θ_1 is the parameter of importance and, without loss of generality, θ_2 is orthogonal to θ_1 in the sense of Chapter 8. Let $\pi \equiv p$ stand for the prior density in both cases. Choose $\theta_{1,\alpha}(X)$, depending on the prior, such that the posterior probability

$$P\{\theta_1 \le \theta_{1,\alpha}(X)|X\} = 1 - \alpha + O(n^{-1}). \tag{9.26}$$

Now choose the prior satisfying any one of the following three conditions.

CONDITION A (No nuisance parameter).

$$P\{\theta_1 \le \theta_{1,\alpha}(X)|\theta_1\} = 1 - \alpha + O(n^{-1}) \tag{9.27}$$

uniformly on compact sets of θ_1.

CONDITION B (Orthogonal nuisance parameter).

$$P\{\theta_1 \le \theta_{1,\alpha}(X)|\theta_1, \theta_2\} = 1 - \alpha + O(n^{-1}) \tag{9.28}$$

uniformly on compact sets of θ.

CONDITION C (Integrated out orthogonal nuisance parameter).

$$\int P\{\theta_1 \le \theta_{1,\alpha}(X)|\theta_1, \theta_2\}\pi(\theta_2|\theta_1)\, d\theta_2 = 1 - \alpha + O(n^{-1}) \tag{9.29}$$

uniformly on compact sets of θ_1.

In each case, the condition is to be satisfied for all α.

The fact that we want $O(n^{-1})$ in these conditions can be explained as follows. If we replace $O(n^{-1})$ by $O(n^{-1/2})$, then by posterior normality, the conditions would hold for all smooth priors. On the other hand, if we replace $O(n^{-1})$ by $O(n^{-3/2})$, we would get two sets of differential equations for p to satisfy and, generally, they will not have a solution. So $O(n^{-1})$ is just right for these one-sided intervals. One the other hand, for similar reasons, $O(n^{-2})$ is right for two-sided intervals as in Section 8.4.

The solution of Condition A owing to Welch and Peers (1963) and rediscovered by Stein (1985), is $p(\theta_1) = \sqrt{I(\theta_1)}$, the Jeffreys prior. Condition A has no obvious generalization to dimension greater than 1. One possible generalization will be considered later, but, in general, one does not get the Jeffreys prior.

Condition B leads to the differential equation [Tibshirani (1989)]

$$\frac{\partial (I^{20})^{1/2}}{\partial \theta_1} = -(I^{20})^{1/2}\frac{\pi_{10}}{\pi}, \tag{9.30}$$

where

$$I^{20} = \frac{1}{I_{20}},$$

$$I_{20} = E\left\{-\frac{\partial^2 \log p(x_1|\theta)}{\partial \theta_1^2}\middle|\theta\right\}$$

and

$$\pi_{10} = \frac{\partial \pi}{\partial \theta_1}.$$

In the absence of orthogonality, (9.30) takes the form

$$\sum_1^2 \frac{\partial}{\partial \theta_i}\left(\pi(\theta)\frac{I^{i1}}{\sqrt{I^{11}}}\right) = 0. \tag{9.30a}$$

The solution of (9.30) is

$$\pi = \{I_{20}(\theta)\}^{1/2} d(\theta_2), \tag{9.31}$$

where $d(\theta_2)$ is an arbitrary function. In general, one cannot increase the dimension of θ_1 since then an orthogonal θ_2, needed for (9.31), may not exist. Tibshirani (1989) notes that the reference prior for the product mean problem in Berger and Bernardo (1989) satisfies this equation. We come back to this point after discussing Condition C.

Condition C leads to the equation

$$\int \left\{\frac{\partial}{\partial \theta_2}\log\left(\frac{\pi(\theta)}{I_{20}^{1/2}}\right)\frac{\pi(\theta_2|\theta_1)}{I_{20}^{1/2}}\right\} d\theta_2 = 0 \qquad \forall\, \theta_1. \tag{9.32}$$

On writing $\pi(\theta) = \pi(\theta_1)\pi(\theta_2|\theta_1)$ above, we get the solution

$$\pi(\theta_1) = \text{const.}\,\frac{1}{\int\left(\pi(\theta_2|\theta_1)/\sqrt{I_{20}}\right) d\theta_2}. \tag{9.33}$$

The solution of Condition B has the following implications noted in Ghosh and Mukerjee (1992b):

1. The reference prior for θ_1 is, in general, not a solution, although it often is in practice.
2. Curiously enough, the reference prior for θ_2 is a solution. In general, because of the requirement of orthogonality, θ_2 is not a "natural" parameter. However, where θ_2 appears in the natural or common parametrization and is orthogonal to θ_1, as in Example 8.2, one may prefer the reference prior for θ_2 even when the parameter of interest is θ_1.
3. The new prior suggested in Section 9.2 with a penalty term added is a solution, provided θ_2 is orthogonal to θ_1. This seems to be an attractive choice if, as in Example 8.2, the natural nuisance parameter is orthogonal to θ_1, but again, it is difficult to say generally.
4. Sometimes a reference or otherwise desirable prior may satisfy (9.30) approximately and hence meet Condition B approximately but satisfactorily for practical purposes. This too needs further study, at least partly through simulation. One particular case involving the MANOVA is contained in a thesis of Anindya De, which is under preparation for submission to Purdue University.

One may ask similar questions on the choice of π based on matching posterior and frequentist probability for "two-sided" intervals as in Section 8 up to $O(n^{2})$:

1. This fails for two-sided symmetric intervals of the kind $\hat{\theta} \pm A(X_1, \dots, X_n)$. The solution depends on α. This is shown in Lee (1989).
2. This works for all other natural two-sided intervals, namely,
   (a) highest posterior density sets,
   (b) equal tail interval (rather than equal cutoff points like notion 1),
   (c) likelihood ratio based intervals.

For 2(a), see Ghosh and Mukerjee (1993a); 2(c) has been discussed earlier. The case 2(b) has been considered in Peers (1968) and unpublished work of Ghosh and Mukerjee.

In many examples, the reference prior is also a probability matching noninformative prior. We list a few such examples.

EXAMPLE 9.1 [The product mean problem of Berger and Bernardo (1989)]. Let $X_i = (X_{i1}, X_{i2})$, X_i's are i.i.d. and X_{i1}, X_{i2} are independent normal $N(\mu_1, 1)$, $N(\mu_2, 1)$, $\mu_1 > 0$, $\mu_2 > 0$. We wish to estimate $\theta_1 = \mu_1 \cdot \mu_2$. Berger and Bernardo (1989) discuss why this example is interesting from a practical point of view. Let $\theta_2 = \mu_2^2 - \mu_1^2$. Then one can check that θ_1, θ_2 are orthogonal. The reference prior with θ_1 as the parameter of importance is

$$\pi(\theta) = \text{const.}\left(\theta_2^2 + 4\theta_1^2\right)^{-1/4}. \tag{9.34}$$

The solution to Condition B is

$$\pi(\theta) = A(\theta_2)\left(\theta_2^2 + 4\theta_1^2\right)^{-1/4}, \tag{9.35}$$

where $A(\cdot)$ is arbitrary.

EXAMPLE 9.2 (Ratio of normal means). Consider the same example as above, but suppose the parameter of interest is $\theta_1 = \mu_1/\mu_2$. Let $\theta_2 = (\mu_1^2 + \mu_2^2)^{1/2}$. Then θ_1 and θ_2 are orthogonal. The solution of Condition B is

$$\pi(\theta) = A(\theta_2) \cdot \theta_2/\left(1 + \theta_1^2\right). \tag{9.36}$$

The reference prior is

$$\pi(\theta) = \text{const.}\left(1 + \theta_1^2\right)^{-1}.$$

EXAMPLE 9.3 (Exponential regression model). We have n i.i.d. random vectors with common density

$$p(x|\theta) = \prod_{i=1}^{p}\left[\{\theta_2 \exp(\theta_1 Z_i)\}^{-1} \exp\left\{-\frac{x_i}{\theta_2} e^{-\theta_1 Z_i}\right\}\right],$$

where $x_1, x_2, \dots, x_p > 0$, $\theta = (\theta_1, \theta_2)$, $-\infty < \theta_1 < \infty$, $\theta > 0$ and $Z_1, Z_2, \dots, Z_p$ are known constants such that $\sum_1^p Z_i = 0$. Here θ_1, θ_2 are orthogonal. A solution of Condition B is

$$\pi(\theta) = A(\theta_2) \tag{9.37}$$

and the reference prior is

$$\pi(\theta) = \text{const.}/\theta_2. \tag{9.37a}$$

We now indicate how (9.30) for Condition B is obtained. (Equations for Conditions A and C are derived in a similar way.) We also suggest a possible generalization of Condition A. We follow the notations of Section 8.4; see (8.36) to (8.41).

We first note (9.28) of Condition B is equivalent to matching the frequentist and posterior probability up to $o(n^{-1/2})$, $\forall\ 0 < \alpha < 1$, of the set

$$A_\alpha = \left\{\sqrt{n}\, D^{-1/2}\left(\theta_1 - \hat{\theta}_1\right) < Z_\alpha\right\}. \tag{9.37b}$$

The posterior density of $\sqrt{n}\,(\theta_1 - \hat{\theta}_1)$ is given by

$$\begin{aligned}\pi(h_1 | X_1, X_2, \dots, X_n)& \\ = \phi(h_1, D^{-1})\Big[1 + n^{-1/2}\Big\{\tfrac{1}{6}h_1^3 b_{30} + h_1\left(\tfrac{1}{2}b_{12}c_{02}^{-1} + \hat{\pi}_{10}\hat{\pi}^{-1}\right)& \\ + h_1^3 R_1 - h_1 R_2\Big\}\Big] + o(n^{-1/2}),&\end{aligned} \tag{9.38}$$

where $\phi(\cdot, D^{-1})$ is the density of a normal with zero mean and variance D^{-1},

$$\begin{gathered} b_{ij} = l_{ij}(\hat{\theta}), \qquad c_{ij} = -b_{ij}, \qquad \hat{\pi}_{ij} = \pi_{ij}(\hat{\theta}), \qquad \hat{\pi} = \pi(\hat{\theta}), \\ R_1 = \tfrac{1}{2}b_{12}c_{11}^2 c_{02}^{-2} - \tfrac{1}{2}b_{21}c_{11}c_{02}^{-1} - \tfrac{1}{6}b_{03}c_{11}^3 c_{02}^{-3}, \\ R_2 = \hat{\pi}_{01}\hat{\pi}^{-1}c_{11}c_{02}^{-1} + \tfrac{1}{2}b_{03}c_{11}c_{02}^{-2} \end{gathered} \tag{9.38a}$$

and, because of orthogonality, $c_{11} = o_p(1)$ and hence so are R_1 and R_2. Hence the posterior probability of A_α is, under θ_0,

$$1 - \alpha + \frac{1}{\sqrt{n}}\,\frac{e^{-Z_\alpha^2/2}}{\sqrt{2\pi}} I_{20}^{-1/2}\,\frac{\pi_{10}}{\pi} + \frac{\psi(\theta_0)}{\sqrt{n}} + o(n^{-1}),$$

where ψ is free of π. We calculate $P_{\theta_0}(A_\alpha)$ as in Section 8.4 by a Bayesian route, that is, we replace π by $\pi_\delta \rightarrow_{w}$ the measure sitting on θ_0, and calculate the limit of the coefficient of $(\sqrt{n}\,)^{-1}$ above. Thus

$$\lim_{\delta \downarrow 0} \int \psi(\theta)\pi_\delta(\theta)\, d\theta = \psi(\theta_0) \tag{9.39}$$

and

$$\lim_{\delta \downarrow 0} \int I_{20}^{-1/2}\left(\frac{\pi_{10,\delta}}{\pi_\delta}\right)\pi_\delta\, d\theta = -\frac{\partial}{\partial\theta_1}\left(I_{20}^{-1/2}\right)\Big|_{\theta_0}. \tag{9.40}$$

Then

$$P_{\theta_0}(A_\alpha) = 1 - \alpha + \frac{e^{-Z_\alpha^2/2}}{\sqrt{n}\sqrt{2\pi}}\left\{\psi(\theta_0)\sqrt{2\pi}e^{z^2/2} - \frac{\partial}{\partial\theta_1}(I_{20}^{-1/2})\bigg|_{\theta_0}\right\} + o(n^{-1/2}). \tag{9.41}$$

Matching the coefficients of $(\sqrt{n})^{-1}$, we get (9.30).

The differential equation corresponding to Condition A is exactly the same, but instead of I_{20}, we have I, since the parameter is one dimensional. The solution will be given by (9.31) with $d(\theta_2)$ being replaced by a constant. This leads to the Jeffreys prior, as noted earlier. The differential equation corresponding to Condition C is obtained similarly.

We now consider a possible generalization of Condition A when θ is of more than one dimension. Arrange the components of θ in decreasing order of importance, that is, θ_1 is the most important, and so forth. *We do not assume orthogonality.* Choose a lower triangular matrix C^* such that

$$C^* C^{-1} C^{*\prime} = \text{identity}, \tag{9.42}$$

where

$$C = (c_{ij}), \qquad c_{ij} = -\frac{\partial^2 \log p(x_1, \dots, x_n|\theta)}{\partial\theta_i \partial\theta_j}\bigg|_{\hat{\theta}}.$$

Let

$$W = C^*\big(\sqrt{n}(\theta - \hat{\theta})\big). \tag{9.43}$$

Now match the posterior and frequentist distribution function of W up to $o(n^{-1/2})$. Note that the first element of W is the standardized version of $\sqrt{n}(\theta_1 - \hat{\theta}_1)$, and the ith element of W is the standardized regression residual of $\sqrt{n}(\theta_i - \hat{\theta}_i)$ on $\sqrt{n}(\theta_j - \hat{\theta}_j)$, $j = 1, 2, \dots, i-1$. The differential equations under this requirement are

$$\sum_i \frac{\partial}{\partial\theta_i}\big[\tilde{I}_{ir}\pi(\theta)\big] = 0 \qquad \forall\theta, 1 \le r \le p, \tag{9.44}$$

where I is the Fisher information matrix, I^* is a lower triangular matrix such that $I^* I^{-1} I^{*\prime}$ is identity and $\tilde{I} = (I^*)^{-1}$.

Example 8.2 (Revisited). We take the normal mean as θ_1 and the variance as θ_2. Then θ_1, θ_2 are orthogonal. Here (9.44) reduces to

$$\frac{\partial}{\partial\theta_1}\big(\sqrt{\theta_2}\,\pi(\theta)\big) = 0, \qquad \frac{\partial}{\partial\theta_2}(\theta_2\pi(\theta)) = 0. \tag{9.45}$$

Then, $\pi(\theta) = \text{const.}(\theta_2)^{-1}$ is a solution, but the Jeffreys prior is not a solution, in spite of orthogonality.

In general one cannot expect the same π to satisfy more than one equation as in (9.44), but in many important cases a solution does exist; see Ghosh and Mukerjee (1993b) for many examples. In particular, for some (but not all) orderings of importance of the components of θ, solutions have been obtained for the variance component problem and compared with the reference priors. Here we are matching the joint distribution of (standardized) $\sqrt{n}(\theta - \hat{\theta})$. In the presence of orthogonality we get the same equations if, as in Tibshirani (1989), we require the marginals of each component to satisfy Condition A.

We conclude this section with a discussion of what is to be done if the object is to match probabilities of two-sided intervals. Here the right order is $o(n^{-1})$, in fact $O(n^{-2})$, because of symmetry.

We first consider the equal tailed case. Choose $\xi^*_{1\alpha}(X_1, \dots, X_n)$, $\xi^*_{2\alpha}(X_1, \dots, X_n)$ such that the posterior probability

$$P\{\theta > \xi^*_{1\alpha} | X_1, X_2, \dots, X_n\} = \frac{\alpha}{2} + o(n^{-1}), \tag{9.46a}$$

$$P\{\theta < \xi^*_{2\alpha} | X_1, X_2, \dots, X_n\} = \frac{\alpha}{2} + o(n^{-1}) \tag{9.46b}$$

and then choose a prior π such that

$$P\{\theta > \xi^*_{1\alpha} | \theta\} = \frac{\alpha}{2} + o(n^{-1}), \tag{9.47a}$$

$$P\{\theta < \xi^*_{2\alpha} | \theta\} = \frac{\alpha}{2} + o(n^{-1}). \tag{9.47b}$$

In the absence of a nuisance parameter θ_2, that is, if $\theta_1 \equiv \theta$, (9.46) and (9.47) lead to

$$\frac{\pi'(\theta)}{I} + \left(L_{11} + \frac{2}{3}L_{001}\right)\frac{\pi(\theta)}{I^2}\theta = c \tag{9.48}$$

for some constant c, each specification of c will provide a solution of (9.46) and (9.47). Here

$$L_{ijk} = E\left\{\left(\frac{d\log p(X_1|\theta)}{d\theta}\right)^i\left(\frac{d\log p(X_1|\theta)}{d\theta}\right)^j\left(\frac{d\log p(X_1|\theta)}{d\theta}\right)^k\middle|\theta\right\}$$

and

$$L_{ij} = L_{ijo}.$$

The case where there is a nuisance parameter can be handled similarly.

If we replace equal tail cutoff points by equidistant cutoff points from $\hat{\theta}$, the resulting differential equation depends on α, which seems to indicate this matching is undesirable.

Possibly the most attractive matching is based on highest posterior density (HPD) sets. For details, see Ghosh and Mukerjee (1993a).

Other alternatives are confidence sets based on the likelihood ratio test, which have been discussed in Chapter 8, and the conditional likelihood ratio test; see Ghosh and Mukerjee (1992a). In the first case, the uniform prior is often a solution. It also appears that the sets have some Bayesian robustness properties.

9.4. Discussion of assumptions and interpretation. For the noninformative priors based on one-sided intervals, we need π is continuously differentiable and $\pi(\theta) > 0$ for all θ, that is, we need Johnson's (1970) condition for all θ for posterior expansion up to $o(n^{-1/2})$ under all P_θ. We need also the regularity condition on p assumed by Johnson. Note that we do not need π to be proper, provided there is an n_0 such that for $n = n_0$, the posterior is proper for all $(x_1, x_2, \dots, x_{n_0})$. We need also (i) valid Edgeworth assumptions on $\sqrt{n}(\hat{\theta} - \theta)$ up to $o(n^{-1/2})$ and (ii) a condition on

$$R'_n = \frac{1}{n^{3/2}} \left\{ \frac{\partial^2 \log p(X_1, X_2, \dots, X_n|\theta)}{\partial\theta_1^i \, \partial\theta_2^j} - nE\left(\frac{\partial^2 \log p(X_1|\theta)}{\partial\theta_i \, \partial\theta_j} \middle| \theta \right) \right\}$$

such that R'_n satisfies (2.2a) with $s = 3$. This would be true if (nR'_n) has a valid Edgeworth expansion up to $o(n^{-1/2})$. Conditions (i) and (ii) above certainly hold under the assumptions made in Section 2.6.

For matching based on two-sided intervals, we need Johnson's (1970) conditions for a posterior expansion up to $o(n^{-1})$ for all P_θ, and the Edgeworth assumptions of Section 2.6.

In both cases, that is, for one-sided and two-sided intervals, it is the Edgeworth assumptions which are most restrictive. In particular, for discrete X's, they are not available. However, such assumptions are not needed if we are prepared to replace, for each θ, a condition like

$$(9.49) \qquad P_\theta(A_{1-\alpha}) = 1 - \alpha + o(n^{-(s-2)/2}), \qquad s = 3 \text{ or } 4$$

by

$$(9.50) \qquad \lim_{\pi_\delta \to_w \delta_\theta} \int P_\theta(A_{1-\alpha}) \pi_\delta(\theta) \, d\theta = 1 - \alpha + o(n^{-(s-2)/2}),$$

where δ_θ is the probability measure with all its mass at θ and π_δ's satisfy the conditions of Chapter 5 in case θ is real valued and analogous conditions on the boundary of its support when θ is multidimensional. For example, it is sufficient if $\theta_1, \theta_2, \dots$ are independent and their marginals support the conditions of Chapter 5. One may think of (9.50) as a smoothed version of (9.49), which agrees with (9.49) only under Edgeworth assumptions.

The above is in the spirit of Woodroofe (1986).

Finally, a remark about interpretation seems desirable. Most of the solutions are improper priors. The matching conditions on the priors try to reflect the following simulation scenario. Fix a θ, draw a sample $X_{i1}, X_{i2}, \dots, X_{in}$ from P_θ, compute a posterior confidence interval $A_{1-\alpha}^{(X_{i1}, \dots, X_{in})}$ for θ_1 (or θ) and see if the frequentist probability

$$N^{-1}\#\{i;\ \theta_1 \text{ (or } \theta) \in A_{1-\alpha}(X_{i1}, \dots, X_{in})\}$$

matches $(1 - \alpha)$, approximately, that is, up to $o(n^{-1/2})$ or $o(n^{-1})$.

9.5. Comments on reference and probability matching noninformative priors. In most examples the reference priors seem to be identical with some probability matching noninformative prior, or close to being so. It would be interesting to know if there is some theoretical basis for this. Our own interest in the probability matching priors came from the hope that the reference priors would satisfy Conditions B or C of Section 9.3. That, unfortunately, has turned out to be false.

Also there is still some doubt as to the order of the steps in the algorithm for a reference prior. If θ_1 is the parameter of interest, should one use the reference prior for θ_1 or the reference prior for θ_2?

We should point out that in examples studied so far, the reference priors seem to be good choices.

We end with two technical questions. Under what conditions can a probability matching prior be obtained by maximizing a functional on the priors so that the differential equations obtained by matching coincide with the Euler equations for the variational problem? Second, are the probability matching priors least favorable in some sense? The answer to the second question would be yes if a functional described in the first question exists and can be interpreted as a Bayes risk.