NONINFORMATIVE BAYESIAN PRIORS. INTERPRETATION AND PROBLEMS WITH CONSTRUCTION AND APPLICATIONS.

ANNE RANDI SYVERSVEEN

1. Introduction

Central in Bayesian statistics is Bayes' theorem, which can be written as follows:

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta).$$

Given the likelihood function $f(x|\theta)$ and the prior $\pi(\theta)$, it is easy to calculate the posterior distribution of θ, $\pi(\theta|x)$, which is used for doing inference. An important problem in Bayesian analysis is how to define the prior distribution. If prior information about the parameter θ is available, it should be incorporated in the prior density. If we have no prior information, we want a prior with minimal influence on the inference. We call such a prior a noninformative prior.

An important question is, how do we construct a noninformative prior? The Bayes/Laplace postulate, stated about 200 years ago says the following: *When nothing is known about θ in advance, let the prior $\pi(\theta)$ be a uniform distribution, that is, let all possible outcomes of θ have the same probability.* This is also known as the *principle of insufficient reason.*

Fisher did not support the Bayes/Laplace postulate. He argued that *Not knowing the chance of mutually exclusive events and knowing the chance to be equal are two quite different states of knowledge.* He accepted Bayes' theorem only for informative priors.

The fundamental problem by using the uniform distribution as our noninformative prior, is that the uniform distribution is not invariant under reparametrization. If we have no information about θ, we also have no information about for example $1/\theta$, but a uniform prior on θ does not correspond to a uniform prior for $1/\theta$. By the transformation formula, the corresponding distribution for a one-to-one function $g(\theta)$ is given below:

$$\pi(\theta) = 1, \quad \phi = g(\theta) \quad \Rightarrow \pi(\phi) = |\frac{d}{d\phi} g^{-1}(\phi)|$$

Another problem with the uniform prior is that if the parameter space is infinite, the uniform prior is improper, which means, it does not integrate to one. This is however not always a serious problem, since improper priors often lead to proper posteriors.

These problems with the uniform prior will be more throughly discussed later. But first we discuss the interpretation of noninformative priors.

2. Interpretation of Noninformative Priors

Kass and Wasserman (1996) stated two different interpretations of noninformative priors: *1) Noninformative priors are formal representations of ignorance. 2) There is no objective, unique prior that represents ignorance, instead noninformative priors are chosen by public agreement much like units of length and weight.* In the second interpretation, noninformative priors are the "default" to use when there is insufficient information to otherwise define the prior. Today, no one use the first interpretation to claim that one particular prior is truly noninformative. The focus is on comparing different priors to see if any is preferable in some sense.

Box and Tiao (1973) define a noninformative prior as a prior which provides little information relative to the experiment. Bernardo and Smith (1994) use a similar definition, they say that noninformative priors have minimal effect relative to the data, on the final inference. They regard the noninformative prior as a mathematical tool, it is not a uniquely noninformative or objective prior. These definitions are related to the second interpretation of Kass and Wasserman (1996).

Pericchi and Walley (1991) have a quite different view. They say that no single probability distribution can model ignorance satisfactory, therefore large classes of distributions are needed. They use the first interpretation of Kass and Wasserman (1996), but they realize that a single distribution is not enough. Therefore they introduce classes of prior distributions.

In the next Section, different methods for finding noninformative priors are presented. Most of the methods are related to the second interpretation of Kass and Wasserman (1996), but a method related to what Pericchi and Walley (1991) say is also presented.

3. Invariant Noninformative Priors

In the introduction, we saw that the fundamental problem by using the uniform distribution as noninformative prior, is that it is not invariant under reparametrization. Now we will see how we can construct invariant noninformative priors.

One approach is to look for an invariance structure in the problem and let the prior have the same invariance structure. Mathematically, this means that the model and the prior should be invariant under action of the same group and we should use the right Haar measure as prior. The right Haar measure is the prior that is invariant to right multiplication with the group. For reasons not to be discussed here, we prefer the right invariant Haar measure instead of the left, as our noninformative prior. See for example Berger (1980) or Robert (1994) for a more throughly discussion of group invariance and Haar measures.

We illustrate the method by two simple examples.

Example 1. Location parameters:

Let X be distributed as $f(x-\theta)$, which is a location invariant density, and θ is called a location parameter. A location invariant density is invariant to linear transformations. This means that $Y = X + a$ is distributed as $f(y-\phi)$ with $\phi = \theta + a$, that is, X and Y have the same distribution, but with different location parameters. Since the model is location invariant, the prior distribution should be location invariant. Therefore:

$$\pi(\theta) = \pi(\theta - a) \quad \forall a \quad \Rightarrow \pi(\theta) = 1.$$

An invariant noninformative prior for a location parameter is the uniform distribution.

Another argument leading to the same result, is that since θ and ϕ are location parameters in the same model, they should have the same prior.

Example 2. Scale parameters:

LetX be distributed as $\frac{1}{\sigma}f(\frac{x}{\sigma})$, which is a scale invariant density with scale parameter σ. That the distribution is scale invariant, means that $Y = cX$ has the same distribution as X, but with a different scale parameter. Since the density is scale invariant, the prior distribution should be scale invariant:

$$\pi(A) = \pi(A/c) \quad \forall A \in (0, +\infty) \quad \textit{and} \quad c > 0.$$

This leads to

$$\pi(\sigma) = \frac{1}{c}\pi(\frac{\sigma}{c}) \quad c > 0 \quad \Rightarrow \pi(\sigma) = \sigma^{-1}$$

so the invariant noninfromative prior for a scale parameter is $\pi(\sigma) = \sigma^{-1}$, which is an improper distribution.

We see that in both cases, the invariant noninformative prior is improper. As we will see in later examples, this is often the case.

A difficulty with this method is that all problems do not have an invariance structure and the right Haar measure does not always exist. In the following we present methods for finding invariant noninformative priors which do not take the structure of the problem into account.

3.1. **Jeffreys' prior.** This method was described by Jeffreys (1946), and it is based on the Fisher information given by

$$I(\theta) = E_\theta(\frac{\partial \log f(x|\theta)}{\partial\theta})^2.$$

Jeffreys prior is defined as

$$\pi(\theta) \propto I^{1/2}(\theta).$$

Jeffreys justified his method by the fact that it satisfies the invariant reparametrization requirement, shown by the following two equations:

$$\begin{aligned} I(\theta) =& I(h(\theta))(h'(\theta))^2 \\ \pi(\theta) \propto& I(h(\theta))^{1/2}|h'(\theta)| = \pi(h(\theta))|h'(\theta)| \end{aligned}$$

In the last equation we recognize the transformation formula.

A motivation for Jeffreys' method is that the Fisher information $I(\theta)$ is an indicator of the amount of information brought by the model (observations) about θ. To favor the values for θ of which $I(\theta)$ is large is equivalent to minimizing the influence of the prior.

When the parameter θ is one-dimensional, the Jeffreys prior coincides with the right Haar measure when it exists.

Jeffreys prior can be generalized to multidimensional parameters θ by letting the prior be proportional to the square root of the determinant of the Fisher information matrix:

$$\pi(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}.$$

However, there are problems with this generalized Jeffreys prior, as the following example, taken from Bernardo and Smith (1994) will show.

Example 3. *We let $x = \{x_1, \dots, x_n\}$ be iid $N(\mu, \sigma^2)$. First, we assume that the mean is known, and equal to 0. Then we have a scale density, and Jeffreys noninformative prior for σ is given by $\pi(\sigma) = \sigma^{-1}$. With this choice of prior, the posterior of σ is such that*

$$[\sum_{i=1}^{n} x_i^2]/\sigma^2 \sim \chi^2_n.$$

Then we assume that the mean is unknown. The two dimensional Jeffreys prior for μ and σ is now

$$\pi(\mu, \sigma) = \sigma^{-2}. \tag{1}$$

With this choice of prior, the posterior of σ is such that

$$[\sum_{i=1}^{n} (x_i - \bar{x})^2]/\sigma^2 \sim \chi^2_n. \tag{2}$$

This is however unacceptable, since we would expect to loose one degree of freedom when we estimate μ.

Jeffreys' advice in this case, and other location-scale families was to assume that μ and σ are independent apriori and use the one-dimensional Jeffreys prior for each of the parameters. Then the prior for (μ, σ) is $\pi(\mu, \sigma) = \sigma^{-1}$, which is also the right invariant Haar measure, and gives us the correct degrees of freedom in expression (2). It can be mentioned that the prior given by equation (1) is the left invariant Haar measure.

3.2. **Reference priors.** Another well-known class of noninformative priors, is the reference prior, first described by Bernardo (1979) and further developed by Berger and Bernardo (1989). The method for deriving the reference prior is also referred to as the Berger-Bernardo method.

The method leads to Jeffreys' prior in the one-dimensional case, but as we see later, it is advantageous to Jeffreys' method in the multidimensional case. The definition of a reference prior is the prior that maximizes the missing information in the experiment. The reference prior is derived as follows. Let $X^n = \{X_1, \ldots, X_n\}$ be iid random variables. Define the Kullback-Leibler distance between the posterior and the prior distribution as

$$K_n(\pi(\theta|x^n), \pi(\theta)) = \int \pi(\theta|x^n) \log(\pi(\theta|x^n)/\pi(\theta)) d\theta.$$

Let K_n^π be the expected Kullback-Leibler distance with respect to X^n:

$$K_n^\pi = E_{X^n}(K_n(\pi(\theta|x^n), \pi(\theta)))$$

The missing information is now given as the limit of K_n^π as the number of observations, n goes to infinity. So we find the prior that maximizes

$$K_\infty^\pi = \lim_{n\to\infty} K_n^\pi.$$

Unfortunately, this limit is usually infinite. To overcome this difficulty, we find the prior π_n maximizing K_n^π and find the limit of the corresponding sequence of posteriors. Then the reference prior is given as the prior that produces the limiting posterior.

The Berger-Bernardo method can be extended to handle nuisance parameters. Then the parameter is given by $\theta = (\phi, \lambda)$, where ϕ is the parameter of interest and λ is the nuisance parameter. We can write the prior for θ as

$$\pi(\phi, \lambda) = \pi(\lambda|\phi)\pi(\phi).$$

The idea is now to first define the conditional prior $\pi(\lambda|\phi)$ to be the reference prior for λ with ϕ fixed. Then we find the marginal model

$$p(x|\phi) = \int p(x|\phi, \lambda)\pi(\lambda|\phi) d\lambda \tag{3}$$

and take the prior for ϕ, $\pi(\phi)$ to be the reference prior based on the marginal model $p(x|\phi)$. There are some technical problems here, because the prior $\pi(\lambda|\phi)$ is often improper, and the integral (3) diverges. To accomplish this, we restrict the integral to a sequence of compact sets.

The method is invariant in choice of nuisance parameter. This seems reasonable, since the parameter of interest is independent of the nuisance parameter.

The method can also be generalized to multidimensional parameter spaces. Then we let the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ be ordered according to importance, with θ_1 being the most important parameter. We write the prior for $\boldsymbol{\theta}$ as

$$\pi(\boldsymbol{\theta}) = \pi(\theta_m|\theta_1, \dots, \theta_{m-1}) \dots \pi(\theta_2|\theta_1)\pi(\theta_1)$$

and use the procedure above recursively. It should be noted that the ordering of the parameters is very important. Different orderings may lead to different reference priors. In some cases it might be difficult to chose the "correct" ordering. However, this method avoids the problem we saw with Jeffreys' multidimensional method.

3.3. **Other methods.** Some methods related to those already discussed are now mentioned. Box and Tiao (1973) described a method based on something they called *Data-translated likelihoods.* The method leads to Jeffreys' prior. A likelihood function is data-translated if it can be written as $L_y(\phi) = f(\phi - t(y))$. They suggested to use a uniform prior when this is satisfied. An "approximate data-translated likelihood" was introduced to motivate for Jeffreys' general rule.

Jaynes (1968) suggested to select the prior that maximizes the entropy. This method is only good for discrete, finite parameter space. If no further constraints are imposed on the problem, this method gives the uniform prior. The method has been used successfully in many problems.

Welch and Peers (1963) developed a method called *Probability matching.* They seek a prior $\pi(\theta)$ so that the posterior confidence interval for θ has coverage error $O(n^{-1})$ in the frequentist sense. This means that the difference between the posterior and frequentist confidence interval should be small. Their method is equivalent to Jeffreys' prior when θ is one-dimensional. Tibshirani (1989) extended the method to be able to handle nuisance parameters.

4. The Bounded Derivative Model

This model for prior ignorance is quite different from the noninformative priors described so far. This model is based on the following: We define a class $\mathcal{M}$ of prior distributions. The class of distributions $\mathcal{M}$ is understood through the upper and lower probabilities, defined by $\overline{P}(A) = \sup\{P_f(A) : f \in \mathcal{M}\}$ and $\underline{P}(A) = \inf\{P_f(A) : f \in \mathcal{M}\}$. The upper and lower probabilities can be interpreted as upper and lower betting rates. We will bet against A on rates larger than the upper probability and on A on rates smaller than the lower probability. For example, if the lower probability is 0.95, you should be willing to bet on A at odds of more than $0.95/0.05 = 19$ to 1. If the class $\mathcal{M}$ is convex and closed, it is fully determined through the upper and lower probabilities. It should be noted that the class is "*Not a class of reasonable priors, but a reasonable class of priors.*" This means that each single member of the class is not a reasonable model for prior ignorance, because no single distribution can model ignorance satisfactory. But the whole class, understood through the upper and lower probabilities is a reasonable model for prior ignorance. When we have little apriori information, $\overline{P}(A)$ should be near 1 and $\underline{P}(A)$ should be near 0, which means that we will never bet on or against A.

The bounded derivative model, defined by Walley (1997) is a model for prior ignorance about a one-dimensional parameter $\theta \in \Theta$. The model is defined as follows:

The bounded derivative model $\mathcal{M}$ *is the convex set of all pdfs* f *satisfying*

1. f *is continuous everywhere in* Θ*.*
2. $f(\theta) > 0 \quad \forall \theta \in \Theta$*.*
3. f *is differentiable and* $|(\ln f)'(\theta)| \leq c$*,*
   i.e. $|f'(\theta)| \leq cf(\theta)$ *for almost all* $\theta \in \Theta$*.*

All members of the class $\mathcal{M}$ are proper distributions. There is no general rule to chose the constant c, but it should be chosen large enough for the model to be highly imprecise, and small enough to produce useful inference.

The model is invariant only under linear transformations. Therefore, we should chose an appropriate transformation of the parameter space such that transformed parameter space is the whole real line. This might sometimes be difficult. We see that the model is not invariant under reparametrization. However, this is not as serious a problem for this model as it is for the uniform distribution, see Walley (1997).

A basic property of the model is this inequality, which holds for any $f \in \mathcal{M}$ and real numbers θ_0 and θ:

$$f(\theta_0)\exp(-c|\theta - \theta_0|) \leq f(\theta) \leq f(\theta_0)\exp(c|\theta - \theta_0|).$$

From this property the following theorem can be proved, which is central for making inference.

Theorem 1. *Assume the likelihood function* L *satisfies*

$$\int_{-\infty}^{\infty} \exp(c|\theta|)L(\theta)d\theta < \infty.$$

Then, for any non-decreasing function $Y : \mathcal{R} \to \mathcal{R}$*, the posterior lower and upper mean of* Y *are given by*

$$\underline{P}(Y|x) = \frac{\int_{-\infty}^{\infty} Y(\theta)\exp(-c\theta)L(\theta)d\theta}{\int_{-\infty}^{\infty} \exp(-c\theta)L(\theta)d\theta}$$

and

$$\overline{P}(Y|x) = \frac{\int_{-\infty}^{\infty} Y(\theta)\exp(c\theta)L(\theta)d\theta}{\int_{-\infty}^{\infty} \exp(c\theta)L(\theta)d\theta}.$$

This theorem can also be generalized to general functions Y. Refer to Walley (1997) for examples on use of the bounded derivative model.

5. Problems with Construction and Applications

In this section we discuss various problems related to construction and application of the noninformative priors described in Section 3. Many of the problems are related to the use of improper noninformative priors.

5.1. **Sample space dependence.** The first problem to be discussed, is sample space dependence. The problem is illustrated by an example.

Example 4. *Let θ be the proportion of successes in a Bernoulli population. Then θ can be estimated in two ways, by observing either y: the number of successes in n trials. The distribution of Y is $\mathrm{Bin}(n, \theta)$. Or we can observe z: the number of trials until r successes. The distribution of Z is $\mathrm{Neg}(r, \theta)$. We will find a noninformative prior for θ. By observing y, Jeffreys' prior is $\pi(\theta) \propto [\theta(1-\theta)]^{-1/2}$. By observing z, Jeffreys' prior is $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$.*

We see that the choice of noninformative prior depends on the sample space. This is also a violation of the likelihood principle, which says that problems with proportional likelihoods should result in the same inference.

5.2. **The marginalization paradox.** This problem is related to the use of improper noninformative priors when the parameter space is multi-dimensional. We illustrate the marginalization paradox by an example from Stone and David (1972).

Example 5. *Let $x = \{x_1, \dots, x_n\}$ be iid $N(\mu, \sigma^2)$. The reference prior for (μ, σ) is $\pi(\mu, \sigma) = \sigma^{-1}$ which is independent of the ordering of the parameters. As said in the discussion of Example 3, this is also the right invariant Haar measure.*

We now assume that we are going to do inference about $\theta = \mu/\sigma$, so the posterior distribution for θ is needed. By using the given reference prior, the posterior for (θ, σ) is given by

$$\pi(\theta, \sigma|x) \propto \exp\Big(-\frac{n\theta^2}{2} + \frac{n\theta\bar{x}}{\sigma} - \frac{R^2}{2\sigma^2}\Big)$$

where $R^2 = \sum x_i^2$. By integrating out σ, we find the marginal posterior for θ,

$$\pi(\theta|x) \propto \exp(-\tfrac{1}{2}n\theta^2) \int_0^\infty \omega^{n-2} \exp\big(-\tfrac{1}{2}\omega^2 + r\theta\omega\big)\, d\omega,$$

where $r = n\bar{x}/R$. We see that the marginal posterior for θ, $\pi(\theta|x)$ is a function of r alone.

We also calculate the distribution of r given the parameters (μ, σ):

$$f(r|\mu, \sigma) \propto \exp(-\tfrac{1}{2}n\theta^2)(1 - r^2/n)^{(n-3)/2} \int_0^\infty \omega^{n-1} \exp\big(-\tfrac{1}{2}\omega^2 + r\theta\omega\big) d\omega$$

and we see that $f(r|\mu, \sigma)$ is a function of θ alone. By Bayes theorem we would expect to be able to find a marginal prior for θ, $\pi(\theta)$ such that

$$\pi(\theta|r) \propto f(r|\theta)\pi(\theta).$$

However, this is impossible. This is what is called the marginalization paradox.

To overcome this problem, we can use a reference prior relative to the ordered partition (θ, σ), which gives:

$$\pi(\theta, \sigma) = (2 + \theta^2)^{-1/2}\sigma^{-1}$$

The marginal posterior for θ is now given by

$$\pi(\theta|x) \propto (2 + \theta^2)^{-1/2}[\exp(-\tfrac{1}{2}n\theta^2) \int_0^\infty \omega^{n-1} \exp(-\tfrac{1}{2}\omega^2 + r\theta\omega) d\omega].$$

Bayes' theorem can now be used, with the marginal prior for θ equal to $(2 + \theta^2)^{-1/2}$.

This shows that no single noninformative prior is universally noninformative. When the parameter space is multi-dimensional, we should chose the noninformative prior in accordance with the inferential problem at hand.

5.3. **Other problems with improper priors.** As mentioned, the reason why the marginalization paradox occurs is that we use an improper prior. In this Subsection, some other problems with noninformative priors are shortly mentioned.

Strong inconsistency: The phenomenon is illustrated by the following example:

Example 6. *Let $x \sim N(\theta, \sigma^2)$, where σ^2 is known. Let B be the event that $|\bar{x}| \geq |\theta|$. From the sampling model we find that*

$$P(B|\theta) = \frac{1}{2} + \Phi(-2|\theta|\sqrt{n}/\sigma) > \frac{1}{2}$$

Since $P(B|\theta) > \frac{1}{2}$ for all values of θ, we conclude that $P(B) > \frac{1}{2}$.

The posterior distribution, using a uniform prior for μ gives

$$P(B|x) = \frac{1}{2} - \Phi(-2|\bar{x}|\sqrt{n}/\sigma) < \frac{1}{2}$$

Since $P(B|x) < \frac{1}{2}$ for all values of x, we conclude that $P(B) < \frac{1}{2}$.

We see that the sampling model and the posterior are inconsistent. This is referred to as *strong inconsistency.*

Inadmissibility: Another important problem is that improper priors can lead to inadmissible Bayes estimators. A Bayes estimator is inadmissible if there is another estimator with less or equal expected loss for all values of the parameter θ, and less expected loss for at least one value of θ. A well-known example is that the posterior mean using a uniform prior is an inadmissible estimator of θ under squared error loss if the number of observations n is greater than or equal to three.

Improper posteriors: In some cases improper priors can lead to improper posteriors. An example, taken from Kass and Wasserman (1996) is the hierarchical model

$$\begin{aligned} Y_i|\mu_i, \sigma &\sim N(\mu_i, \sigma^2) \\ \mu_i|\tau &\sim N(\mu, \tau^2) \end{aligned}$$

for $i = 1, \ldots, n$, and σ is known. A natural choice of prior is $\pi(\mu, \tau) = \tau^{-1}$, but this leads to an improper posterior.

5.4. **Stein's paradox.** As mentioned in the foregoing Subsections, many problems occur for improper priors. An idea, in order to overcome problems with improper priors is to use proper approximations to improper priors. Examples are normal distributions with large variance, or a uniform distribution on a compact set. However, this is not always a good solution, as this example, taken from Bernardo and Smith (1994), shows.

Example 7. *Let $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a random sample from $N_k(\boldsymbol{\mu}, \boldsymbol{I}_k)$. Let $t = \sum_i \bar{x}_i^2$ where $\bar{x}_i$ is the mean of the n observations from coordinate i. Since $\boldsymbol{\mu}$ is a location parameter, the natural choice of noninformative prior for $\boldsymbol{\mu}$ is the uniform distribution. We approximate $\pi(\mu_1, \ldots, \mu_k) = 1$ by the product of k normal densities with large variance. Inferences about $\phi = \sum_i \mu_i^2$ is desired. We shall see that in this case the prior strongly dominates the data. With the given choice of prior, the posterior of $n\phi$ is $\chi_k^2(nt)$ with*

$$\mathrm{E}(\phi|\boldsymbol{x}) = t + k/n \qquad \text{and} \qquad \mathrm{Var}(\phi|\boldsymbol{x}) = \frac{2}{n}[2t + k/n].$$

The sampling distribution for nt is $\chi_k^2(n\phi)$ with $\mathrm{E}(t|\phi) = \phi + k/n$. By setting $k = 100$, $n = 1$ and $t = 200$ we have that $\mathrm{E}(\phi|\boldsymbol{x}) \approx 300$ and $\mathrm{Var}(\phi|\boldsymbol{x}) \approx 32^2$.

The unbiased estimator based on sampling distribution is

$$\hat{\phi} = t - k/n \approx 100$$

which is far from the posterior mean.

If we instead use the reference prior for $\{\phi, \omega_1, \ldots, \omega_{k-1}\}$, where the ω's are nuisance parameters, the marginal prior for ϕ is $\pi(\phi) = \phi^{-1/2}$. With this choice of prior, the posterior for ϕ is

$$\pi(\phi|\boldsymbol{x}) \propto \phi^{-1/2}\chi^2(nt|k, n\phi)$$

with mode close to $\hat{\phi}$.

Again we see that the prior should be chosen according to the inference problem at hand. In addition the example illustrates two problems. The first one is that noninformative priors can dominate the data, and the second is that proper approximations to noninformative priors do not solve all problems with noninformative priors.

6. Some Open Problems

An important question is, when does noninformative priors lead to proper posteriors? There are no general rules for finding out this.

An even more important problem is, how is one to know whether a posterior is data dominated? Some solutions to this problem are discussed by Kass and Wasserman (1996), but they all have some disadvantages.

Finally it should be noted that in many models, it is difficult to compute the prior. We have seen situations where it is simple, but for other situations, such as non-normal hierarchical models, it may not be clear how to compute the prior.

References

Berger, J. (1980). *Statistical Decision Theory.* Springer-Verlag.

Berger, J. and J. Bernardo (1989). Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statistical Association 89*, 200–207.

Bernardo, J. (1979). Reference posterior distributions for Bayesian inference. *J. R. Stat. Soc. B 41*, 113–147.

Bernardo, J. and A. Smith (1994). *Bayesian Theory.* John Wiley & Sons.

Box, G. and G. Tiao (1973). *Bayesian Inference in Statistical Analysis.* John Wiley & Sons.

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

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

Kass, R. and L. Wasserman (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association 91*(431), 1343–1370.

Pericchi, L. and P. Walley (1991). Robust bayesian credible intervals and prior ignorance. *Int. Statist. Rev. 58*(1), 1–23.

Robert, C. (1994). *The Bayesian Choice.* Springer-Verlag.

Stone, M. and A. David (1972). Un-Bayesian implications of improper Bayes inference in routine statistical problems. *Biometrika 59*(2), 369–375.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika 76*(3), 604–608.

Walley, P. (1997). A bounded derivative model for prior ignorance about a real-valued parameter. *Scand. J. of Stat. 24*, 463–483.

Welch, B. and H. Peers (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. R. Stat. Soc. B 25*, 318–329.