Journal of Statistical Planning and Inference 41 (1994) 37–60
North-Holland

Jeffreys' prior is asymptotically least favorable under entropy risk

Bertrand S. Clarke

Department of Statistics, The University of British Columbia, 2021 West Mall, Vancouver, BC, Canada V6T 122

Andrew R. Barron*

Yale University, New Haven, CT, USA

Received 1 December 1992; revised manuscript received 8 September 1993

Abstract

We provide a rigorous proof that Jeffreys' prior asymptotically maximizes Shannon's mutual information between a sample of size n and the parameter. This was conjectured by Bernardo (1979) and, despite the absence of a proof, forms the basis of the reference prior method in Bayesian statistical analysis. Our proof rests on an examination of large sample decision theoretic properties associated with the relative entropy or the Kullback–Leibler distance between probability density functions for independent and identically distributed random variables. For smooth finite-dimensional parametric families we derive an asymptotic expression for the minimax risk and for the related maximin risk. As a result, we show that, among continuous positive priors, Jeffreys' prior uniquely achieves the asymptotic maximin value. In the discrete parameter case we show that, asymptotically, the Bayes risk reduces to the entropy of the prior so that the reference prior is seen to be the maximum entropy prior. We identify the physical significance of the risks by giving two information-theoretic interpretations in terms of probabilistic coding.

AMS Subject Classification: Primary 62C10, 62C20; secondary 62F12, 62F15.

Key words: Bayes risk; minimax risk; Kullback–Leibler information; Jeffreys' prior; Fisher information; Shannon's mutual information; parametric density estimation; data compression; reference priors, least favorable priors.

1. Introduction

In Bayesian statistics much attention has been focussed on how to choose a prior. Various criteria have been proposed. For instance, Jeffreys (1961), George and McCulloch (1989), and Chang and Eaves (1990), amongst others, have advocated

Correspondence to: Dr. B.S. Clarke, Department of Statistics, The University of British Columbia, 2021 West Mall, Vancouver, BC, Canada V6T 122.

*Supported, in part, by the Office of Naval Research, contract N 00014-89-J-1811.

0378-3758/94/$07.00 © 1994—Elsevier Science B.V. All rights reserved.
SSDI 0378-3758(93)E0083-S

invariance conditions; matching frequentist coverage probabilities has been proposed by various authors including Welch and Peers (1966), Tibshirani (1989), and Bickel and Ghosh (1990). Others such as Bernardo (1979), Berger and Bernardo (1989, 1991, 1992a, b), Berger et al. (1989), and Polson and Wassermann (1989) have advocated optimizing functionals based on information-theoretic quantities. Also, Hartigan (1975) has examined asymptotic bias as a functional to be optimized. The earliest information-theoretic approach (Bernardo, 1979) advocated a criterion based on Shannon's mutual information, a special case of the relative entropy, so as to obtain what he called a 'reference prior', i.e. a prior against which alternative priors should be judged. This paper is intended to be a contribution to the development of the reference prior method.

Information-theoretic methods are attractive because, in the context of probabilistic coding, relative entropies have a well-defined role. They can be used to characterize the supremal rate of transmission for information-theoretic channels and to identify the redundancy of a noiseless source code. As a result, Bernardo's relative entropy criterion that we examine here can be given a physical interpretation.

Bernardo (1979) distinguished between the case where all parameters are of interest and the case where nuisance parameters are present. In the first of these he used a heuristic argument to support his conjecture that for smooth parametric families Jeffreys' prior maximizes an asymptotic expression for the Shannon mutual information, $I(\Theta; X^n)$. We give a rigorous justification for his conjecture: We prove that Jeffreys' prior is the unique continuous prior for which the Bayes strategy achieves the asymptotically maximum Bayes risk with relative entropy loss (Theorem 1). This maximin risk coincides with the minimax risk. Thus, we prove that Jeffreys' prior is asymptotically least favorable in smooth finite-dimensional parametric families in a formal decision-theoretic sense. Consequently, we can identify an asymptotically minimax estimator and its risk, an asymptotically minimax code and its redundancy, and the distribution achieving the capacity of certain channels. Also, obtaining the least favorable prior is useful because its density indicates which values of the parameter are the hardest to estimate.

This approach is in contrast to Jeffreys' original motivation which was based on invariance considerations; see Jeffreys (1961, Section 3.10). There, he observed that $(\det I(\theta))^{1/2}$, where $I(\theta)$ is the Fisher information matrix, is the Jacobian of the transformation of the parameter space that makes Hellinger and relative entropy distances locally Euclidean. This led him to propose $w^*(\theta) = (\det(I(\theta))^{1/2}/c$ as a choice of prior on the basis of its invariance under reparametrization.

Bernardo's framework for identifying reference priors is extended to problems with nuisance parameters in Berger and Bernardo (1989, 1992a, b). In that context, the results here have already been used to provide a formal justification for the prior they use in the presence of nuisance parameters; see Ghosh and Mukerjee (1991).

The same criterion can be applied when the parameter is assumed to take values in a discrete space. In this case we show that the relative entropy criterion reduces to maximum entropy.

Jeffreys' prior has been demonstrated to have other desired properties. Using a decision-theoretic formulation of unbiasedness, Hartigan (1965) demonstrated that, for loss functions with constant second derivatives, Jeffreys' prior is mean unbiased (Hartigan, 1965, p. 1139, also see Theorem 3) when the parameter space is a subset of the real line. Also, in the unidimensional case, Welch and Peers (1966) demonstrated that Jeffreys' density gives one-sided credible regions which match confidence intervals more closely than do the intervals from any other prior; see Hartigan (1983).

Turning to formalities, we assume we are given a parametric family of probability density functions $\{p_\theta: \theta\in\Omega\}$, $\Omega\subset\mathbb{R}^d$, $\theta=(\theta_1,\ldots,\theta_d)$, with respect to a fixed dominating measure $\lambda(\mathrm{d}x)$ on a separable metric space X, with probability measures assumed to be defined on the Borel subsets of X. We denote the density of n independent outcomes $x^n=(x_1,\ldots,x_n)$ by $p_\theta^n(x^n)=\prod_{i=1}^n p_\theta(x_i)$. Let $D(p\|q)$ denote the relative entropy or the Kullback–Leibler distance, which for densities p and q is defined to be

$$D(p\|q)=E_p\log\frac{p(X)}{q(X)}.$$

The main quantity of interest here is the relative entropy $D(p_\theta^n\|q_n)$, between the density functions $p_\theta^n(x^n)$ and an arbitrary joint probability density function $q_n(x^n)$, with respect to the same dominating measure $\lambda^n(x^n)$.

A game-theoretic interpretation is that one player, Nature, picks $\theta\in\Omega$ and assigns the joint density p_θ^n for each n, while a second player, the Statistician, chooses q_n for each n. Then, the relative entropy $D(p_\theta^n\|q_n)$ can be regarded as the risk to the Statistician or, in game-theoretic terminology, the 'payoff' to Nature. For prior probability density functions $w(\theta)$, $\theta\in\Omega$ with respect to the Lebesgue measure on $\mathbb{R}^d$, the Bayes strategy, which is to minimize $\int_\Omega w(\theta)D(p_\theta^n\|q_n)\,\mathrm{d}\theta$ over densities q_n, is achieved by choosing $q_n(x^n)=m_n^w(x^n)$; see Aitchison (1975), where $m_n^w(x^n)=\int_\Omega p_\theta^n(x^n)w(\theta)\,\mathrm{d}\theta$. For general prior distributions $w(\mathrm{d}\theta)$ on Ω, the definitions are the same, with integration with respect to $w(\mathrm{d}\theta)$ in place of $w(\theta)\,\mathrm{d}\theta$.

We obtain the asymptotics associated with the Bayes strategy. The quantities that we examine in this paper include the risk of the Bayes strategy, which, for priors w supported on a compact subset K in the interior of Ω is

$$R_n(\theta, w)=D(p_\theta^n\|m_n^w), \tag{1.1}$$

its corresponding Bayes risk,

$$R_n(w)=\int_K R_n(\theta, w)w(\theta)\,\mathrm{d}\theta, \tag{1.2}$$

and the minimax value, for θ in K

$$R_n = \inf_{q_n} \sup_{\theta \in K} D(p_\theta^n \| q_n). \tag{1.3}$$

where the infimum is over all probability densities of n on X^n.

The quantity $D(p_\theta^n \| m_n^w)$ can be given a statistical interpretation also. It is the cumulative risk of a sequence of Bayes estimators. Indeed, let $\hat{p}_k(x)$ be the predictive density given by

$$\hat{p}_k(x) = m_n^w(X_k = x | X^{k-1}) = m_k^w(X^k)/m_{k-1}^w(X^{k-1})$$

for $k=2,\ldots,n$. For $n=1$, $\hat{p}_1(x)=m_1^w(x)$. Then, as in Aitchison (1975) or Clarke and Barron (1990), it is seen that $\hat{p}_k$ is the Bayes estimator of the density of X_k based on X^{k-1}, under relative entropy loss. Since the expression in (1.1) may be written as

$$D(p_\theta^n \| m_n^w) = \sum_{k=1}^{n} E D(p_\theta \| \hat{p}_{k-1}),$$

the sum of the risks for each outcome, the quantities $R_n(\theta, w)$, $R_n(w)$, and R_n, may be interpreted as the cumulative risk of the Bayes estimator sequence, the cumulative Bayes risk of the Bayes estimator sequence, and the cumulative minimax risk, in an on-line estimation context.

In this paper we have three main goals. The first is to give asymptotic formulae for (1.2) and (1.3) for the continuous and discrete parameter cases. The second is to find the asymptotically least favorable prior corresponding to the maximin risk. The third is to give an information-theoretic interpretation for the quantities (1.1), (1.2) and (1.3).

When θ is a continuous parameter we show, formally, that the risk of the Bayes estimator, $R_n(\theta, w) = D(p_\theta^n \| m_n^w)$, satisfies the asymptotic expression

$$R_n(\theta, w) = \frac{d}{2} \log \frac{n}{2\pi e} + \log \det I(\theta) + \log \frac{1}{w(\theta)} + o(1), \tag{1.4}$$

in which the error, $o(1)$, tends to zero as $n \to \infty$, uniformly on compact sets in the interior of the support of the prior. The pointwise validity of (1.4) was verified in Clarke and Barron (1990). Here we use the uniformity of (1.4) over compact sets as the main tool for deriving the decision-theoretic asymptotics of (1.2) and (1.3).

The Bayes risk, $R_n(w)$, is obtained by averaging the risk $R_n(\theta, w)$ with respect to the prior w. It is seen that this Bayes risk is the relative entropy distance between the joint density $w(\theta)p_\theta(X^n)$ and the product of marginals $w(\theta)m_n^w(X^n)$. This latter quantity is Shannon's mutual information $I(\Theta; X^n)$ between the parameter Θ and the sample $X_1, \ldots, X_n$ which may be interpreted as an average relative entropy distance between the densities p_θ^n and m_n^w. Equivalently, application of Bayes rule shows that it is also the average relative entropy distance between $w(\theta | X^n)$ and $w(\theta)$, the posterior and prior densities for Θ. Provided (1.4) is valid it is seen that the asymptotic expression we obtain for Bayes risk is

$$R_n(w) = \frac{d}{2} \log \frac{n}{2\pi e} + \frac{1}{2} \int_K w(\theta) \log \det I(\theta)\, d\theta + H(w) + o(1), \tag{1.5}$$

where $H(w)=\int w(\theta)\log(1/w(\theta))\,d\theta$ is the entropy of the prior density w, and $o(1)\to 0$ as $n\to\infty$. This actually holds more generally; see Ibragimov and Hasminsky (1973) and Efroimovich (1980).

The maximum of this mutual information over choices of prior distribution is denoted as

$$R_n^* = \sup_w R_n(w). \tag{1.6}$$

where the supremum is over all distributions (discrete and continuous) supported on the given $K\subset\Omega$. Since $R_n(w)$ is the Bayes risk of the Bayes estimator, R_n^* is the maximin risk. The reference prior method of Bernardo is to choose that prior w^* which achieves the maximum mutual information in (1.6), or achieves the maximum in an asymptotic expression for $R_n(w)$. Typically, the choice of w^* that (asymptotically) maximizes the information is in fact the (asymptotically) least favorable prior with respect to the relative entropy loss, i.e. the reference prior and the least favorable prior are the same. Indeed, for the models p_θ treated here, the minimax and the maximin values agree for finite n, $R_n^*=R_n$. This equality follows from Davisson and Leon-Garcia (1980, Theorem 3). Moreover, when n is finite, Berger et al. (1989) and Zhang (1994) have demonstrated that the least favorable prior typically is discrete, i.e. for finite n, the supremum is achieved by a discrete prior, w_n, for which $R_n(w_n)=R_n^*$.

By contrast, as n increases, we show that the asymptotically minimax risk is achieved by Jeffreys' prior, $w^*(\theta)=(\det I(\theta))^{1/2}/c$, which is continuous, where $c=\int_K \sqrt{\det I(\theta)}\,d\theta$. The asymptotic form for the minimax risk is, by using (1.4),

$$\inf_{q_n}\sup_{\theta\in K} D(p_\theta^n\|q_n)=\frac{d}{2}\log\frac{n}{2\pi e}+\log\int_K\sqrt{\det I(\theta)}\,d\theta+o(1). \tag{1.7}$$

In addition, Jeffreys' prior is the unique asymptotically least favorable continuous positive prior. Moreover, although sequences of discrete priors can achieve the same asymptotic maximin value, no prior w or sequence of priors $\langle w_n\rangle|_{n=1}^{\infty}$ can achieve an asymptotically higher value for the information $R_n(w)$ than is achieved by Jeffreys' prior. This means that Jeffreys' prior is globally least favorable, but not necessarily uniquely so. Our results only give uniqueness among positive continuous priors.

When the parameter takes discrete values we can obtain asymptotic formulae for (1.1) and (1.2) which are analogous to (1.4) and (1.5). We show that for discrete θ

$$R_n(\theta,w)=D(p_\theta^n\|m_n)=\log\frac{1}{w(\theta)}+o(1), \tag{1.8}$$

and then that

$$R_n(w)=\sum_\theta w(\theta)R_n(\theta,w)=H(w)+o(1). \tag{1.9}$$

If the parameter space is finite, then the asymptotically least favorable distribution is uniform over the values of Θ. If the parameter space is noncompact and no constraint

is imposed on the class of priors then, as is typical in the continuous case, the maximum entropy is infinite. Maximizing the entropy in (1.9) is independent of the parametric family and depends only on the support of the prior. This is different from the result of the optimization in (1.7) which gives a prior proportional to $(\det I(\theta))^{1/2}$ and so depends on the parametric family. (Attempts to deal with an appropriate choice of reference prior in the discrete nonfinite case are in Berger et al. (1989) and Berger and Bernardo (1991).)

It can readily be verified that our hypotheses are satisfied in many examples. In the continuous case there are three main hypotheses. They are expected supremum conditions, that the Fisher information is a positive real number on the parameter space, and that the parametrization of the family of densities is one-to-one. As a consequence, it can be seen that our results apply to many commonly occurring parametric families including the Normal (μ, σ^2), the Gamma (p, λ), the Binomial (n, p), and the Poisson (λ). For these families our results demonstrate that when the parameter space is restricted to a compact set, Jeffreys' prior is asymptotically least favorable under relative entropy loss and that Jeffreys' is the reference prior.

In the information theory context of universal data compression, the quantities $R_n(\theta, w)$, $R_n(w)$, and R_n can be interpreted as the redundancy, average redundancy, and minimax redundancy of universal source codes; see Davisson (1973) and Csiszar (1993). Krichevsky and Trofimov (1981) studied minimax redundancy in the multinomial case, obtaining $R_n = \frac{1}{2} d \log n + o(1)$ as its asymptotic expression. Rissanen (1986, 1987) showed that the redundancy $R_n(\theta, w)$ equals $\frac{1}{2} d \log n + o(\log n)$ for smooth parametric families. Here, we extend the more exact asymptotics for $R_n(\theta, w)$ derived in Clarke and Barron (1990) to give the asymptotics for the average and minimax redundancies $R_n(w)$ and R_n.

Also in the information theory context, the characterization of $R_n(w)$ as a special case of Shannon's mutual information $I(\Theta; X^n)$ leads to implications for channel coding with one sender and many receivers. In this case R_n^* is the capacity of the channel and w^* is the source distribution which achieves the capacity of the channel.

The outline for the remainder of this paper is as follows. In Section 2 we formally state our main results which are subsequently proved in Sections 3 and 4. In Section 5 we give the information-theoretic and statistical interpretations of the quantities we have examined.

2. Formal statements of conditions and main results

So as to facilitate the statement of our main results we give a list of conditions to which it will be convenient to refer to.

Expectations, E or E_θ, are taken with respect to p_θ unless denoted otherwise. We denote the density of the mixture distribution by $m = m_n = m_{n,w}$, where W is the probability with density w. A similar notation is used for the mixture distribution. We

regard the relative entropy as a function of probabilities rather than densities and write $p(X|\theta)$ for $p_\theta(X)$ when convenient. Also, we assume the parameter space Ω has nonvoid interior and that its boundary has d-dimensional Lebesgue measure zero.

Condition 1. The density $p(x|\theta)$ is twice continuously differentiable in θ for almost every x, and there is a $\delta=\delta(\theta)$ so that for each j,k from 1 to d

$$E \sup_{\{\theta' \| \theta'-\theta\|<\delta\}} \left| \frac{\partial^2}{\partial\theta'_j \partial\theta'_k} \log p(X|\theta') \right|^2$$

is finite and continuous as a function of θ and for each j from 1 to d

$$E\left| \frac{\partial}{\partial\theta_j} \log p(X|\theta) \right|^{2+\xi}$$

is finite and continuous as a function of θ.

There are two information matrices, which typically coincide, which we use here. One is the Fisher information which we take to be defined by

$$I(\theta)=E\left[\frac{\partial}{\partial\theta_j} \log p(X|\theta) \frac{\partial}{\partial\theta_k} \log p(X|\theta) \right]_{j,k=1,\cdots,d},$$

and the other is the second derivative of the relative entropy

$$J(\theta)=\left[\frac{\partial^2}{\partial\theta'_j \partial\theta'_k} D(p_\theta \| p_{\theta'})|_{\theta'=\theta} \right]_{j,k=1,\cdots,d}.$$

When Condition 1 is satisfied the relative entropy is twice continuously differentiable and $J(\theta)$ is seen to equal the matrix with entries $-E_\theta(\partial^2/\partial\theta_i\partial\theta_j)\log p(X|\theta)$.

Condition 2. $I(\theta)$ is positive definite and coincides with $J(\theta)$.

Under Condition 1, Condition 2 will be satisfied provided that $\int(\partial^2/\partial\theta_i\partial\theta_j)\times p(X|\theta)\lambda(dx)=0$. See, for instance, Lehmann (1983, Lemma 2.6.1).

We next give a condition on the parametrization of the parametric family.

Condition 3. The parametrization of the family $\{p_\theta\}$ is one-to-one, i.e. for $\theta\neq\theta'$ we have that the corresponding probabilities P_θ and $P_{\theta'}$ are distinct.

The next condition is used for the results (2.1), (2.2) and (2.3) that require a probability density $w(\theta)$. It is not required for the results (2.4) and (2.5) that involve optimization over choices of prior W or distributions q_n. Fix a compact set K in the interior of Ω.

Condition 4. The prior w is positive, continuous and supported on K.

For the sake of completeness we state a result from Clarke and Barron (1990) which summarizes the asymptotics for the risk of the Bayes estimator, (Theorem 0 is stated there more generally, for possibly noncompactly supported priors under a soundness condition.)

Theorem 0. *Assume that Conditions* 1–4 *are satisfied. Then, for each* θ *in the interior of the support of* w,

$$\log \frac{p_\theta(X^n)}{m(X^n)} + \frac{1}{2} S_n^t J(\theta)^{-1} S_n - \frac{d}{2} \log \frac{n}{2\pi e} \to \log \frac{1}{w(\theta)} + \frac{1}{2} \log \det J(\theta), \quad (2.1a)$$

in P_θ^n *probability and in* $L_1(P_\theta)$ *as* $n \to \infty$, *where* $S_n = (1/\sqrt{n}) \nabla \log p(x^n|\theta)$. *Consequently,*

$$R_n(\theta, w) = \frac{d}{2} \log \frac{n}{2\pi e} + \frac{1}{2} \log \det I(\theta) + \log \frac{1}{w(\theta)} + o(1). \quad (2.1b)$$

Equation (2.1b) guarantees that (1.4) is true pointwise in θ. The first theorem extends that result by giving the minimax and maximin asymptotics for continuous parameters taking values in a compact parameter space.

Theorem 1. *Assume that Conditions* 1–4 *are satisfied. Then the risk of the Bayes strategy satisfies the following asymptotics uniformly on compact subsets in the interior of the support of* w, *i.e., for each* K_0 *in the interior of* K

$$\lim_{n \to \infty} \sup_{\theta \in K} \left| R_n(\theta, w) - \left(\frac{d}{2} \log \frac{n}{2\pi e} + \log \frac{\sqrt{\det I(\theta)}}{w(\theta)} \right) \right| = 0. \quad (2.2)$$

Again, assume Conditions 1, 2, 3, *and* 4. *The Bayes risk of the Bayes estimator satisfies*

$$\lim_{n \to \infty} \left| R_n(w) - \left(\frac{d}{2} \log \frac{n}{2\pi e} + \int_K w(\theta) \log \frac{\sqrt{\det I(\theta)}}{w(\theta)} d\theta \right) \right| = 0. \quad (2.3)$$

Now, assume only, that Conditions 1, 2, *and* 3 *are satisfied. The asymptotic minimax risk* $R_n = \inf_{Q_n} \sup_{\theta \in K} D(P_\theta^n \| Q_n)$ *satisfies*

$$\lim_{n \to \infty} \left[R_n - \frac{d}{2} \log \frac{n}{2\pi e} \right] = \log \int_K \sqrt{\det I(\theta)}\, d\theta, \quad (2.4)$$

and, the maximin risk $R_n^* = \sup_w R_n(w)$ (*where the supremum is over all distributions supported on* K) *has the same asymptotics*:

$$\lim_{n \to \infty} \left[R_n^* - \frac{d}{2} \log \frac{n}{2\pi e} \right] = \log \int_K \sqrt{\det I(\theta)}\, d\theta \quad (2.5)$$

Jeffreys' prior, $w^*(\theta) = \sqrt{\det I(\theta)}/c$ *with* $c = \int_K \sqrt{\det I(\theta)}\, d\theta$, *is the unique continuous and positive prior on* K *for which the Bayes strategy achieves the asymptotic maximin value, i.e.*

$$\lim_{n \to \infty} \left[R_n(w^*) - \frac{d}{2} \log \frac{n}{2\pi e} \right] = \log \int_K \sqrt{\det I(\theta)}\, d\theta. \quad (2.6)$$

No other prior w, or sequence of priors $\langle w_n \rangle|_{n=1}^{\infty}$, discrete or continuous, achieves an asymptotically larger value of $R_n(w)$ than does Jeffreys' prior.

Remark. To show that w^* has the indicated optimally properties it would suffice by standard decision-theoretic arguments to upper bound the maximum risk $\sup_{\theta \in K} R_n(\theta, w)$ by $\frac{1}{2} d \log[n/2\pi e] - \log[w(\theta)/\sqrt{\det I(\theta)}] + o1$, at least for $w = w^*$, and to lower bound the Bayes risk $R_n(w)$ by the average of the same asymptotic expression. The conditions required may be less than those used here to get the uniformity of the asymptotics of the risk $R_n(\theta, w)$; cf. Ibragimov and Hasminski (1973) and Efroimovich (1980) for the asymptotics of $R_n(w)$ under weaker conditions.

The asymptotic expression for the mutual information for continuous priors takes the form

$$\lim_{n \to \infty} \left[R_n(w) - \frac{d}{2} \log \frac{n}{2\pi e} \right] = \log c - D(w \| w^*), \tag{2.7}$$

where w^* is Jeffreys' prior. Therefore, it is apparent from (2.7) that the unique (continuous) prior maximizing the limit of the mutual information is Jeffreys' prior (since $D(w \| w^*) \geq 0$ with equality if and only if $w = w^*$). Recall that if a Bayes procedure has (asymptotically) constant risk then it is (asymptotically) minimax. Indeed, to complete the demonstration of the desired asymptotic properties it would be enough to show that $R_n(\theta, w^*) - (d/2) \log (n/2\pi e)$ is asymptotically constant for $\theta \in K$ with value $\log c$. The result of (2.2) gives this desired uniformity for compact sets in the interior of K. However, on the boundary of K, the behavior of the risk $R_n(\theta, w^*)$ may be different from what (2.1b) suggests. Nevertheless, in the proof we identify the asymptotic minimax value. The trick is to bound the risk using Jeffreys' prior on sets K' slightly larger than K. The asymptotic minimax value is the same as the Bayes risk achieved by Jeffreys' prior on K. Consequently, Jeffreys' prior is asymptotically least favorable and no sequence of priors has an asymptotically larger Bayes risk.

In noncompact cases lower bounds on the maximin risk can be obtained by restricting the application of Theorem 1 to compact subsets of the parameter space. When $\int_{\Omega} \sqrt{\det I(\theta)}\, d\theta$ is infinite, the value of $\lim_{n \to \infty} R_n^* - \frac{1}{2} d \log n$ is seen to be infinite by applying Theorem 1 to a sequence of compacta with divergent $\int_K \sqrt{\det I(\theta)}\, d\theta$.

It is possible to extend the present result to include parametrized families of densities whose parameter does not vary a compact set contained in the interior of the parameter space. In particular, Clarke and Barron (1991) give conditions which imply the convergence of the Bayes risk in (2.3) while permitting the parameter to vary over a set which may contain boundary points of the parameter space.

An examination of the proof of the results in the continuous case shows that the $\frac{1}{2}d\log n$ behavior comes out of a Laplace's method argument; see Clarke and Barron (1990, Lemma 4.1). This cannot work in the discrete case and the asymptotic form is a constant independent of n, to order o(1). Our result is the following.

Theorem 2. *Suppose that the members of the parametric family to which the prior distribution $W(\cdot)$ with density $w(\cdot)$ (with respect to counting measure) assigns positive mass are distinct and there is a Kullback–Leibler neighborhood of P_θ which excludes all other members of the parametric family which have positive mass under the prior. Then*

$$D(p_\theta \| M_n) = \log \frac{1}{w(\theta)} + o(1)$$

as $n \to \infty$.

Corollary 1. *Assume that the entropy of the prior is finite and that the parameter values in the support of the prior are isolated (are not the limits of other parameter points). Then*

$$R_n(w) = \sum_\theta D(p_\theta^n \| M_n) w(\theta) = H(w) + o(1)$$

as $n \to \infty$, where $H(w) = \sum_\theta w(\theta) \log 1/w(\theta)$.

Remark. In the case that the parameter space is infinite, the corollary is already well known and follows from Fano's inequality; see Blahut (1987).

Before giving rigorous details in Section 3, here we give the heuristic argument for the validity of the asymptotic expansion for $R_n(\theta, w)$. We use Laplace's method to approximate the integral $m(x^n) = \int_\Omega p(x^n|\theta')\mathrm{d}\theta'$, which defines Bayesian's marginal distribution for the data. Laplace's method is to apply a second-order Taylor expansion to $\log p(x^n|\theta)$ so as to reduce to an approximate Gaussian integration. The appropriate Taylor expansion is

$$\log \frac{p(x^n|\theta')}{p(x^n|\theta)} = \sqrt{n}(\theta' - \theta)^t S_n(\theta) - \tfrac{1}{2} n(\theta' - \theta)^t I^*(\theta^*)(\theta' - \theta),$$

where $S_n(\theta) = (1/\sqrt{n})\nabla \log p(x^n|\theta)$ and $I^*(\theta)$ is the empirical Fisher information with entries $-(1/n)(\partial^2/\partial\theta_j \partial\theta_i)\log p(x^n|\theta)$ and θ^* is a point on the line segment joining θ' and θ.

If a consistency argument is used to restrict the integration in m to a neighborhood of θ and to argue that the score $S_n(\theta)$ is near its mean value of zero and if the error in the replacement of $I^*(\theta^*)$ by the theoretical Fisher information $I(\theta)$ is ignored, then

$$\frac{m(x^n)}{p(x^n|\theta)} \sim w(\theta)(2\pi)^{d/2}\left(\sqrt{\det nI(\theta)}\right)^{-1} \mathrm{e}^{(1/2)S_n(\theta)I^{-1}(\theta)S_n(\theta)}. \tag{2.8}$$

Since

$$R_n(\theta, w) = \underset{\sim}{E} \log \frac{p(x^n|\theta)}{m(x^n)},$$

the expected logarithm of the right-hand side of (2.8) yields

$$\frac{d}{2}\log\frac{n}{2\pi e} + \frac{1}{2}\log\det I(\theta) + \log\frac{1}{w(\theta)} \tag{2.9}$$

as the desired asymptotic expression.

The proof of the validity of this expansion (pointwise in θ) in Clarke and Barron (1990) demonstrates that the sources of error in Laplace's method can be controlled when taking the expected logarithm. There it is shown (cf. (4.10), (4.11), p. 464) that for each positive ε, δ the remainder defined by $\mathrm{Rem}_n(\theta, w) = R_n(\theta, w) - \frac{1}{2}d\log[n/2\pi e] + \log\sqrt{\det I(\theta)}/w(\theta))]$ satisfies

$$\begin{aligned}\mathrm{Rem}_n(\theta, w) \geqslant\ & P_\theta^n((A_n \cap B_n)^c)\log P_\theta^n((A_n \cap B_n)^c) \\ & - P_\theta^n((A_n \cap B_n)^c)\left(\frac{d}{2}\log\frac{n}{2\pi e} + \log\frac{\sqrt{\det I(\theta)}}{w(\theta)}\right) \\ & - \frac{d\varepsilon}{2(1-\varepsilon)} + \frac{(d-1)}{2}\log(1-\varepsilon) - \rho(\delta, \theta)\end{aligned} \tag{2.10}$$

when $P_\theta^n((A_n \cap B_n)^c) \leqslant e^{-1}$, where $\rho(\delta, \theta) = \sup_{|\theta'-\theta|<\delta} |\log w(\theta')/w(\theta)|$ goes to zero as δ goes to zero, for $\theta \in \mathrm{Int}\,(\mathrm{support}(w))$ and the events A_n and B_n are defined by

$$A_n = A_n(\theta, \delta, \varepsilon) = \left\{ \int_{N_\delta^c} p(X^n|\theta) w(\theta)\, d\theta \leqslant \varepsilon \int_{N_\delta} p(X^n|\theta)\, w(\theta)\, d\theta \right\},$$

$$\begin{aligned}B_n = B_n(\theta, \delta, \varepsilon) = \{ & (1-\varepsilon)(\theta'-\theta)^t I(\theta)(\theta'-\theta) \leqslant (\theta'-\theta)^t I^*(\theta'')(\theta'-\theta) \\ & \leqslant (1+\varepsilon)(\theta'-\theta)^t I(\theta)(\theta'-\theta) \text{ for all } \theta', \theta'' \in N_\delta\},\end{aligned}$$

in which $N_\delta = \{\theta' : |\theta'-\theta| < \delta\}$ and the inner product defining the norm is with respect to $I(\theta)$, i.e. $|\theta-\theta'|^2$ is defined to be $(\theta-\theta')^t I(\theta)(\theta-\theta')$. Conditions 1 and 2 imply the equivalence of neighborhoods defined by the relative entropy $D(P_\theta \| P_{\theta'})$ and neighborhoods defined by $|\theta-\theta'|^2$. Indeed, it can be shown that there exist constants c' and δ_0 (possibly depending on K_0) such that

$$\{\theta' : |\theta-\theta'|^2 \leqslant \delta^2/c'^2\} \subset \{\theta' : D(P_\theta \| P_{\theta'}) \leqslant \delta^2\} \subset \{\theta' : |\theta-\theta'|^2 \leqslant \delta^2 c'^2\} \tag{2.11}$$

for $0 < \delta < \delta_0$ and all θ in K_0.

An analogous lower bound on the remainder is possible. We have from Clarke and Barron (1990, (4.12)–(4.15), p. 464),

$$\begin{aligned}\mathrm{Rem}_n(\theta, w) \leqslant & nP_\theta^n((B_n \cap C_n)^c) Eg(X, \theta, \delta) \\ & + (nP_\theta^n((B_n \cap C_n)^c))^{1/2} (Eg^2(X, \theta, \delta))^{1/2} \\ & + P_\theta^n((B_n \cap C_n)^c)\left(\frac{d}{2}\log\frac{n}{2\pi e} + \log\frac{\sqrt{\det I(\theta)}}{w(\theta)}\right) \\ & + P_\theta^n((B_n \cap C_n)^c)|\log W(N_\delta)| + ES_n(\theta)^t I(\theta) S_n(\theta) 1_{(B_n \cap C_n)^c} \\ & + \frac{d\varepsilon}{2(1-\varepsilon)} - \frac{d}{2}\log(1+\varepsilon) - \log(1 - 2^{d/2} e^{-\varepsilon^2 n\delta^2/8}) + \rho(\delta, \theta), \quad (2.12)\end{aligned}$$

where the event $C_n(\theta, \varepsilon)$ is defined by

$$C_n(\theta, \varepsilon) = \{S_n^t I^{-1}(\theta) S_n(\theta) \leqslant n\varepsilon\},$$

and the function $g(x, \theta, \delta)$ is defined by

$$g(x, \theta, \delta) = \sup_{\theta', \theta'' \in N_\delta} (\theta' - \theta)^t \nabla \log p(x \mid \theta'').$$

Theorem 1 will be established once the remainder terms in (2.10) and (2.12) are shown to be asymptotically negligible uniformly over K_0 in the interior of the support of w. In the course of the proof we will show that $P_\theta((A_n \cap B_n)^c) = o(1/\log n)$ uniformly on K_0, which implies the lower bound. For the upper bound we must obtain three results, namely that $P_\theta((B_n \cap C_n)^c) = O(1/n)$ uniformly on K_0, the Eg and Eg^2 tend to zero as $\delta \to 0$ and that $E_\theta S_n^t I(\theta) S_n(\theta) 1_{(B_n \cap C_n)^c} \to 0$ under the second part of Condition 1. We note that the slightly higher moment required there is not atypical since it provides uniform rates of convergence in the central limit theorem for S_n. The proofs presented here are a refinement of the methods used in the doctoral dissertation of Clarke (1989).

3. Proof of Theorem 1

We start by noting that $P_\theta(C^c(\theta, \varepsilon)) = O(1/n)$. Indeed, by applying Markov's inequality we have that

$$P_\theta(C^c(\theta, \varepsilon)) \leqslant \frac{1}{n\varepsilon} E_\theta S_n(\theta) I^{-1}(\theta) S_n(\theta) = \frac{d}{n\varepsilon}. \quad (3.1)$$

Next, we observe that $P_\theta(B^c(\theta,\delta,\varepsilon))=O(1/n)$. Following Clarke and Barron (1990, p. 465) we have that

$$P_\theta(B^c(\theta,\delta,\varepsilon)) \leqslant \sum_{j,k=1}^{d} P_\theta^n\left(\sup_{\theta'\in N_\delta} |I^*_{j,k}(\theta')-I_{j,k}(\theta)|>\varepsilon/d^2\right)$$

$$\leqslant \sum_{j,k=1}^{d}\left[P_\theta^n\left(\sup_{\theta'\in N_\delta}|I^*_{j,k}(\theta')-I^*_{j,k}(\theta)|>\varepsilon/2d^2\right)\right.$$

$$\left.+P_\theta^n(|I^*_{j,k}(\theta)-I_{j,k}(\theta)|>\varepsilon/d^2)\right]. \tag{3.2}$$

Each of the first terms in the summands of (3.2) is bounded by a function which is $O(1/(n\varepsilon^2))$, uniformly for $\theta\in K_0$. Indeed, by Chebyshev's inequality each is bounded by $(16d^2/(n\varepsilon^2))E((\sqrt{n}(\bar{Y}_{j,k}-EY_{j,k,1}))^2)=(16d^2/(n\varepsilon^2))\operatorname{Var}_\theta Y_{j,k,1}$ in which $\bar{Y}_{j,k}$ is the mean of $Y_{j,k,i}(\theta,\delta)=\sup_{|\theta'-\theta|<\delta}|(\partial^2/\partial\theta_j\partial\theta_k)\log p(X_i|\theta)-(\partial^2/\partial\theta_j\partial\theta_k)\log p(X_i|\theta')|$ and δ, depending on ε, is so small that $EY_{j,k,i}<\varepsilon/4d^2$. By Condition 1 and the compactness of K_0, $\operatorname{Var}_\theta Y_{j,k,1}$ is bounded uniformly for $\theta\in K_0$. The second terms in the summands of (3.2) can be bounded from above in a similar fashion. Since there are finitely many terms, expression (3.2) is $O(1/n\varepsilon^2)$, as required.

Next we deal with $P_\theta(A_n^c)$ by a more involved argument. Consider the set

$$U_n(\theta)=\left\{\int_{N_\delta} w(\theta')p(X^n|\theta')\,d\theta' > e^{-nr'}\,p(X^n|\theta)\right\}$$

for $r'>0$. For $r>0$ we have that $P_\theta(A_n^c)$ is bounded above by

$$P_\theta\left(\int_{N_\delta} w(\theta')p(X^n|\theta')\,d\theta < e^{nr}\int_{N_\delta^c} w(\theta')p(X^n|\theta')\,d\theta'\right)$$

$$\leqslant P_\theta\left(p(X^n|\theta)<e^{n(r+r')}\int_{N_\delta^c} w(\theta')p(X^n|\theta')\,d\theta'\right)$$

$$+P_\theta\left(e^{nr}\int_{N_\delta} w(\theta')p(X^n|\theta')\,d\theta' < p(X^n|\theta)\right) \tag{3.3}$$

by intersection with U_n and U_n^c. The second term is uniformly $O(1/n)$ on compact sets: this term is upper bounded by taking an integral over a smaller neighborhood of θ with δ replaced by $\delta_n=1/\sqrt{n}$. One can show that $W(N_{\delta_n})$, where W is the measure defined by w, is bounded below by a constant times $(1/n)^{d/2}$, uniformly for θ in K_0. Using this, one can apply Markov's inequality to the absolute value of the logarithm of the density ratio so as to obtain an upper bound on the second term of the form $2(D(P_\theta^n\,\|\,M_n(\cdot\,|N_{\delta_n}))+2e^{-1})/nr'$, for n large enough, where $M_n(\cdot\,|N_{\delta_n})$ is the distribution with density given by $m(x^n|N_{\delta_n})=\int_{N_{\delta_n}} w(\theta')p(x^n|\theta')\,d\theta'/W(N_{\delta_n})$. The relative

entropy in the numerator of the upper bound can be bounded from above by a constant c by a convexity argument. Now one has $2c(1+2e^{-1})/nr'$ as the upper bound on the second term, uniformly for θ in K_0. For details, see Clarke and Barron (1990, pp. 469–470). Now, it is enough to show that the first term on the right-hand side is of order $O(e^{-nr''})$ for some $r''>0$.

Let $G=\{F_1, F_2, \dots\}$ be a countable field of sets generated by balls of the form $\{x: d_X(x, s_j)\leqslant 1/k\}$, where d_X is the metric on the separable space $\boldsymbol{X}$, the sequence $s_1, s_2, \dots$ is a countable sequence of points which is dense in $\boldsymbol{X}$, and $j, k=1, 2, \dots$ Let

$$d_G(P, Q)=\sum_{i=1}^{\infty} 2^{-i}|P(F_i)-Q(F_i)|.$$

Here, d_G is a metric on the collection of probabilities on $\boldsymbol{X}$ for which $d_G(P_n, P)\to 0$ implies that P_n converges to P in distribution; see Gray (1988, pp. 251–253). Now let $G(\theta)=\{X^n | d_G(P_\theta, \hat{P})<\xi\}$. We have that there is an r_ξ so that $P_\theta(G(\theta)^c)\leqslant e^{-nr_\xi}$ and for θ' with $d_G(P_\theta, P_{\theta'})>2\xi$ that $P_{\theta'}(G(\theta))\leqslant e^{-nr_\xi}$, where r_ξ does not depend on θ. Indeed, by Hoeffding's inequality (Hoeffding, 1963) we have that for any probability P

$$P^n(X^n: d_G(P, \hat{P})>\xi)\leqslant P^n\left(X^n: \sum_{i=1}^{\infty} 2^{-i}|P(F_i)-\hat{P}(F_i)|>\xi\right)$$
$$\leqslant 2k_\xi e^{-n\xi^2/2} \tag{3.4}$$

for $k_\xi\geqslant 1+\log 2/\xi$, as shown in Clarke and Barron (1990, p. 469). Thus, $P_\theta(G^c(\theta))\leqslant 2k_\xi e^{-n\xi^2/2}$ and, when $d_G(P_{\theta'}, P_\theta)>2\xi$, the triangle inequality gives $P_{\theta'}(G(\theta))\leqslant P_{\theta'}(d_G(P_{\theta'}, \hat{P})>\xi)\leqslant 2k_\xi e^{-n\xi^2/2}$.

Now we can bound the first term on the right-hand side in (3.3). Intersecting the event in this term with $G(\theta)$ and $G(\theta)^c$ and applying Markov's inequality to the first of these two quantities gives

$$e^{n(r+r')}E_\theta\int_{N_\delta^c} 1_{G(\theta)}P_{\theta'}(G(\theta))\,d\theta+P_\theta(G(\theta)^c).$$

The second term is exponentially small by (3.4). The first term is bounded from above by

$$e^{n(r+r')}\int_{N_\delta^c} P_{\theta'}(G(\theta))\,d\theta'.$$

Choosing r and r' so small that $r+r'-r_\xi<0$, and using (3.4), gives that this last expression is exponentially small as well, provided N_δ contains $\{\theta': d_G(P_\theta, P_{\theta'})\leqslant 2\xi\}$.

Now we show that there is a $\delta=\delta_\xi$ independent of θ for which this containment is valid and such that $\delta_\xi\to 0$ when $\xi\to 0$. First we note that by Conditions 1 and 2, the map $\phi: \theta\to P_\theta$ is continuous. The continuous image of a compact set is again compact. So, when θ is restricted to lie in a compact set K_0, the corresponding collection of

probability measures $\phi(K_0)$ is compact in the topology of weak convergence. Furthermore, since ϕ is one-to-one (by Condition 3) as well as continuous, the restriction of ϕ to a compact set has a continuous inverse. Indeed, this inverse map is uniformly continuous since it is only defined on a compact set. That is, there is a uniformly continuous mapping that recovers θ from P_θ where continuity in the collection of probabilities is defined by weak convergence. Since convergence in d_G implies weak convergence, the existence of the desired δ_ξ follows. So, for given $\varepsilon>0$ we choose ξ so small that δ_ξ is not greater than the δ required for the bounding of the summands of (3.2). Thus, the first term on the right-hand side of (3.3) is exponentially small, so we have that $P_\theta(A^c)$ is $O(1/n)$.

From the bounds on $P_\theta(A_n^c)$ and $P_\theta(B_n^c)$ it is clear that the error terms in (2.10) vanish uniformly in θ on compact sets as $n\to\infty$ and then $\varepsilon\to 0$. Thus, the lower bound (2.10) goes to zero uniformly over K_0.

For the upper bound (2.12), we start by noting that the uniform bounds on $P_\theta(B_n^c)$ and on $P_\theta(C_n^c)$ imply

$$nP_\theta((B_n\cap C_n)^c)\leqslant\frac{C_K}{\varepsilon^2}, \tag{3.5}$$

where C_{K_0} is a constant depending on the compact set K_0. In the first two terms on the right-hand side of (2.12) we choose $\delta=\delta(\varepsilon)$ so small that for given $\varepsilon>0$ we have that

$$E_\theta g^2(X,\theta,\delta)<\varepsilon^3. \tag{3.6}$$

Then the first two terms on the right-hand side (2.12) are less than a constant times $\sqrt{\varepsilon}$ uniformly in θ on K_0. The third and fourth terms on the right-hand side of (2.12) converge to zero as $n\to\infty$ by using (3.5).

It remains to show that $E_\theta S_n(\theta)^t I(\theta) S_n(\theta) 1_{(B_n\cap C_n)^c}$ goes to zero uniformly for θ in K_0. By rearranging under the trace function, tr, we have

$$\begin{aligned}E_\theta S_n(\theta)^t I^{-1}(\theta) S_n(\theta) 1_{(B_n\cap C_n)^c} &= \operatorname{tr} I^{-1}(\theta) E_\theta\, E_\theta S_n(\theta)^t S_n(\theta) 1_{(B_n\cap C_n)^c}\\ &= \operatorname{tr} I^{-1}(\theta)\left[E_\theta\left(\frac{1}{n}\frac{\partial}{\partial\theta_i}\log p_\theta(X^n)\frac{\partial}{\partial\theta_j}\log p_\theta(X^n)\right)_{i,j} 1_{(B_n\cap C_n)^c}\right].\end{aligned}$$

We show that the (i,j)th entry in the matrix goes to zero. By the Cauchy–Schwartz inequality it is bounded above by

$$\sqrt{E_\theta\left((1/\sqrt{n})\left(\frac{\partial}{\partial\theta_i}\log p_\theta(X^n)\right)\right)^2 1_{(B_n\cap C_n)^c} E_\theta\left((1/\sqrt{n})\left(\frac{\partial}{\partial\theta_j}\log p_\theta(X^n)\right)\right)^2 1_{(B_n\cap C_n)^c}}. \tag{3.7}$$

The first expectation in (3.7) is bounded, by the Holder inequality, by

$$\left(E_\theta\left((1/\sqrt{n})\left(\frac{\partial}{\partial\theta_i}\log p_\theta(X^n)\right)\right)^{2(1+\varepsilon)}\right)^{1/(1+\varepsilon)} P_\theta((B_n\cap C_n)^c)^{\varepsilon/(1+\varepsilon)}. \tag{3.8}$$

The expectation in (3.8) is bounded. By Lemma 3.1 in Ibragimov and Hasminsky (1981, p. 186), it is bounded above by a constant (depending on $(1+\varepsilon)$) times $E_\theta((\partial/\partial\theta_i)\log p_\theta(X^n))^{2(1+\varepsilon)}$. The probability in (3.8) goes to zero by using (3.5). The second expectation in (3.7) is similar.

Thus, letting $n\to\infty$ followed by $\varepsilon\to 0$ where $\delta=\delta(\varepsilon)$ as prescribed shows that the upper bound (2.12) goes to zero uniformly on K. Thus, we have established that

$$\sup_{\theta\in K}|\mathrm{Rem}_n(\theta,w)|\to 0$$

as $n\to\infty$, and (2.2) is proved.

Next, we show how (2.2) implies the other statements of the theorem. For (2.3), recall that K equals the support of w and that (2.2) only holds for interior points of K. Consequently, we consider a continuous extension of w say w' which has support equal to K' a compact set which contains K in its interior; such an extension exists by the Tietze Extension Theorem, see Royden (1968). Now, for $\theta\in K$, we have $w'(\theta)=w(\theta)$ and by the continuity of w' there is a K' so that on $\theta\in K'-K$, w' is positive. Let the standardized form of w' be $w''(\theta)=w'(\theta)/\int_{K'}w'(\theta)\,\mathrm{d}\theta$. The denominator is of the form $1+\alpha$ where $\alpha=\int_{K'-K}w'(\theta)\,\mathrm{d}\theta$, and $W^*(K)=1/(1+\alpha)$ where W^* is the probability associated to the density w''.

Now, for w'' we can use the uniformity of (2.2), since K is a compact set in the interior of K'. Thus, we have that

$$\lim_{n\to\infty}\left|\int_K w(\theta)D(p_\theta^n||m_{w'',n})\,\mathrm{d}\theta-\left(\frac{d}{2}\log\frac{n}{2\pi e}+\int_K w(\theta)\log\frac{\sqrt{\det I(\theta)}}{w''(\theta)}\,\mathrm{d}\theta\right)\right|=0 \tag{3.9}$$

Consequently, to establish (2.3) it is sufficient to bound

$$\int_K w(\theta)D(p_\theta^n\,\|\,m_{w'',n})\,\mathrm{d}\theta-\int_K D(p_\theta^n\,\|\,m_{w,n})\,\mathrm{d}\theta \tag{3.10}$$

and

$$\int_K w(\theta)\log\frac{\sqrt{\det I(\theta)}}{w''(\theta)}\,\mathrm{d}\theta-\int_K w(\theta)\log\frac{\sqrt{\det \mathrm{I}(\theta)}}{w(\theta)}\,\mathrm{d}\theta. \tag{3.11}$$

The quantity in (3.10) is non-negative and equal to

$$\int_K w(\theta)(D(p_\theta^n\,\|\,m_{w'',n})-D(p_\theta^n\,\|\,m_{w,n}))\,\mathrm{d}\theta=D(m_{w,n}\,\|\,m_{w'',n})\leqslant\log(1+\alpha), \tag{3.12}$$

where the bound follows from noting that $m_w(x^n)/m_{w''}(x^n)$ can be written as $(1+\alpha)(\int_K w(\theta)p(x^n|\theta)\,\mathrm{d}\theta)/\int_{K'}w'(\theta)p(x^n|\theta)\,\mathrm{d}\theta$, which is less than $1+\alpha$ since w' is an extension of w.

The quantity in (3.11) is equal to

$$\int_K w(\theta)\log\frac{w(\theta)}{w''(\theta)}\,d\theta=\log(1+\alpha), \tag{3.13}$$

which tends to zero as $K'\to K$, because $\alpha=\int_{K'-K} w'(\theta)\,d\theta$. [Here it is assumed that we hold the continuous extension $w'(\theta)$ fixed while we let $K'\to K$, so that $\alpha\to 0$.]

The desired conclusion (2.3) for the asymptotic Bayes risk follows from these bounds, since the limit of

$$\left|\int_K w(\theta)D(p_\theta^n||m_{w,n})\,d\theta-\left(\frac{d}{2}\log\frac{n}{2\pi e}+\int_K w(\theta)\log\frac{\sqrt{\det I(\theta)}}{w(\theta)}\,d\theta\right)\right|$$

is bounded from above by the sum of the bounds from (3.9), (3.12), and (3.13). Letting $n\to\infty$ and then $K'\to K$ gives the claimed result.

For the determination of the minimax and maximin conclusions below, we will only need to use (2.3) for priors proportional to $\sqrt{\det I(\theta)}$, which is defined on Ω. For such priors the extension from K to K' is automatic so there is no need to appeal to an extension theorem.

For (2.4), the conclusion about the minimax value, we recall that the minimax risk is

$$R_n=R(n,K,\{P_\theta\})=\inf_{Q_n}\sup_{\theta\in K} D(P_\theta^n\|Q_n),$$

and for (2.5), (2.6) that the maximin risk is

$$\begin{aligned}R_n^*=R^*(n,K,\{P_\theta\})&=\sup_W\inf_{Q^n}\int_K D(P_\theta^n\|Q^n)\,W(d\theta)\\&=\sup_W\int_K D(P_\theta^n\|M_{n,W})\,W(d\theta),\end{aligned}$$

in which $M_{n,W}$ is the distribution with density $m_n(x^n)=\int p_\theta(x^n)\,W(d\theta)$. We can upper bound the minimax risk by replacing Q_n with any other estimator, the mixture distribution with respect to Jeffreys' prior for instance. We consider M_{n,w_K^*} in which $w_K^*(\theta)$ is given by standardizing $\sqrt{\det\ I(\theta)}$ on a compact set K' which contains K in its interior. So, we have that

$$R_n-\tfrac{1}{2}d\log n\leqslant\sup_{\theta\in K}[D(P_\theta^n\|M_{n,w_K^*})-\tfrac{1}{2}d\log n],$$

in which case the right-hand side is upper bounded by

$$\frac{d}{2}\log\frac{1}{2\pi e}+\log\int_{K'}\sqrt{\det I(\theta)}\,d\theta+o(1),$$

by (2.2), for large n. Letting K' decrease to K gives the stated result. The minimax risk is lower bounded by the maximin risk. In turn, the maximin risk is lower bounded by

replacing w with Jeffreys' prior w^*. So, we have that

$$R_n - \tfrac{1}{2} d \log n \geq R_n^* - \tfrac{1}{2} d \log n$$
$$\geq \int_K w^*(\theta) D(P_\theta \| M_{n,w^*}) \, \mathrm{d}\theta - \tfrac{1}{2} d \log n.$$

By (2.3), the right-hand side has

$$-\tfrac{1}{2} d \log 2\pi e + \log \int_K \sqrt{\det I(\theta)} \, \mathrm{d}\theta + o(1)$$

as an asymptotic lower bound. Since the upper and lower bounds agree (2.4), (2.5) and (2.6) are proved. Jeffreys' prior is asymptotically least favorable, and we have identified the asymptotic minimax value, and a sequence of procedures which achieves it. Consequently, after subtracting $\frac{1}{2} d \log n/(2\pi e)$, the quantities $R_n, R_n^*, R_n(w^*)$, and $R_n(\theta, w^*)$ all have the same limiting value, $\log \int_K \sqrt{I(\theta)} \, \mathrm{d}\theta$, the latter uniformly for θ in K_0.

Finally, we note that no sequence of prior probabilities $\langle W_n \rangle|_{n=1}^{\infty}$ achieves an asymptotically larger value of $R_n(w)$ than Jeffreys' prior. If $\langle W_n \rangle|_{n=1}^{\infty}$ gives a higher value than $R_n(w^*)$ for each fixed n, then we have that

$$R_n(w^*) < \int D(P_\theta^n \| M_{n,W_n}) \, W_n(\mathrm{d}\theta) \leq \sup_W \int D(P_\theta^n \| M_{n,W}) \, W(\mathrm{d}\theta) = R_n^*.$$

However, the difference between the left- and right-hand sides is asymptotically negligible, as we have shown.

4. Proof of Theorem 2

Here we give the proof of Theorem 2 and its corollary. We use θ_0 to denote a fixed value of the parameter.

Proof. We can rewrite the Kullback–Leibler number as

$$E_{\theta_0} \log \frac{p_{\theta_0}(x^n)}{m(x^n)} = \log \frac{1}{w(\theta_0)} - E_{\theta_0} \log \left[1 + \sum_{\theta \neq \theta_0} \frac{w(\theta)}{w(\theta_0)} \frac{p(x^n|\theta)}{p(x^n|\theta_0)} \right]. \tag{4.1}$$

By using the inequality $-\log(1+x) \leq 0$ for x positive we have the bound

$$D(P_{\theta_0}^n \| M_n) \leq \log \frac{1}{w(\theta_0)}$$

which we hope is attained in the limit. To get a lower bound it is enough to upper bound the positive quantity

$$E_{\theta_0} \log \left[1 + \sum_{\theta \neq \theta_0} \frac{w(\theta)}{w(\theta_0)} \frac{p(x^n|\theta)}{p(x^n|\theta_0)} \right]$$

by something which shrinks to zero as n increases.

Consider the set defined by

$$D_n = D_n(\theta_0) = \left\{ x^n : \sum_{\theta \neq \theta_0} \frac{w(\theta)}{w(\theta_0)} \frac{p(x^n|\theta)}{p(x^n|\theta_0)} \leqslant \varepsilon_n \right\}.$$

The set D_n is the discrete analog of the set A_n defined in Section 3. Consequently, by the reasoning in Section 3, $P^n_{\theta_0}(D_n^c) \to 0$ when ε_n is chosen to be of the form e^{-rn} for small enough $r > 0$.

Decomposing the sample space into D_n and D_n^c, the second member of the right-hand side of (4.1) can be written as a sum of two terms. The first is

$$E_{\theta_0} \chi_{D_n} \log \left[1 + \sum_{\theta \neq \theta_0} \frac{w(\theta)}{w(\theta_0)} \frac{p(x^n|\theta)}{p(x^n|\theta_0)} \right] \leqslant E_{\theta_0} \chi_{D_n} \log(1 + e^{-rn})$$

$$= P_{\theta_0}(D_n) \log(1 + e^{-rn}) \leqslant \log(1 + e^{-rn}),$$

which clearly tends to zero as n increases. The other term tends to zero also. It is

$$\begin{aligned} & E_{\theta_0} \chi_{D_n} \log \left[1 + \sum_{\theta \neq \theta_0} \frac{w(\theta)}{w(\theta_0)} \frac{p(x^n|\theta)}{p(x^n|\theta_0)} \right] \\ & \leqslant P_{\theta_0}(D_n^c) \log E_{\theta_0} \frac{\chi_{D_n^c}}{P_{\theta_0}(D_n^c)} \left[1 + \sum_{\theta \neq \theta_0} \frac{w(\theta)}{w(\theta_0)} \frac{p(x^n|\theta)}{p(x^n|\theta_0)} \right] \\ & = -P_{\theta_0}(D_n^c) \log P_{\theta_0}(D_n^c) + P_{\theta_0}(D_n^c) \log E_{\theta_0} \chi_{D_n^c} \left[1 + \sum_{\theta \neq \theta_0} \frac{w(\theta)}{w(\theta_0)} \frac{p(x^n|\theta)}{p(x^n|\theta_0)} \right] \\ & \leqslant -P_{\theta_0}(D_n^c) \log P_{\theta_0}(D_n^c) + P_{\theta_0}(D_n^c) + \log \left[1 + \frac{W(\{\theta | \theta \neq \theta_0\})}{w(\theta_0)} \right]. \end{aligned} \tag{4.2}$$

We see that both terms in (4.2) go to zero since $P_{\theta_0}(D_n^c) \to 0$. Thus, we have that the second term on the right-hand side of (4.1) goes to zero. □

Proof of the corollary. We have that for each θ

$$0 \leqslant D(P_\theta^n \| M_n) \leqslant \log \frac{1}{w(\theta)},$$

and that the quantity in the middle tends pointwise to its upper bound, which is integrable with respect to the prior. The result follows from the dominated convergence theorem. □

5. Interpretations of the results

Here we briefly restate the content of our results in information-theoretic terms and then state a few implications for statistical inference. Bernardo (1979), like Ibragimov and Hasminsky (1973), interpreted the Shannon mutual information as a measure of the information in channel coding and source coding. The channel coding context

permits the interpretation of a reference prior as that source distribution which achieves the maximal rate of data transmission in bits per unit time. The use of any other prior gives a lower rate of transmission. In this sense, Jeffreys' prior serves as a reference. Furthermore, use of the reference prior method suggests that one is implicitly assuming there is some agent which is transmitting the data to the experimenter. When this assumption is valid it provides a physical justification for the prior so that prior selection becomes another aspect of statistical modeling.

5.1. Channel capacity

An information-theoretic channel is a conditional distribution which specifies the probability distribution of the output received given the input sent. The input is an encoded representation of the message. The output has a probabilistic description because it is possible that the transimission was corrupted, by background noise, for instance. We want a high rate of transmission and we want the output received to be decodable so as to give the right message with high probability. The capacity of a channel is the supremal rate of transmission of data across a communication channel. Shannon's channel coding theorem states that for any rate below the capacity there will exist a coding scheme which, over repeated uses of the channel, will achieve that rate with an arbitrarily small probability of decoding error.

Let X be the input to a channel defined by $p(\cdot \mid x)$ and let the output be denoted by Y. We recall that the mutual information between two random variables X and Y is

$$I(X; Y) = \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \, dx \, dy,$$

where $p(x, y)$ is the joint density of (X, Y) and $p(x)$, $p(y)$ denotes the marginal densities for X, Y. The capacity of the channel defined by $p(y|x)$ is

$$C = \sup_{p(x)} I(X; Y).$$

Suppose that we have one broadcaster sending the same encoded message X to each of k receivers $Y_1, \ldots, Y_k$, which are conditionally independent and identically distributed given X. Intuitively, this means that the noise which interferes with the signal received by any one receiver is independent of the noise that interferes with the signal received by any other receiver. Thus, the conditional distribution defining the channel is

$$p(y_1, \ldots, y_k|x) = \prod_{i=1}^{k} p(y_i|x).$$

When a block of coded data $x_1, \ldots, x_n$ is sent, the ith receiver, i between 1 and k, picks up $y_1^i, \ldots, y_n^i$. Suppose the k receivers decode cooperatively, i.e. they pool their data and then estimate the message sent. Then, the capacity of the resulting channel is

$$C_k = \sup_{p(x)} I(X; Y^k).$$

We can relate the present case to the statistical context by letting X correspond to the parameter and Y^k correspond to the random sample. Thus, $p(x)$ takes the role of the density of the prior, and $p(y|x)$ takes the role of the density of the i.i.d. random variables. Denoting the Fisher information for the density by $I(x)$, with $x=(x_1, \ldots, x_d)$ varying over a compact set in $\mathbb{R}^d$, and the entropy of X by $H(X)$, we have that

$$I(X; Y^k)=\frac{d}{2}\log\frac{k}{2\pi e}+H(X)+\int p(x)\log\det I(x)\,dx+o(1),$$

under the same assumptions as in Theorem 1 except for a change in notation. As a result, we have that the capacity of k receivers, C_k, is

$$C_k=\frac{d}{2}\log\frac{k}{2\pi e}+\log c,$$

where $c=\int\sqrt{\det I(x)}\,dx$. Observe that this formula is asymptotic in k, the number of receivers, and not in the length of the data stream.

When the input arises from a continuous distribution, it is seen that Jeffreys' prior is the source distribution which achieves the capacity. Using a different source distribution would give the rate of transmission as the corresponding mutual information and be strictly less than the capacity. Since the capacity increases as the logarithm of the number of receivers, there are coding schemes which achieve rates of transmission arbitrarily close to $(d/2)\log(k/2\pi e)+c$ when k is large. A similar interpretation can be given in the discrete case provided the entropy can be maximized.

5.2. *Universal noiseless source coding*

An alternative interpretation of $R_n(\theta, w)$, $R_n(w)$, and R_n can be stated in terms of the redundancy of universal noiseless source coding. For ease of exposition we consider the case where the X_i's are discrete. Suppose we want a variable length binary code so as to encode a block of data X^n for transmission, but that the underlying density governing X^n is only known to be a member of the smooth parametric family $\{p_\theta\}$. We seek a code which is universal in the sense that it will perform well no matter which element of the parametric family is true.

In this context, a code's performance is assessed by its expected codelength, which we want to be small. It is well known that the lower bound on the expected codeword length is the entropy of the distribution. Cosequently, we minimize the redundancy which is the difference between the expected codelength of the code we use and the entropy bound which can be achieved to within one bit when the true distribution is known. Indeed, if θ were the true value of the parameter, and were known, we would use the Shannon code which has codelengths given by $\log 1/p_\theta(X)$: a code with these lengths is guaranteed to exist by the Kraft–McMillan theorem (see Blahut (1987, p. 50)) and achieves the entropy bound to within one-bit roundoff error. If θ is not known then a Bayesian would use a code with lengths $\log 1/m(X^n)$.

Indeed, typically θ is not known. To verify the Bayesian's intuition, consider a code $\{\phi\}$ with codelengths given by $l(\phi(X^n))$. Then $Q_n(X^n)=2^{-l(\phi(X^n))}$ is a subprobability mass function and, for a fixed θ, the redundancy of $\{\phi\}$ is $E_\theta\, l(\phi(X^n))-H(P_\theta^n)=D(P_\theta^n \| Q_n)$. If we integrate this redundancy with respect to a prior density $w(\theta)$ then we obtain the Bayes redundancy of the code $\{\phi\}$. By definition, the Bayes code is the code which achieves the minimal Bayes redundancy. Performing this minimization it is seen that the Bayes code has codelengths $\log 1/m(X^n)$. Consequently, $R_n(\theta, w)= D(P_\theta^n \| M_n)$ is the pointwise redundancy of the Bayes code, and $R_n(w)= \int w(\theta)R(\theta, w)\,\mathrm{d}\theta$ is the Bayes redundancy of the Bayes code. Theorem 1 gives an asymptotic expression for the pointwise redundancy $R_n(\theta, w)$ which is uniformly good on compact sets. Also, Theorem 1 gives an asymptotic expression for the Bayes redundancy $R_n(w)$.

Analogously, the minimax code achieves the minimax redundancy $\min_{Q_n}\max_\theta D(P_\theta^n \| Q_n)$ and the maximin code achieves the maximin redundancy $\max_w \min_{Q_n} \int w(\theta)D(P_\theta^n \| M_n)\,\mathrm{d}\theta$. Theorem 1 shows that the minimax code and the maximin code are asymptotically the same, and gives an asymptotic expression for the common redundancy. Also, it is seen that the asymptotically minimax (or maximin) code has codelengths specified by $\log 1/m_{n,w^*}(X^n)$ to within one-bit roundoff error, where w^* is Jeffreys' prior.

If the X_i's are continuous and can be quantized arbitrarily finely, this interpretation holds in a limiting sense. Also, if the parameter is discrete, Theorem 2 says that the number of extra bits required by coding based on m_n rather than p_θ is the Shannon codelength of the true parameter value under the prior.

5.3. *Bounds on the cumulative risk*

Suppose that we have a parametric family indexed by θ and that we want to identify the true density p_{θ_0}, but that it is not the true value of the parameter that interests us. One natural estimator of $p(x \mid \theta_0)$ at any given x is the predictive density $\hat{p}_n(\cdot)$, the posterior mean of $p(x \mid \Theta)$. If the relative entropy is used as the loss function for parametric density estimation, one can examine the behavior of the cumulative risk.

Let δ_k for $k=0, \ldots, n-1$ be a sequence of density estimators. Each δ_k estimates the density of X_{k+1}, given the data X^k. When θ_0 is true, the risk associated with $\delta_k=\delta_k(X^k)$ is $E_{\theta_0} D(P_{\theta_0} \| \delta_k)$, and the cumulative risk over the first n uses of the sequence of estimators is the sum of the individual risks $C(n, \theta_0, \delta)= \sum_{k=0}^{n-1} E_{\theta_0} D(P_{\theta_0} \| \delta_k)$. The sum of the risks plays an important role in universal coding theory, sequential estimation, hypothesis testing and portfolio selection theory; see Clarke and Barron (1990).

Proposition 1. *The cumulative risk of the sequence of Bayes $\hat{p}_n$ is*

$$C(n, \theta, \hat{p}_n)= \sum_{k=0}^{n-1} E_{\theta_0} D(P_{\theta_0} \| \hat{p}_k)=D(P_{\theta_0}^n \| M_n),$$

under the convention that $\hat{p}_0(x)=m_1(x_1)$. *Its cumulative Bayes risk is*

$$\int w(\theta)\, D(P_{\hat{\theta}}^n \| M_n)\, \mathrm{d}\theta.$$

Under the assumptions of Theorem 1, *the minimax risk is asymptotically realized by choosing* w *to be the Jeffreys' prior which is asymptotically least favorable. Consequently, the cumulative risk, Bayes risk and minimax risk are asymptotically approximated by expressions of the form* $\frac{1}{2}d\log n+c$.

Proof. This is a restatement of Theorems 0 and 1. □

An analogous result can be stated for the risk and Bayes risk in the discrete case using Theorem 2. A general result for the minimax risk can only be stated when the entropy can be maximized.

Alternatively, if the parameter value is of interest, then estimating it can be regarded as a special case of density estimation where we restrict the estimator of the density to be of the form $p(x \mid \theta(X^n))$. Enlarging the class of estimators we see that the Bayes risk in parametric density estimation lower-bounds the Bayes risk in parametric estimation:

$$\inf_{\delta} E_w E_\theta D(\theta \| \delta) \geq \inf_{Q} E_w E_\theta D(P_\theta \| Q).$$

Similarly, for the minimax risk we have

$$\inf_{\delta} \sup_{\theta} E_\theta D(\theta \| \delta) \geq \inf_{Q} \sup_{\theta} E_\theta D(P_\theta \| Q),$$

where δ is an estimator of the parameter, Q is an estimator of the density and $D(\theta \| \delta)=D(P_\theta \| P_\delta)$ is the relative entropy loss for parameter estimation. (A similar statement holds for the maximin risk.) Thus, Theorems 1 and 2 are seen to give asymptotic lower bounds on the minimax (and maximin) cumulative risk of parameter estimation.

References

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62**, 547–554.

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

Berger, J.O. and J.M. Bernardo (1991). On the development of the reference prior method. In: *Bayesian Statistics 4: Proc. 4th Valencia Internat. Meeting on Bayesian Statistics.* Clarendon Press, Oxford.

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

Berger, J.O. and J.M. Bernardo (1992b). Reference priors in a variance components problem. In: P.K. Goel and N.S. Iyengar, Eds., *Bayesian Analysis in Statistics and Econometrics.* Springer, New York, 177–194.

Berger, J.O., J.M. Bernardo and M. Mendoza (1989). On priors that maximize expected information. In: J.P. Klein and J.C. Lee, Eds., *Recent Developments in Statistics and Their Applications.* Freedom Academy Publishing, Seoul.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B* **41**, 113–147.

Bickel, P.J. and J.K. Ghosh (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction — a Bayesian argument. *Ann. Statist.* **18**, 1070–1090.

Blahut, R.E. (1987). *Principles and Practice of Information Theory.* Addison-Wesley, Reading, MA.

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

Clarke, B. (1989). Asymptotic cumulative risk and Bayes risk under entropy loss with applications. Ph.D. thesis, Department of Statistics, University of Illinois at Urbana-Champaign.

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

Clarke, B. and A.R. Barron (1991). Entropy risk and the Bayesian central limit theorem. *Technical Report* 91-5, Department of Statistics, Purdue University.

Csiszar, I. (1993) Stanford course notes, private communication.

Davisson, L. (1973). Universal noiseless coding. *IEEE Trans. Inform. Theory* **19**, 783–795.

Davisson, L. and A. Leon-Garcia (1980). A source matching approach to finding minimax codes. *IEEE Trans. Inform. Theory* **26**, 166–174.

Efroimovich (1980). Information contained in a sequence of observations. *Problems Inform. Transmission* **15**, 24–39.

George, E. and R. McCulloch (1989). On obtaining invariant prior distributions. *Technical Report*, Graduate School of Business, University of Chicago.

Ghosh, J.K. and R. Mukerjee (1992). Noninformative priors. In: *Bayesian Statistics 4: Proc. 4th Valencia Internat. Meeting on Bayesian Statics.* Clarendon Press, Oxford.

Gray, R.M. (1988). *Probability, Random Processes, and Ergodic Properties.* Springer, New York.

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

Hartigan, J.A. (1983). *Bayes Theory.* Springer, New York.

Ibragimov, I.A. and R.Z. Hasminsky (1973). On the information in a sample about a parameter. In: *Proc. 2nd Internat. Symp. on Information Theory*, Akademiai, Kiado, Budapest, 295–309.

Ibragimove, I.A. and R.Z. Hasminskii (1981). *Statistical Estimation: Asymptotic Theory.* Springer, New York.

Jeffreys, H. (1961). *Theory of Probability.* Oxford Univ. Press, New York.

Krichevsky, R.E. and V.K. Trofimov (1981). The performance of universal encoding. *IEEE Trans. Inform. Theory* **27**, 199–207.

Lehmann, E.L. (1983). *Theory of Point Estimation.* Wiley, New York.

Polson, N. and L. Wasserman (1989). Prior distributions for the bivariate binomial. *Technical Report 460*, Department of Statistics, Carnegie-Mellon University.

Rissanen, J. (1986). Stochastic complexity and modeling. *Ann. Statist.* **14**, 1080–1110.

Rissanen, J. (1987). Stochastic complexity. *J. Roy. Statist. Soc. Ser. B* **49**, 223–239; *IEEE Trans. Inform. Theory* **30**, 629–636.

Royden, H.L. (1968). *Real Analysis.* MacMillan Publishing, New York.

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

Zhang, Z. (1994). *Discrete Noninformative Priors.* Ph.D. Dissertation, Department of Statistics, Yale University.