The Annals of Statistics
1998, Vol. 26, No. 6, 2083–2103

THE MAXIMUM LIKELIHOOD PRIOR[1]

BY J. A. HARTIGAN

Yale University

Consider an estimate θ^* of a parameter θ based on repeated observations from a family of densities f_θ evaluated by the Kullback–Leibler loss function $K(\theta, \theta^*) = \int \log(f_\theta/f_{\theta^*})f_\theta$. The maximum likelihood prior density, if it exists, is the density for which the corresponding Bayes estimate is asymptotically negligibly different from the maximum likelihood estimate. The Bayes estimate corresponding to the maximum likelihood prior is identical to maximum likelihood for exponential families of densities. In predicting the next observation, the maximum likelihood prior produces a predictive distribution that is asymptotically at least as close, in expected truncated Kullback–Leibler distance, to the true density as the density indexed by the maximum likelihood estimate. It frequently happens in more than one dimension that maximum likelihood corresponds to no prior density, and in that case the maximum likelihood estimate is asymptotically inadmissible and may be improved upon by using the estimate corresponding to a least favorable prior. As in Brown, the asymptotic risk for an arbitrary estimate "near" maximum likelihood is given by an expression involving derivatives of the estimator and of the information matrix. Admissibility questions for these "near ML" estimates are determined by the existence of solutions to certain differential equations.

1. Introduction. A common assumption, in both frequentist and Bayesian approaches to statistical inference, is that observations are sampled according to a family $\{f_\theta\}$ of probability densities indexed by a parameter θ lying in some subset Θ of R^p. In the Bayesian approach, the model is supplemented by a prior distribution for the parameter θ, and then inferences are much simplified by the presence of a full probabilistic model for the joint distribution of observations and parameter. The selection and justification of these priors is an important part of the Bayesian approach. But even in a classical decision theoretic view, the admissibility of any statistical technique is usually explored by examining various types of Bayesian approximations to the technique. Thus choosing to use a particular technique may be cast as choosing to use a particular prior distribution or a particular class of prior distributions for which that technique is optimal. These "technical" priors offer a useful bridge between frequentist and Bayesian approaches. In this paper, we describe a prior distribution corresponding asymptotically to maximum likelihood, as the number of observations n from f_θ approaches ∞

Received September 1996; revised July 1998.
[1]Supported in part by NSF Grant DMS-86-17919.
AMS 1991 *subject classifications*. Primary 62F15; secondary 62C15.
Key words and phrases. Uninformative priors, maximum likelihood, Kullback–Leibler distance, the Jeffreys prior, asymptotic admissibility.

and show that maximum likelihood is asymptotically inadmissible when there is no such prior.

The most widely used prior distribution of this "technical" type was proposed for estimation problems by Jeffreys (1946). The Jeffreys density, with respect to p-dimensional Lebesgue measure, is

$$J(\theta) = \det^{1/2}(L(\theta)),$$

where $L(\theta)$ is the information matrix $\mathbb{P}_\theta[-\partial^2/\partial\theta\,\partial\theta' \log f_\theta]$ and $\mathbb{P}_\theta$ denotes expectation, given θ.

There are several justifications for the Jeffreys density. Jeffreys (1946) uses the Kullback–Leibler distance

$$K(f_\theta, f_{\theta'}) = \mathbb{P}_\theta \log(f_\theta/f_{\theta'});$$

balls of K-radius r are given equal probability, as $r \to 0$, by the Jeffreys density. Perks (1947) notes that a maximum likelihood-based confidence region of given confidence size has volume asymptotically inversely proportional to $J(\theta)$ as θ varies. Thus the Jeffreys prior states that the parameter value will lie in any one of these confidence regions with asymptotically equal probabilities. Welch and Peers (1963) show that the one-sided posterior intervals of posterior probability α are confidence intervals of size $\alpha + O(n^{-1/2})$ for any smooth prior, but are of size $\alpha + O(n^{-1})$ for the Jeffreys prior. This justification does not extend to two-sided intervals or to many dimensions. The Jeffreys density has wide acceptance in one dimension, but is not used so much in many dimensions; Jeffreys (1961) recommends against its use in regression problems where the Jeffreys density produces degrees of freedom for the error sum of squares not in accord with classical calculations.

Bernardo (1979) offers a justification again based on the Kullback–Leibler distance. Determine a prior distribution H to maximize, as the amount of data x increases, the expected K–L distance between the posterior distribution H_x and the prior distribution H:

$$I(H) = \int \log(dH_x/dH)\, f_\theta\, dx\, dH(\theta).$$

Bernardo argues that this prior distribution is least informative, because the data is allowed the greatest weight when the largest possible difference between prior and posterior is expected. Under certain conditions, for example Clarke and Barron (1994), the asymptotically optimal prior is the Jeffreys prior. The Bernardo prior is the least favorable prior in the no-data decision problem in which the object is to select a single density f, with loss function $K(f_\theta, f)$. The least favorable prior maximizing $\inf_f \int K(f_\theta, f)\, dH(\theta)$, the Bernardo prior, generates the Bayes density $f_H = \int f_\theta\, dH$ which is also minimax, minimizing $\max_\theta K(f_\theta, f_H)$. See also Berger and Bernardo (1992) for some modifications of the method when nuisance parameters are present.

None of these justifications directly support Jeffreys' original program for using the Jeffreys density in parameter estimation problems. In Hartigan

(1964, 1965), I proposed the "asymptotically locally invariant" prior M solving

$$\frac{\partial \log M}{\partial \theta_i} = q_i = \sum_{jk} L^{-1}_{j,k} \mathbb{P}_\theta \left[\frac{\partial^2 \log f_\theta}{\partial \theta_i \partial \theta_j} \frac{\partial \log f_\theta}{\partial \theta_k} \right]$$

and a one-parameter family of invariant priors, including Jeffreys at $\alpha = \frac{1}{2}$,

$$\frac{\partial \log M_\alpha}{\partial \theta_i} = \sum_{jk} L^{-1}_{j,k} \mathbb{P}_\theta \left[\alpha \frac{\partial \log f_\theta}{\partial \theta_i} \frac{\partial \log f_\theta}{\partial \theta_j} \frac{\partial \log f_\theta}{\partial \theta_k} + \frac{\partial^2 \log f_\theta}{\partial \theta_i \partial \theta_j} \frac{\partial \log f_\theta}{\partial \theta_k} \right].$$

It is not necessarily true that any M exists satisfying the first differential equation in more than one dimension. The necessary and sufficient condition for a solution is that $dq_i/\partial\theta_j = dq_j/\partial\theta_i$ for each i, j.

For the exponential family in the canonical form

$$f(\mathbf{x} \mid \boldsymbol{\theta}) = \exp\Big(\sum_i \theta_i Y_i(\mathbf{x}) + c(\boldsymbol{\theta}) \Big),$$

the prior M sets $\boldsymbol{\theta}$ uniform. This prior was proposed for exponential families by Hurzubazar [Jeffreys (1961)].

In this paper, I will argue that choosing maximum likelihood estimation when using truncated Kullback–Leibler loss is asymptotically equivalent to choosing the prior M. We use asymptotic expansions of risk functions in which the first term $p/2n$ is the same for maximum likelihood and for all Bayes estimates with sufficiently smooth prior densities. Comparisons between the different methods are based on examination of the $1/n^2$ terms.

The study of "third order" asymptotic behavior of the maximum likelihood estimator $\hat{\theta}$ was begun by Rao (1961); see Ghosh and Subramanyan (1974), Pfanzagl and Wefelmeyer (1978), Ghosh (1994). A typical result is that an efficient estimator T_n with bias $b(\theta)/n$ has asymptotic risk, up to terms in order $o(n^{-2})$, no less than a "bias-adjusted" likelihood estimator of $\theta^* = \hat{\theta} + c(\hat{\theta})/n$ where $c(\hat{\theta})$ is chosen to match the bias in T_n.

Brown (1979) suggests a general program for evaluating the admissibility of an estimator δ, for a general class of estimators and loss functions, by looking at the behavior of risk differences between the estimator δ and a family of competing estimators δ' near δ. In the maximum likelihood case, the competing estimators are the "bias-adjusted" maximum likelihood estimators studied by Ghosh (1994). Admissibility questions are now expressed in terms of certain differential operators that first appeared in Stein (1956) for the multivariate normal location problem.

The asymptotic behavior of Bayes estimators near maximum likelihood estimators have been studied for loss functions of form $L(d, \theta) = w(d - \theta)$ by Levit (1982, 1983, 1985); in particular, he shows that the Bayes estimators form a complete class within the class of estimators near maximum likelihood under certain regularity conditions.

Eguchi and Yanagimoto (private communication) study bias-adjusted maximum likelihood estimators under reverse Kullback–Leibler loss $K(\theta^*, \theta) =$

$\int \log(f_{\theta^*}/f_\theta) f_{\theta^*}$; the risk difference between a Bayes estimator and maximum likelihood is expressed in terms of the Laplace–Beltrami operator $\Delta u = \sum_{i,j} J^{-1}(\partial/\partial\theta_j)(JL_{i,j}^{-1}(\partial/\partial\theta_i)u)$, and so maximum likelihood is asymptotically inadmissible when there exist *superharmonic* positive nonconstant u satisfying $\Delta u \leq 0$, generalizing Brown's (1971) result for normal location families. The Bayes estimators considered by Eguchi and Yanagimoto are chosen to maximize the posterior density rather than to minimize posterior expected loss; under the usual Bayes rule, it turns out that the maximum likelihood prior with this loss function is Jeffreys density.

The present paper follows Brown and Levit for a particular estimator (maximum likelihood) and a particular loss function (Kullback–Leibler). The Bayes estimate θ_h corresponding to h minimizes the truncated posterior expectation given n observations of $K(f_\theta, f_{\theta_h})$. The maximum likelihood estimate $\hat{\theta}$ based on n observations, maximizes $\Pi f_\theta(x_i)$. We will show that θ_h differs from $\hat{\theta}$ by order n^{-1} for every smooth prior density h, but that θ_M differs from $\hat{\theta}$ by order $n^{-3/2}$. The truncated risk difference $\mathbb{P}_\theta^\varepsilon[K(f_\theta, f_{\theta_h}) - K(f_\theta, f_{\hat{\theta}})]$ is order n^{-2} for every smooth density, but of order n^{-3} for M. We identify, via Levit's complete class theorem modified for Kullback–Leibler loss, maximum likelihood estimators as inadmissible when they do not satisfy the simple differential equation necessary for the maximum likelihood prior to exist. It remains true, as in Stein's problem, that Bayes estimators may also be inadmissible, and this question is settled by the existence of solutions of certain elliptical differential equations, depending on the prior, the information matrix and the parameter space.

One application of the maximum likelihood prior is to density estimation. The maximum likelihood *estimative* density for the next observation is $f_{\hat{\theta}}$. We compare this estimate to the *predictive* h-estimate

$$f_h(x \mid \mathbf{x}) = \int f_\theta(x) h(\theta \mid \mathbf{x})\, d\theta.$$

The K–L distance $K(f_{\hat{\theta}}, f_h)$ is asymptotically $O(n^{-2})$ for any smooth prior h. The limiting distance, for each true value θ, is minimized by the choice M, when it is possible to find a prior of form M. Measuring the loss of a density estimate $\hat{f}$ by $K(f_\theta, \hat{f})$, the asymptotic risk for $\hat{f} = f_{\hat{\theta}}$ is no less than the asymptotic risk of $\hat{f} = f_M$ for every θ. [See Aitchison (1975), Komaki (1996) for similar comparisons for exponential families.] It turns out that the ratio of predictive to estimative density is asymptotically the same for all priors, and this ratio may be used to improve any estimative density based on an efficient estimator, with the same improvement in the risk.

A referee suggests considering invariant loss functions other than Kullback–Leibler, for example, the Amari (1982) divergences $L(\theta^*, \theta) = (\beta - \beta^2)^{-1}[1 - \int f_{\theta^*}^\beta f_\theta^{1-\beta}\, dx]$, which includes Kullback–Leibler ($\beta = 0$), Hellinger ($\beta = \frac{1}{2}$), reverse Kullback–Leibler ($\beta = 1$), chi square ($\beta = 2$). For each choice of β, the maximum likelihood prior is the member of the invariant family of priors proposed in Hartigan (1964) with $\alpha = \frac{1}{2}\beta$. The choice $\beta = 1$ corresponding to reverse Kullback–Leibler loss (studied by Eguchi and

Yanagimoto) is the Jeffreys density; this is the one member of the family that always satisfies the condition for existence of the prior. There are some reasons for preferring Kullback–Leibler loss to the other invariant measures. One is that Kullback–Leibler loss has historically been the principal device for developing noninformative priors. The corresponding predictive Bayes density estimate for the next observation is just the posterior density of the next observation. Also, maximum likelihood is exactly the Bayes estimate for the maximum likelihood prior under Kullback–Leibler loss in exponential families.

Following Brown (1979), the same kinds of asymptotic computations are possible for other than invariant loss functions and other than maximum likelihood estimator sequences, even in nonparametric settings. We can ask whether or not an estimator is asymptotically Bayes in the family of bias-adjusted estimators and show that it can be beaten asymptotically by another bias-adjusted estimator if it is not.

2. Truncated expectations and regularity conditions A. We need expectations to evaluate risks to terms in $O(n^{-2})$, but we must work with Taylor series for the loss of form $A(\hat{\theta} - \theta)^2 + B(\hat{\theta} - \theta)^3 + \cdots$ where we do not know the detailed tail behavior of $\hat{\theta} - \theta$ and do not know the detailed behavior of the remainder terms in the Taylor series expansion.

Say a sequence of random variables Y_n is *truncatable* (denoted $Y_n = t_n$) if for each $\varepsilon > 0$, $\mathbb{P}\{|Y_n| > \varepsilon\} = o(n^{-2})$; *tail-bounded* (denoted $Y_n = T_n$) if for some A, $\mathbb{P}\{|Y_n| > A\} = o(n^{-2})$. The two orders t_n, T_n are adaptations of the usual stochastic orders o_p, O_p, with $o(n^{-2})$ rates necessary for asymptotic risk approximation. For any random variable Y, define its *truncated* expectation

$$\mathbb{P}^{\varepsilon}Y = \mathbb{P}\big[-\varepsilon\{Y < -\varepsilon\} + Y\{|Y| \le \varepsilon\} + \varepsilon\{Y > \varepsilon\}\big].$$

It may be verified that truncated expectations of truncatable sequences of random variables satisfy the usual rules of finitely additive expectation up to $o(n^{-2})$, namely, that Y_n, Z_n truncatable implies $|Y_n|$, $aY_n + bZ_n$ truncatable, and

$$\mathbb{P}^{\varepsilon}Y_n = Y_n + o(n^{-2}) \quad \text{for } Y_n \text{ constant},$$

$$\mathbb{P}^{\varepsilon}|Y_n| \ge 0,$$

$$\mathbb{P}^{\varepsilon}(aY_n + bZ_n) = a\mathbb{P}^{\varepsilon}Y_n + b\mathbb{P}^{\varepsilon}Z_n + o(n^{-2}).$$

Observations $x_1, \ldots, x_n$ are independent and identically distributed from a density f_θ, $\theta \in \Theta$, a subset of R^p. We consider asymptotic behavior at the true value θ_0, an interior point of Θ. For sequences of indices $i_1, i_2, \ldots, i_r$, and for subsets of indices $S_1, S_2, \ldots, S_t$, define

$$l_{i_1 i_2 \cdots i_r} = \sum_{j=1}^{n} \frac{\partial}{\partial\theta_{i_1}} \frac{\partial}{\partial\theta_{i_2}} \cdots \frac{\partial}{\partial\theta_{i_r}} \log f_\theta(x_j) \quad \text{at } \theta = \theta_0,$$

$$L_{S_1, S_2, \ldots, S_t} = \mathbb{P}_{\theta_0}\big[l_{S_1} l_{S_2} \cdots l_{S_t}\big] \quad \text{for } n = 1.$$

The i, j element of the inverse of the matrix $L_{i,j}$ will be denoted $L_{i,j}^{-1}$.

The maximum likelihood estimate $\hat{\theta}$ is a value of θ maximizing the log likelihood $l(\theta) = \log \Pi f_\theta(x_j)$ over Θ. Likelihood derivatives and expectations of their products evaluated at $\hat{\theta}$ will be denoted $\hat{l}_{i_1 i_2 \cdots i_r}$, $\hat{L}_{S_1, \ldots, S_t}$. Prior densities on Θ with respect to Lebesgue measure will be denoted by h and $h_i = (\partial/\partial\theta_i)\log h$, $\hat{h}_i = h_i$ evaluated at $\hat{\theta}$. The maximum likelihood prior density will be denoted by M, and M_i, $\hat{M}_i$ are defined as for h. The posterior density is $h(\theta \mid \mathbf{x}) = f(\mathbf{x} \mid \theta)h(\theta)k(\mathbf{x})$ where $k(\mathbf{x})$ is chosen so that the posterior density integrates to 1. The prior h may be improper, but the posterior will be assumed to be proper except for a set of data values $\mathbf{x}$ which will be assumed to have asymptotically zero probability given θ_0. Posterior truncated moments will be written $\mathbb{P}^\varepsilon[\theta \mid \mathbf{x}]$, $\mathbb{P}^\varepsilon[\theta^2 \mid \mathbf{x}]$ and so forth.

ASSUMPTIONS A.

A1. All derivatives of $\log f_\theta(x)$ up to order 5 exist for each observation x and for θ in a neighborhood of θ_0.
A2. All moments exist for the first four derivatives and for the maximum squared fifth derivatives in a neighborhood of θ_0. The moments are differentiable in a neighborhood of θ_0.
A3. The information matrix $L_{i,j}$ is positive definite.
A4. $\mathbb{P}_{\theta_0}[((\partial/\partial\theta_{i_1})(\partial/\partial\theta_{i_2})\cdots(\partial/\partial\theta_{i_r})f)/f]_{\theta=\theta_0} = 0$, $1 \le r \le 4$.
A5. The prior density h is positive and has two derivatives in a neighborhood of θ_0.
A6. $|\hat{\theta} - \theta_0| = t_n$.
A7. Posterior tail probabilities are negligible: $\int_{|\theta-\theta_0|>\varepsilon} \sup_x f_\theta(x)h(\theta \mid \mathbf{x})\, d\theta = t_n/n^2$.
A8. θ_0 is K-identifiable; for each $\varepsilon > 0$, there exists $\delta > 0$ such that $K(\theta, \theta_0) < \delta$ implies $|\theta - \theta_0| < \varepsilon$.

These are versions of the usual Wald regularity conditions that ensure that maximum likelihood behaves well asymptotically. They are stronger than usual here because we need to specify the risk behavior neglecting only terms in $o(n^{-2})$. The assumption A6 specifies that $\hat{\theta} - \theta_0$ be truncatable; it will have to be checked by considering global behavior of the likelihood. Likewise, the assumption A7 ensures that we only need parameter values near θ_0 in evaluating posterior truncated moments and in evaluating the posterior density of a new observation; if $f_\theta(x)$ is bounded, it states that the tail probability for a given $\mathbf{x}$ is $o(n^{-2})$, excluding a set of $\mathbf{x}$ values which has probability $o(n^{-2})$ given θ_0. Also the assumption A8 is necessary to ensure that the Bayes estimate will be close to θ_0. All the other conditions depend only on the behavior of the likelihood in the neighborhood of θ_0. The condition A4 is the generalization of the information equality $L_{1,2} + L_{12} = 0$ to three and four derivatives; for example, the third order equation is $L_{1,2,3} + L_{12,3} + L_{13,2} + L_{23,1} + L_{123} = 0$, which is seen to be constructed, as are the other equations, by summing expectations of products of derivatives of log likelihood over all the partitions of the set 1, 2, 3. In the calculations we will adopt

the tensor analysis convention of implicitly summing expressions over the range of an index that appears twice.

3. Asymptotic behavior of maximum likelihood and Bayes procedures. The asymptotic risk of the maximum likelihood estimator

THEOREM 1. *Under assumptions* A, *when* θ_0 *is true*,

$$\mathbb{P}^{\varepsilon}_{\theta_0} K\big(\theta_0, \hat{\theta}\big) = \frac{p}{2n} + \frac{1}{n^2}\left\{ -\frac{1}{2}p + L^{-1}_{i,j}L^{-1}_{k,l}\left[L_{i,k,jl} + \frac{3}{2}L_{ik,jl} + \frac{3}{2}L_{i,jkl} + \frac{3}{8}L_{ijkl}\right] - L^{-1}_{i,j}L^{-1}_{k,l}L^{-1}_{m,n}\left[2L_{ikm}L_{n,jl} + \frac{5}{12}L_{ikm}L_{jln} + \frac{3}{8}L_{ikl}L_{jmn} + \frac{3}{2}L_{ikl}L_{m,nj} + \frac{3}{2}L_{i,jk}L_{m,nl} + \frac{3}{2}L_{m,ik}L_{j,nl}\right]\right\} + o(n^{-2}).$$

Our various Bayes estimates will have asymptotic risks differing only in the $O(n^{-2})$ term.

We will develop an asymptotic formula for the posterior density when the true parameter value is θ_0. Such formulas have been given previously, under similar assumptions, for example Strassen (1977); we need to make sure that we can compute the Bayes estimate sufficiently accurately to evaluate risks to terms in $O(n^{-2})$. It will turn out to be sufficient to know the truncated posterior moments to terms in $O(n^{-1})$.

THEOREM 2. *Let* $\delta = \theta = \hat{\theta}$. *Under assumptions* A, *for* $|\hat{\theta} - \theta_0| < \varepsilon$, $|\theta - \hat{\theta}| < \varepsilon$,

$$h(\theta \mid \mathbf{x}) = \det{}^{1/2}\big(-\hat{l}_{ij}/2\pi\big)\exp\Big[\tfrac{1}{2}\delta_i\delta_j\hat{l}_{ij} + \big(1 + n|\delta|^2\big)^2 R_n(\theta)/n\Big] \times\Big[1 + \delta_i\hat{h}_i + \tfrac{1}{6}\delta_i\delta_j\delta_k\hat{l}_{ijk}\Big],$$

where $\sup_{|\theta - \hat{\theta}| < \varepsilon} |R_n(\theta)| = T_n$.

The truncated posterior moments of $\theta \mid \mathbf{x}$ *satisfy, for* $|\hat{\theta} - \theta_0| < \varepsilon$,

$$\mathbb{P}^{\varepsilon}\big[\,\delta_i \mid \mathbf{x}\big] = -\hat{l}^{-1}_{ij}\hat{h}_j + \tfrac{1}{2}\hat{l}^{-1}_{ij}\hat{l}^{-1}_{kl}\hat{l}_{jkl} + n^{-2}T_n,$$

$$\mathbb{P}^{\varepsilon}\big[\,\delta_i\delta_j \mid \mathbf{x}\big] = -\hat{l}^{-1}_{ij} + n^{-2}T_n.$$

All higher order moments of δ *are* $n^{-2}T_n$.

The Bayes estimate θ^h minimizes the truncated expected posterior loss $\mathbb{P}^{\varepsilon}[K(\theta, \theta^h) \mid \mathbf{x}]$. Let $q_i = L^{-1}_{j,k}L_{j,ik}$. The maximum likelihood prior density M satisfies $M_i = q_i$. Such a twice differentiable prior exists for all $\theta \in \Theta$ if and only if for each i, j, $\partial q_i/\partial\theta_j = \partial q_j/\partial\theta_i$. When M exists, it produces a Bayes estimate negligibly different from maximum likelihood.

THEOREM 3. *Under assumptions* A, *when* θ_0 *is true*, *with* h, q *evaluated at* θ_0, *then* $n(\theta_j^h - \hat{\theta}_j) \to L_{i,j}^{-1}(h_i - q_i)$ *in probability*. *Thus*, $\theta^h - \hat{\theta} = o_p(n^{-1})$ *precisely when the prior is* M.

We compare the asymptotic risk of Bayes estimators with the asymptotic risk of the maximum likelihood estimator specified in Theorem 1.

THEOREM 4. *Under Assumptions* A, *when* θ_0 *is true*,

$$\mathbb{P}_{\theta_0}^{\varepsilon}\left[K(\theta_0, \theta^h) - K(\theta_0, \hat{\theta})\right] = n^{-2}\left\{\frac{\partial}{\partial\theta_i}\left(L_{i,j}^{-1}\left[h_j - q_j\right]\right) + \frac{1}{2}L_{i,j}^{-1}(h_i h_j - q_i q_j)\right\} + o(n^{-2}).$$

More generally, *bias-adjusted* estimators θ^* of form

$$\theta_i^* = \hat{\theta}_i + \hat{L}_{i,j}^{-1}(\hat{q}_j^* - \hat{q}_j)/n,$$

where the q_j^* are differentiable functions on Θ, have *asymptotic risk* $A(\theta_0, \theta^*)$:

$$\lim_{n\to\infty} n^2 \mathbb{P}_{\theta_0}^{\varepsilon}\left[K(\theta_0, \theta^*) - \frac{p}{2n}\right] + c(\theta_0) = \frac{\partial}{\partial\theta_i}\left(L_{i,j}^{-1}q_j^*\right) + \frac{1}{2}L_{i,j}^{-1}q_i^* q_j^*.$$

The derivative term may be thought of as the change in variance term, the square term as the change in bias term. The constant $c(\theta_0)$ is irrelevant in comparing the different estimators. We can now develop a decision theory for bias-adjusted estimators based on their asymptotic risks.

Define the *Bayes asymptotic risk* for θ^* by

$$B(\theta^*) = \int \lim_{n\to\infty} n^2\left[\mathbb{P}_{\theta_0}^{\varepsilon} K(\theta_0, \theta^*) - \frac{p}{2n}\right]h(\theta_0)\, d\theta_0.$$

The Bayes risk is not necessarily the same as the limit of the finite sample Bayes risks, since contributions from near the boundary of the parameter space may dominate that limit; in particular Bayes estimates may differ from maximum likelihood estimates by $O_p(n^{-1/2})$ rather than $O_p(n^{-1})$ near the boundary.

THEOREM 5. *If* Θ *is compact with an open interior*, *assumptions* A *hold on an open set that includes* Θ, *and the prior density* h *vanishes on the boundary of* Θ,

$$B(\theta^*) - B(\theta^h) = \tfrac{1}{2}\int L_{i,j}^{-1}(h_i - q_i^*)(h_j - q_j^*)h\, d\theta.$$

Thus the bias-adjusted estimate θ^* of minimum Bayes asymptotic risk is just the Bayes estimate θ^h when h vanishes on the boundary of Θ. When the boundary condition is not met, θ^h may be shown to be the limit of Bayes estimates θ^{h^*} for priors h^* vanishing on the boundary.

4. Exponential families. Suppose the density f is of canonical exponential form

$$\log f_\theta(x) = \sum_{i=1}^{p} \theta_i Y_i(x) - a(\theta)$$

with respect to some x-measure ν. Let a_i and a_{ij} denote first and second derivatives of a. The natural parameter set Θ is the convex subset of R^p for which $\int \exp(\sum_{i=1}^{p} \theta_i Y_i(x))\, d\nu$ is finite. Another useful parametrization is in terms of the *expectation* parameters $\phi_i = a_i$.

Assume that Θ has a nonnull interior. The maximum likelihood prior M is the uniform density over this parameter set, since $l_i = Y_i - a_i$ and $l_{ij} = -a_{ij}$, so that $L_{k,ij} = 0$. This density may be improper.

For observations $\mathbf{x} = x_1, \ldots, x_n$ from f_θ, the maximum likelihood estimate solves the equations

$$\sum_j \left(Y_i(x_j) - a_i(\hat{\theta}) \right) = 0, \qquad 1 \le i \le p.$$

The loss function $K(\theta, d) = (\theta_i - d_i)a_i(\theta) - a(\theta) + a(d)$ is a convex function in d. The Bayes estimator $d = \theta^M$ has the minimum posterior risk when

$$\mathbb{P}_M \left[\frac{\partial K}{\partial d_i} \,\middle|\, \mathbf{x} \right] = \mathbb{P}_M [\, a_i(d) - a_i(\theta) \mid \mathbf{x}] = 0.$$

For those $\mathbf{x}$ for which $f_\theta(\mathbf{x}) = 0$ whenever θ is on the boundary of Θ,

$$0 = \mathbb{P}_M \frac{\partial f_\theta}{\partial \theta_i} = \mathbb{P}_M \left[\sum_j \left(Y_i(x_j) - a_i(\theta) \right) \,\middle|\, \mathbf{x} \right].$$

Thus $n^{-1} \sum_j Y_i(x_j) = a_i(\theta_M)$, and the maximum likelihood estimate and the M-Bayes estimate coincide. It was noted in Hartigan (1983) that the prior M has posterior mean $n^{-1} \sum Y_i(x_j) = \mathbb{P}_M[\, a_i(\theta) \mid \mathbf{x}]$. In the asymptotic theory, we need the estimate θ_ε^M that minimizes *truncated* expected posterior Kullback–Leibler loss $\mathbb{P}^\varepsilon[\, K(\theta, d) \mid x]$; it may be shown that $\theta_\varepsilon^M = \theta^M + O_p(\sqrt{n}\, e^{-n\varepsilon})$.

For example, in the p-dimensional normal location problem

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \sum [\, Y_i(x) - \theta_i]^2 \right),$$

the parameter space is R^p, M is uniform, $\hat{\theta}_i = n^{-1} \sum_j Y_i(x_j)$, and the posterior mean is $n^{-1} \sum_j Y(x_j)$ coinciding with the maximum likelihood estimator.

In the binomial case, n observations from a Bernoulli,

$$f_\theta(x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x},$$

the parameter space is the open unit interval $(0, 1)$, M is $\log(\theta/1 - \theta)$ uniform, $\hat{\theta} = x/n$ and the posterior mean is x/n, provided $x \neq 0$, $x \neq n$; in the extreme cases the likelihood does not vanish on the boundary, and in fact

the posterior densities are improper. The extreme cases have negligible asymptotic probability for $0 < \theta < 1$.

5. Asymptotic admissibility. We have seen that *bias-adjusted* estimators $\theta_i^* = \hat{\theta}_i + \hat{L}^{-1}_{i,j}(\hat{q}_j^* - \hat{q}_j)/n$ have asymptotic risk $(\partial/\partial\theta_i)(L^{-1}_{i,j}q_j^*) + \frac{1}{2} L^{-1}_{i,j}q_i^*q_j^*$. Thus the asymptotic admissibility of θ^*, following Brown (1979) and Levit (1982, 1983, 1985), is determined by a simple differential operator, depending only on the functions q_{i*}, the information matrix L and the parameter space Θ. First, let us show that we need only consider Bayes estimates as candidates for asymptotically admissible θ^*.

The complete class theorem due to Levit (1983) holds under conditions A and for certain symmetric bounded loss functions w.

LEVIT'S THEOREM. *For each bias-adjusted estimator θ^*, there exists a Bayes estimator θ^h of the same asymptotic risk.*

Although the form of the loss function is usually not crucial in these asymptotic studies, I have been able to prove only a weaker version of Levit's theorem for Kullback–Leibler loss.

THEOREM 6. *For each bias-adjusted estimator θ^*, there exists a Bayes estimator θ^h of no greater asymptotic risk.*

One consequence of Theorem 6, because the asymptotic risk is convex as a function of θ^*, is that if θ^* is not a Bayes estimator, it is inadmissible. Suppose that θ^h is the Bayes estimator of no greater asymptotic risk. The estimator $\theta' = \frac{1}{2}(\theta^* + \theta^h)$ has asymptotic risk

$$A(\theta', \theta) = \tfrac{1}{2}\big[A(\theta^*, \theta) + A(\theta^h, \theta)\big] + \tfrac{1}{2}L^{-1}_{ij}(\theta^* - \theta^h)_i(\theta^* - \theta^h)_j,$$

which is never greater than $A(\theta^*, \theta)$ and sometimes less, since L is negative definite and $\theta^* \neq \theta^h$ by assumption.

In particular, maximum likelihood is asymptotically inadmissible unless the maximum likelihood prior exists, that is for each i, j, $\partial q_i/\partial\theta_j = \partial q_j/\partial\theta_i$. These conditions are satisfied for exponential families, for location parameter models, for mixture models with parameters the mixing probabilities. An example where the conditions fail has observations drawn from $N(0, 1)$ with probability p and $N(0, \sigma^2)$ with probability $1 - p$.

When maximum likelihood fails, I would suggest using instead the least favorable prior beating it, which may be obtained from the Dirichlet minimization $\alpha = \inf_{h \in H}[D(h) = \int (L^{-1}_{i,j}(h_i - q_i)(h_j - q_j)h\, d\theta]$ over the class H of priors with $\int_\Theta h\, d\theta = 1$ and $h = 0$ on the boundary of Θ. Suppose $0 \in \Theta$. As in the proof of Theorem 6, we define the least favorable prior h^* as the limit of $h^k/h^k(0)$ as $D(h^k) \to \alpha$, if that limit exists.

The asymptotic risk of the Bayes estimate θ^h is

$$A(\theta_0, \theta^h) = 2h^{-1/2}\Delta h^{1/2} = 2h^{-1/2}\frac{\partial}{\partial\theta_i}\left[L^{-1}_{i,j}\frac{\partial}{\partial\theta_j}\right]h^{1/2},$$

where Δ denotes an elliptical differential operator. Thus, θ^h is admissible by definition if and only if there is no nonzero nonnegative function ψ for which there is a positive solution $u = h^{1/2}_{\psi}$ to the elliptic differential equation in θ,

$$\Delta u + (\psi + h^{-1/2}\Delta h^{1/2})u = 0.$$

Consider now asymptotic admissibility of maximum likelihood in the exponential family case where the maximum likelihood prior is the uniform in the canonical parameter. The qualification *asymptotic* will be dropped in the following discussion. Maximum likelihood is admissible if and only if there is no nonconstant positive *superharmonic* function u for which $\Delta u \leq 0$ in Θ [Brown (1971), Levit (1982, 1983)].

In one dimension, the maximum likelihood estimate is the solution to $\sum Y(x_i) = na'(\hat{\theta})$; the above condition reduces to $\hat{\theta}$ being admissible if and only if the range of $a'(\theta)$ (the range of expected values of the maximum likelihood estimate), is the whole real line. Thus for the Poisson, $a'(\theta) = \exp(\theta)$ and the maximum likelihood estimate $\sum Y(x_i)/n$ is inadmissible, being beaten in asymptotic risk by $\sum Y(x_i)/(n-1)$ by the amount $\frac{1}{2}e^{-\theta}$ corresponding to the prior density $h = e^{\theta}$. Similarly, maximum likelihood is inadmissible for the binomial and for normal scale parameters and admissible for normal location. However, if a single point, say $\theta = 0$, in the normal location family, is excluded from Θ, then the maximum likelihood estimate becomes inadmissible.

The *expectation parameters* $\phi_i = (\partial/\partial\theta_i)a$ provide a 1–1 transformation from Θ into Φ say, because a is convex. In view of the one-dimensional results, we examine admissibility questions in the Φ space rather than the Θ space. Now $\Delta = a_{ij}(\partial/\partial\phi_i)(\partial/\partial\phi_j)$ and following Brown (1971) and Levit (1985), the admissibility of maximum likelihood is equivalent to the recurrency of a diffusion in Φ with local variance–covariance a_{ij}.

Suppose that there exists a hyperplane $H(\alpha, \beta)$ such that $\alpha'\phi > \beta$ for $\phi \in \Phi$; then the prior density $u^2 = (\alpha'\phi - \beta)$ satisfies

$$\Delta u = -\tfrac{1}{4}\alpha_i\alpha_j a_{ij}/u^3$$

and so beats maximum likelihood, moving the maximum likelihood estimate away from the hyperplane.

We might ask, for a parameter space Φ within which maximum likelihood is admissible, which subspaces Φ' retain admissibility? If we delete a set A which has the property for some spheres S, S', S'' with S', $S'' \subset S \subset \Phi$ that all paths between S' and S'' that lie in S pass through A, then $\Phi - A$ will be inadmissible. (Since the recurrent diffusion in Φ will occasionally pass from S' to S'' in S and so pass through A, the diffusion in $\Phi - A$ will be absorbed by the boundary A.) For example, with normal location in two

dimensions, removal of a slit from the plane is enough to render maximum likelihood inadmissible. Removal of a finite number of points will leave maximum likelihood admissible.

The classic Stein (1956) result has maximum likelihood admissible in one and two dimensions, and inadmissible in three or more dimensions, for multivariate normal location in the finite sample case. How do these results generalize to asymptotic admissibility in arbitrary exponential families? Let us define the *condition number* $c(a)$ to be the ratio of the largest to the smallest eigenvalue of a_{ij}. Let Θ be R^p; I conjecture that a has the same admissibility status as $a = I$ when $c(a)$ is uniformly bounded. The condition bounds the partial correlations corresponding to a away from 1 and bounds the ratios of variances in any two coordinate directions away from zero. Thus the diffusion acts as if the coordinate directions are proceeding independently with the same variance.

How do the asymptotic admissibility results relate to finite sample admissibility results? In general, not very well. The risks converge only pointwise, not uniformly. Thus, near the boundary it is quite possible for the finite n risks to be quite far from the asymptotic risk and for a procedure that is everywhere beaten asymptotically at interior points to have superior performance close to the boundary for finite n. For example, with normal location, $\Theta = (0, \infty)$, the prior density θ has asymptotic risk $-1/2\theta^2$ beating maximum likelihood everywhere; but at $\theta = 0$, the finite sample risk for this prior is $1/n$, and the finite sample risk for maximum likelihood is $1/4n$. Thus in the neighborhood of 0, maximum likelihood is better in the first-order term. It is possible to produce an asymptotic risk smaller than that of maximum likelihood at every interior point and also at the point $\theta = 0$ by adding any positive atom at 0 to the prior density θ. However, for any finite n, maximum likelihood has smaller risk somewhere near 0. For any finite n, maximum likelihood is inadmissible, since, not being differentiable, it is not generalized Bayes, and there will be some smaller risk Bayes solution depending on n.

There is a special case, Brown (1971), where the asymptotic results do carry over to finite sample sizes. Consider the multivariate normal location problem in p dimensions, and take the loss function for an estimate $\theta^* = x + q(x)$ based on observation x to be $\frac{1}{2}W_{i,j}(x)(\theta^* - \theta)_i(\theta^* - \theta)_j$ where $W_{i,j}$ is a nonnegative definite function of x. Then the risk difference from maximum likelihood may be expressed, using Stein's estimating principle, by

$$\mathbb{P}_\theta\left[\frac{\partial}{\partial x_i}(W_{ij}q_j) + \frac{1}{2}W_{ij}q_iq_j\right].$$

If we can choose q to make the differential expression negative, we will beat maximum likelihood. We need only consider q corresponding to prior distributions h on R^p. Thus we can show that maximum likelihood is inadmissible in the finite sample case if we can discover a nonconstant positive solution u to $(\partial/\partial x_i)(W_{ij}(\partial/\partial x_j)u) \leq 0$, just as in the asymptotic risk case.

6. Density estimation. We have defined the loss for an estimate θ^h as the Kullback–Leibler distance between the densities f_θ and f_{θ^h}; when a prior density h is available, the optimal Bayes density estimate is the *predictive* density $\hat{f} = f_h(x \mid \mathbf{x})$ which minimizes the expected Kullback–Leibler distance given $\mathbf{x}$, $\int K(f_\theta, \hat{f}) h(\theta \mid \mathbf{x})\, d\theta$. The *estimative density* f_{θ^h} turns out to be the density restricted to the family $f_\theta(x)$ that is closest to f_h in Kullback–Leibler loss $L(f_\theta, f_h)$. It is of interest to compare the risk of the predictive density with that of the estimative density, $\mathbb{P}_\theta[K(f_\theta, f_{\theta^h}) - K(f_\theta, f_h)]$.

Let

$$f_{i_1 \cdots i_r} = \frac{\partial^r f}{\partial\theta_{i_1} \cdots \partial\theta_{i_r}}.$$

THEOREM 7. *Under Assumptions* A, *when* θ_0 *is true*,

$$\mathbb{P}^{\varepsilon}_{\theta_0}\left[K(f_{\theta_0}, f_{\theta^h}) - K(f_{\theta_0}, f_h)\right] = \tfrac{1}{8} n^{-2} \operatorname{var}\left\{L^{-1}_{i,j}\left(f_{ij} - L^{-1}_{k,l} f_l (L_{i,j,k} + L_{ij,k})\right)/f\right\} + o(n^{-2}).$$

For example, in the normal location case, the gain is $p/4n^2$. Thus there is a loss in using the estimative density rather than the predictive density which is equal to $1/8n^2$ times the residual variance after predicting $L^{-1}_{i,j} f_{ij}/f$ as a linear function of f_i/f. The difference between the estimative log density and the predictive log density is

$$\log f_h - \log f_{\theta^h} + t_n/n = \frac{1}{2n}\left\{\hat{L}^{-1}_{i,j}\left(\hat{f}_{ij} - \hat{L}^{-1}_{k,l}\left(\hat{L}_{i,j,k} + \hat{L}_{i,jk}\right)\hat{f}_l\right)/\hat{f}\right\},$$

which is orthogonal to densities f_θ in the neighborhood of the maximum likelihood estimate; that is, the expectation of the expression and of its multiple with $\hat{f}_m$ is zero at $\theta = \hat{\theta}$. The difference in K–L risk is half the average square of this difference in log densities evaluated at θ_0. It is remarkable that the improvement due to using the predictive density does not depend on the prior used, nor does the difference in log density. The improvement is positive except when $L^{-1}_{i,j} f_{ij}$ is a linear function of the f_i; there is no improvement in a bounded open set Θ' if and only if f_θ is a mixture of the densities on the boundary of Θ', with mixing probabilities (depending on θ) all being solutions of the elliptic differential equation $L^{-1}_{i,j} f_{ij} = a_i f_i$.

Indeed, we can improve the estimative density for **any** bias-adjusted estimator by the same addition in the log density, resulting in the same decrease in Kullback–Leibler risk.

Aitchison (1975) argues for the value of predictive rather than estimative densities in general and gives particular arguments in the case of gamma and multivariate normal exponential families. Komaki (1996) considers optimal adjustments of estimative to predictive estimators for exponential families.

7. Proofs.

PROOF OF THEOREM 1 (Asymptotic risk of maximum likelihood). Algebra is simplified by assuming that $\theta_0 = 0$; we show that the maximum likelihood estimate $\hat{\theta}$ may be adequately approximated, with truncatable errors, by a polynomial in $l_i, l_{ij}, l_{ijk}, l_{ijkl}$. Then the Kullback–Leibler loss is approximated by a polynomial in $\hat{\theta}$, thus by a polynomial in the first four l's, from which the truncated expectation is computed.

At the maximum likelihood estimate, the first derivatives of the log likelihood are zero, for $|\hat{\theta}| < \varepsilon$,

$$0 = l_i(\hat{\theta}) = l_i + \hat{\theta}_j l_{ij} + \tfrac{1}{2}\hat{\theta}_i\hat{\theta}_j l_{ijk} + \tfrac{1}{6}\hat{\theta}_j\hat{\theta}_k\hat{\theta}_l l_{ijkl} + n|\hat{\theta}|^4 R,$$

where $\sup_{\hat{\theta} < \varepsilon} R = T_n$. From A2, R is bounded by an average of n independent identically distributed terms having all moments; the probability that R exceeds K for some K is $o(n^{-2})$, which justifies the T_n term necessary for the truncated expectation calculations.

Standardize the l's by

$$nm_i = L^{-1}_{i,j} l_j,\ nm_{ij} = L^{-1}_{i,k}(l_{kj} - nL_{kj}),\ nm_{ijk} = L^{-1}_{i,l} l_{ljk},\ nm_{ijkl} = L^{-1}_{i,r} l_{rjkl}.$$

Note that m_i, m_{ij} are $O_p(n^{-1/2})$ and that m_{ijk}, m_{ijkl} are $O_p(1)$. Successively solving the polynomial equations in $\hat{\theta}$ of increasing degree, obtain

$$\hat{\theta}_i = m_i + \left[m_{ij}m_j + \tfrac{1}{2}m_{ijk}m_j m_k\right] + m_{ij}\left[m_k m_{jk} + \tfrac{1}{2}m_{jkl}m_k m_l\right] + m_{ijk}m_k\left[m_l m_{jl} + \tfrac{1}{2}m_{jrs}m_r m_s\right] + \tfrac{1}{6}m_{ijkl}m_j m_k m_l + nT_n(m_r m_r)^2.$$

The error term is arrived at by substituting this expression into the Taylor series expansion of $l_i(\hat{\theta})$.

The Taylor series expansion of $\log f_{\hat{\theta}}(x)$ about $\theta = 0$ gives, for $|\hat{\theta}| < \varepsilon$,

$$-K(\theta_0, \hat{\theta}) = \tfrac{1}{2}\hat{\theta}_i\hat{\theta}_j L_{ij} + \tfrac{1}{6}\hat{\theta}_i\hat{\theta}_j\hat{\theta}_k L_{ijk} + \tfrac{1}{24}\hat{\theta}_i\hat{\theta}_j\hat{\theta}_k\hat{\theta}_l L_{ijkl} + t_n n^{-2}.$$

Replacing $\hat{\theta}$ by its polynomial approximation in $m_i, m_{ij}, m_{ijk}, m_{ijkl}$, $-K(\theta_0, \hat{\theta})$ is approximated by a polynomial in those terms with error $t_n n^{-2}$, for $|\hat{\theta}| < \varepsilon$. The truncated expectation of each term in the polynomial differs negligibly from the expectation. Contributions to the truncated expectation from $|\hat{\theta}| > \varepsilon$ are negligible by A6.

In evaluating the expectations of polynomials in the m's, we make frequent use of the fundamental cumulant identity [McCullagh (1987)], that an expectation of a product of random variables is the sum, over all partitions of variables into components, of the product of the cumulants of the sets of variables in the different components. Note that the cumulant of a set of sums of independent identical random variables equals n times the cumulant of the individual random variables. Only cumulants in the m of order 4 or less (having four or fewer variables) appear in the final expression. For example, one of the terms in the polynomial is $\frac{1}{6}m_i L_{ir} m_{rjkl} m_j m_k m_l$. None of the terms involving cumulants of $m_i m_{rjkl}$ need to be considered because they

are of order more than 4. The cumulant $\kappa_{m_i, m_j, m_k, m_l}$ is negligible because it is order n/n^4. Thus the only nonnegligible term in the expectation is $\frac{1}{2}M_{rjkl}L_{ir}M_{i,j}M_{k,l}$. Similar computations show that the asymptotic risk $\mathbb{P}^{\varepsilon}_{\theta_0}\hat{K}(\theta_0, \hat{\theta})$ is as in Theorem 1. □

PROOF OF THEOREM 2 (Asymptotic posterior distribution and moments). To ease notation, take $\hat{\theta} = 0$. Note that the event $\hat{l}_{ij}$ negative definite occurs with probability $1 - o(n^{-2})$. For $|\theta_0| < \varepsilon$, $|\theta| < \varepsilon$,

$$\log\bigl(h(\theta)f_\theta(\mathbf{x})\bigr) = \hat{l} + \log h(\hat{\theta}) + \theta_i\hat{h}_i + \tfrac{1}{2}\theta_i\theta_j\hat{l}_{ij} + \tfrac{1}{6}\theta_i\theta_j\theta_k\hat{l}_{ijk} + \bigl(|\theta|^2 + n|\theta|^4\bigr)R_n(\theta),$$

where $\sup_{|\theta|<\varepsilon} R_n(\theta) = T_n$,

$$\int_{|\theta|<\varepsilon} h(\theta)f_\theta(\mathbf{x})\,d\theta = \hat{h}\exp(\hat{l})(2\pi)^{p/2}\det{}^{-1/2}\bigl(-\hat{l}_{ij}\bigr)\bigl[1 + T_n/n\bigr].$$

It follows from A7, which bounds the tail probability, that $\int h(\theta)f_\theta(\mathbf{x})\,d\theta$ satisfies the same condition, so that

$$h(\theta \mid \mathbf{x}) = \det{}^{1/2}\bigl(-\hat{l}_{ij}/2\pi\bigr)\exp\Bigl[\tfrac{1}{2}\theta_i\theta_j\hat{l}_{ij} + E\Bigr]\Bigl[1 + \theta_i\hat{l}_i + \tfrac{1}{6}\theta_i\theta_j\theta_k\hat{l}_{ijk}\Bigr].$$

where $E = (1 + n|\theta|^2)^2R_n(\theta)/n]$, $\sup_{|\theta|<\varepsilon} R_n(\theta) = T_n$ as required.

Since the expansion of the posterior density holds only for $|\hat{\theta} - \theta_0| < \varepsilon$, $|\theta - \hat{\theta}| < \varepsilon$, we may compute only the posterior truncated moments, which requires integration of the Edgeworth expansion over the region $|\theta - \hat{\theta}| < \varepsilon$. First, ignoring the error term, the averages of polynomials in θ over this region differ by exponentially small terms from the Gaussian averages over R^p, producing the formulas for truncated moments given in Theorem 2, without the error terms.

Again setting $\hat{\theta} = 0$, we evaluate the effect on posterior truncated moments of the error term $E = (1 + n|\theta|^2)^2R_n/n$ by breaking the integrals into two regions, an outer one where $\sqrt{n}\,\varepsilon > \sqrt{n}|\theta| > A\sqrt{\log n}$, and an inner one where $\sqrt{n}|\theta| \le A\sqrt{\log n}$. Note that $|E| < n|\theta|^2\varepsilon^2T_n$. Since L_{ij} is negative definite by A3, the quadratic term $\frac{1}{2}\theta_i\theta_j\hat{l}_{ij}$ is less than $-Kn|\theta|^2$ for all θ with probability $1 - o(n^{-2})$. This term dominates the error term for ε small enough, and thus the integral over the outer region will be of size n^{-A^2k}. Choosing A large enough, the contribution to the truncated moments from the outer region is negligible.

In the inner region, the error term is bounded by $T_n(A\log n)^2/n$, and we use the bound $|e^E - 1 - E| < K_\alpha E^2$ for $|E| < \alpha$. The term $K_\alpha E^2$ makes a contribution $T_n(\log n/\sqrt{n})^{d+4}$ to a moment of order d, which is less than T_n/n^2 for all moments. Replace e^E by $1 + E$, and expand E more accurately as $(T_n^0P_0 + T_n^2P_2 + T_n^4P_4 + T_n^6P_6)/n + (|\theta|^3 + n|\theta|^5)R_n(\theta)$, where P_i is a fixed polynomial of degree i in $\sqrt{n}\,\theta$, and the T_n^i denote different tail bounded random variables not depending on θ. Considering the truncated posterior mean, the truncated posterior expectation of $\theta T_nP_i/n$ with respect to the

quadratic exponent is T_n/n^2 for each even polynomial, and the posterior expectation of $\theta(|\theta|^3 + n|\theta|^5)R_n$ is also T_n/n^2. The contributions from E will be a fortiori bounded by T_n/n^2 for all higher truncated posterior moments, as required. □

PROOF OF THEOREM 3 (Asymptotic equivalence of $\hat{\theta}$ and Bayes estimate). The Bayes estimate θ^h minimizes $\mathbb{P}^\varepsilon[K(\theta, \theta^h) \mid \mathbf{x}]$. Assume that $|\hat{\theta} - \theta_0| < \varepsilon$. Then $g(\theta \mid \mathbf{x})$ gives probability t_n/n^2 to the region $|\theta - \hat{\theta}| > \varepsilon$. By A8, $|\theta - \theta^h| < \varepsilon$ when $K(\theta, \theta^h) < \delta$. Thus if $|\theta^h - \hat{\theta}| > 2\varepsilon$, the region $|\theta - \theta^h| > \varepsilon$ has probability $1 - t_n/n^2$, and the posterior risk exceeds $\delta(1 - t_n/n^2)$. We will show that an estimate in $|\theta^h - \hat{\theta}| < \varepsilon$ is available with smaller risk.

The Bayes estimate θ^h will solve

$$\mathbb{P}\left[\{K < \varepsilon\}\frac{\partial}{\partial\theta_i^h}K(\theta, \theta^h) \mid \mathbf{x}\right] = 0,$$

$$\begin{aligned}
-\frac{\partial}{\partial\theta_i^h}K(\theta, \theta^h) &= (\theta^h - \theta)_j L_{ij} + \frac{1}{2}(\theta^h - \theta)_j(\theta^h - \theta)_k L_{ijk} + |\theta^h - \theta|^3 M \\
&= (\theta^h - \theta)_j \hat{L}_{ij} + (\theta^h - \theta)_j(\theta - \hat{\theta})_k\left[\hat{L}_{ijk} + \hat{L}_{ij,k}\right] \\
&\quad + \frac{1}{2}(\theta^h - \theta)_j(\theta^h - \theta)_k \hat{L}_{ijk} + \delta^3 M,
\end{aligned}$$

where $\delta = \max(|\theta - \theta^h|, |\theta - \hat{\theta}|)$ and M is bounded for $\delta < \varepsilon$.

Using the truncated posterior moments of θ given in Theorem 2, we obtain

$$\theta_i^h = \hat{\theta}_i - \hat{l}_{ij}^{-1}\hat{h}_j + \tfrac{1}{2}\hat{l}_{ij}^{-1}\hat{l}_{kl}^{-1}\hat{l}_{jkl} + \tfrac{1}{2}\hat{L}_{ij}^{-1}\hat{l}_{lk}^{-1}\left(\hat{L}_{j,l,k} + \hat{L}_{j,lk}\right) + T_n n^{-2},$$

which simplifies, using A4 and replacing $\hat{L}, \hat{l}$ by L, to

$$\theta_i^h - \hat{\theta}_i + T_n n^{-3/2} = L_{i,j}^{-1}\left[h_j - L_{k,l}^{-1}L_{jk,l}\right]/n = L_{i,j}^{-1}\left[h_j - q_j\right]/n,$$

which establishes Theorem 3. Note that the risk of this estimate is $O(n^{-1})$, ruling out any estimate with $|\theta^h - \hat{\theta}| > 2\varepsilon$. □

PROOF OF THEOREM 4 (Asymptotic risk of Bayes estimators). Let $\theta_0 = 0$. Define $\delta_i = \theta_i^h - \hat{\theta}_i = L_{i,j}^{-1}(h_j - q_j)/n + T_n/n^{3/2}$,

$$\begin{aligned}
-K\left(\theta_0, \theta^h\right) &= \tfrac{1}{2}\theta_i^h\theta_j^h L_{ij} + \tfrac{1}{6}\theta_i^h\theta_j^h\theta_k^h L_{ijk} + R(\theta^h), \\
-K\left(\theta_0, \hat{\theta}\right) &= \tfrac{1}{2}\hat{\theta}_i\hat{\theta}_j L_{ij} + \tfrac{1}{6}\hat{\theta}_i\hat{\theta}_j\hat{\theta}_k L_{ijk} + R(\hat{\theta}), \\
-K\left(\theta_0, \theta^h\right) + K\left(\theta_0, \hat{\theta}\right) &= \tfrac{1}{2}\delta_i\delta_j L_{ij} + \delta_i\hat{\theta}_j L_{ij} + \tfrac{1}{2}\delta_i\hat{\theta}_j\hat{\theta}_k L_{ijk} + T_n n^{-5/2}.
\end{aligned}$$

Only the constant term $L_{i,j}^{-1}(h_j - q_j)/n$ in δ needs to be considered in computing truncated expectations, except for the term $\mathbb{P}^\varepsilon\delta_i\hat{\theta}_j L_{ij}$.

EXPECTATION LEMMA. *Let S be any set of indices containing more than one member. Then* $\mathbb{P}^\varepsilon[\hat{\theta}_i(\hat{l}_S - n\hat{L}_S)] = O(n^{-1/2})$.

PROOF.

$$
\begin{aligned}
n\hat{\theta}_i &= L^{-1}_{i,j} l_j + T_n,\\
\hat{l}_S &= l_S + \hat{\theta}_j n L_{Sj} + T_n,\\
\hat{L}_S &= L_S + \hat{\theta}_j n\big(L_{Sj} + L_{S,j}\big) + T_n,\\
n\hat{\theta}_i\big(\hat{l}_S - n\hat{L}_S\big) &= L^{-1}_{i,j} l_j\big[l_S + L^{-1}_{u,k} l_k L_{Su} - nL_S - nL^{-1}_{u,k} l_k (L_{Su} + L_{S,u})\big]\\
&\quad + T_n\sqrt{n}\,,\\
\mathbb{P}^{\varepsilon}\hat{\theta}_i\big(\hat{l}_S - n\hat{l}_S\big) &= L^{-1}_{i,j}\big(L_{j,S} + L_{jS} - L_{j,S} - L_{jS}\big) + O(n^{-1/2}) = O(n^{-1/2}).
\end{aligned}
$$

This lemma is useful because it permits replacement of $\hat{l}$ by $n\hat{L}$ terms in δ when computing $\mathbb{P}^{\varepsilon}\delta_i\hat{\theta}_j L_{ij}$ with error $o(n^{-2})$;

$$
\delta_i = -\hat{l}^{-1}_{ij}\hat{h}_j + \tfrac{1}{2}\hat{l}^{-1}_{ij}\hat{l}^{-1}_{kl}\hat{l}_{jkl} + \tfrac{1}{2}\hat{L}^{-1}_{ij}\hat{l}^{-1}_{lk}\big(\hat{L}_{j,l,k} + \hat{L}_{j,lk}\big) + T_n n^{-2}
$$

is replaced by

$$
\hat{L}^{-1}_{ij}\big(\hat{q}_i - \hat{h}_i\big)/n = L^{-1}_{i,j}[h_i - q_i]/n + \hat{\theta}_k\frac{\partial}{\partial\theta_k}\big(L^{-1}_{i,j}[h_i - q_i]\big)/n + T_n n^{-2}.
$$

It then follows that

$$
\begin{aligned}
&\mathbb{P}^{\varepsilon}_{\theta_0}\big[K\big(\theta_0,\theta^h\big) - K\big(\theta_0,\hat{\theta}\big)\big]\\
&\quad = n^{-2}\left[\frac{\partial}{\partial\theta_i}\big(L^{-1}_{i,j}\big[h_j - q_j\big]\big) + \frac{1}{2}L^{-1}_{i,j}\big(h_i h_j - q_i q_j\big)\right] + o(n^{-2}). \qquad \square
\end{aligned}
$$

PROOF OF THEOREM 5 (Asymptotic Bayes risk). The asymptotic risk of θ^* is given by: $(\partial/\partial\theta_i)(L^{-1}_{i,j}q^*_j) + \frac{1}{2}L^{-1}_{i,j}q^*_i q^*_j$. Since $h \to 0$ on the boundary, fixing all coordinates except θ_i,

$$
0 = hL^{-1}_{i,j}q^*_j\Big|^{\max\theta_i}_{\min\theta_i} = \int\left[\frac{\partial}{\partial\theta_i}\big(L^{-1}_{i,j}q^*_j\big) + L^{-1}_{i,j}q^*_j h_i\right]h\,d\theta_i.
$$

Thus the difference in Bayes risks becomes the integral of the quadratic in h_i, q^*_i,

$$
B(\theta^*) - B(\theta^h) = \tfrac{1}{2}\int L^{-1}_{i,j}(h_i - q^*_i)(h_j - q^*_j)h\,d\theta.
$$

PROOF OF THEOREM 6 (Bayes estimates form a complete class). The proof follows Levit (1982, 1983) with adjustments to allow for Kullback–Leibler loss. In the first case, assume that Θ is open bounded and connected with a smooth boundary [see Levit (1982)], and that conditions A hold uniformly for all points in Θ; that is, each moment of log likelihood derivatives and the first derivatives of the moments are uniformly bounded, and the information matrix $-L$ is bounded below by a positive definite matrix. Define the differ-

ential operator $\Delta = (\partial/\partial\theta_i)(L^{-1}_{i,j}(\partial/\partial\theta_j))$. Let $r = (\partial/\partial\theta_i)(L^{-1}_{i,j}q^*_j) + \frac{1}{2}L^{-1}_{i,j}q^*_i q^*_j$. Let $u = h^{1/2}$. From Levit (1982), Theorem 5, the minimum α of the Bayes risk difference $B(\theta^*) - B(\theta^h)$ for prior densities h that satisfy $h = 0$ on the boundary of Θ and $\int_\Theta h(\theta)\,d\theta = 1$ occurs for a unique positive "least favorable" twice differentiable h^* satisfying the elliptic differential equation

$$\Delta u^* + (\alpha - r)u^* = 0.$$

Since $\alpha \geq 0$ by definition of the Bayes risk, and

$$\Delta u^* - ru^* = u^*\big[A(\theta, \theta^{h^*}) - A(\theta, \theta^*)\big],$$

it follows that Theorem 6 holds for this particular Θ.

For general Θ, we follow Levit (1983). If Θ decomposes into connected open sets, we can prove the theorem separately within each open component, so it is sufficient to prove the theorem for a connected open set. Assuming $0 \in \Theta$, we approximate Θ by a sequence of connected bounded open sets Θ_i with smooth boundaries, such that $0 \in \Theta_i$, $\overline{\Theta}_i < \Theta_{i+1}$, $\bigcup \Theta_i = \Theta$. Now consider the least favorable solutions u_i for the parameter set Θ_i, standardized so that $u_k(0) = 1$; since the minimization is over a larger range of priors h for Θ_{i+1} than for Θ_i, the minimum Bayes risks α_i are nonincreasing in i and converge to some α^*, say. We need to show that the sequence of prior root solutions u_k converges on some subsequence in Θ_i; we follow a standard procedure using a version of Harnack's inequality [e.g., Taylor (1996), page 349] to bound the variation of the solutions u_k within Θ_i.

A version of Harnack's inequality. Suppose that Θ is a connected bounded open set on which $\Delta u - ru = 0$. Suppose that Θ' is a connected open set with $\overline{\Theta}' < \Theta$. Suppose that there exist positive definite matrices L^1, L^2 and a bound C such that, for all $\theta \in \Theta$, $L^1 \leq L \leq L^2$, $|(\partial/\partial\theta_i)L^{-1}_{j,k}| \leq C$, $|r| \leq C$. Then for ε small enough, whenever $|\theta_1 - \theta_2| < \varepsilon$, and $\theta_1, \theta_2 \in \Theta'$,

$$w(\theta_1) < w(\theta_2)k(\varepsilon, L^1, L^2, C),$$

where $k(\varepsilon) \to 1$ as $\varepsilon \to 0$.

For each $\delta > 0$, cover Θ_i with a finite number of spheres within each of which $u_k(\theta_1)/u_k(\theta_2) < (1 + \delta)$, for $k > i$. Since $u_k(0) = 1$, for each $\theta \in \Theta_i$, $u_k(\theta)$ is bounded away from 0 and infinity and converges to some positive $u^*(\theta)$ on a subsequence. Convergence to, say, u^* may be then obtained on some subsequence for a set of θ values dense in Θ_i. Harnack's inequality shows that the function u^* obtained is continuous on the dense set, and so defines a continuous function on Θ_i. Now the solution u to the elliptic equation $\Delta u + (\alpha - r)u = 0$ is characterized by satisfying for each sphere of radius r in Θ,

$$u(\xi) = \int_{|\theta - \xi| \leq r} \Delta u k(d\theta, \xi) = \int_{|\theta - \xi| \leq r} (\alpha - r)u(\theta)k(d\theta, \xi)$$

for a kernel function k (which may be interpreted as the density at θ of the expected number of times a diffusion beginning at ξ passes through the element $d\theta$ before reaching the boundary). The limiting u^* will satisfy these equations with $\alpha = \alpha^*$ and so will satisfy on Θ_i the elliptic equation

$$\Delta u^* + (\alpha^* - r)u^* = 0.$$

A subsequence of such functions defined on each Θ_i will converge to a solution satisfying the elliptic equation on Θ, which concludes the proof. □

PROOF OF THEOREM 7 (Predictive densities better than estimative densities). We need a generic bound $M(x, \hat{\theta}, \theta)$ for Taylor series expansions dependent on the new observation x, the maximum likelihood estimate $\hat{\theta}$ and a parameter θ. By definition, $\sup_{|\theta - \hat{\theta}| < \varepsilon, |\theta - \theta_0| < \varepsilon} M(x, \hat{\theta}, \theta)$ is assumed to have finite moments, averaging over $x, \mathbf{x}$ when θ_0 is true. Assume $|\hat{\theta} - \theta_0| < \varepsilon$.

Then, expanding about $\hat{\theta} = 0$, and using A2, A7,

$$\begin{aligned}
f_\theta(x) &= \hat{f}\Big[1 + \theta_i \hat{f}_i/\hat{f} + \tfrac{1}{2}\theta_i\theta_j \hat{f}_{ij}/\hat{f} + \tfrac{1}{6}\theta_i\theta_j\theta_k \hat{f}_{ijk}/\hat{f} + M|\theta|^4\Big], \\
f_h(x) &= \int f_\theta(x) h(\theta \mid \mathbf{x})\, d\theta \\
&= \int \{\theta - \theta_0 \mid < \varepsilon\} f_\theta(x) h(\theta \mid \mathbf{x})\, d\theta \big(1 + Mt_n n^{-2}\big) \\
&= \hat{f}\Big[1 + \mathbb{P}^\varepsilon[\theta_i \mid \mathbf{x}]\hat{f}_i/\hat{f} + \tfrac{1}{2}\mathbb{P}^\varepsilon\big[\theta_i\theta_j \mid \mathbf{x}\big]\hat{f}_{ij}/\hat{f} + MT_n n^{-2}\Big], \\
f_h(x) &= f_{\theta^h}\Big[1 + a_i \hat{f}_i/\hat{f} + a_{ij}\hat{f}_{ij}/\hat{f} + MT_n n^{-2}\Big],
\end{aligned}$$

where $a_i = -\frac{1}{2}\hat{l}^{-1}_{lk}\hat{L}^{-1}_{ij}[\hat{L}_{j,k,l} + \hat{L}_{j,lk}]$, $a_{ij} = -\frac{1}{2}\hat{l}^{-1}_{ij}$, from the posterior moments computed in proving Theorem 4. Note that the ratio of the predictive to the estimative densities does not involve the prior to this order of accuracy.

Define $\delta(x) = f_h(x)/f_{\theta^h}(x) - 1 = MT_n/n$. Note that if $\delta_0(x)$ is defined by replacing $\hat{l}$ in $\delta(x)$ with $n\hat{L}$, then $\int \delta_0 \hat{f}_i\, dx = 0$. This identity will enable us to show that the difference in log densities is orthogonal to $\hat{f}_i/\hat{f}$ to the required order of accuracy,

$$\begin{aligned}
K(f_\theta, f_{\theta^h} - K(f_\theta, f_h) &= \int \log(f_h/f_{\theta^h}) f_\theta(x)\, dx \\
&= \int \big[\delta - \tfrac{1}{2}\delta^2\big]\Big[1 + \theta_i \hat{f}_i/\hat{f} + \tfrac{1}{2}\theta_i\theta_j\hat{f}_{ij}/\hat{f}\Big]\hat{f}\, dx + t_n n^{-2}.
\end{aligned}$$

The only term that makes a random contribution to the expectation as $\hat{\theta}$ varies is $\int \delta(\theta_i - \hat{\theta}_i)\hat{f}_i\, dx$. Using the Expectation lemma of Theorem 4, we may replace $\hat{l}$ terms in δ by corresponding $n\hat{L}$ terms with eventual error $o(n^{-2})$, producing δ_0 satisfying $\int \delta_0 \hat{f}_i\, dx = 0$, so this term makes no contribu-

tion. The remaining terms, again using the orthogonality, give

$$\mathbb{P}^{\varepsilon}\left[K\left(f_\theta, f_\theta^h\right) - K\left(f_\theta, f_h\right)\right] = \frac{1}{2n^2}\mathbb{P}_{\theta_0}^{\varepsilon}\left[\int \delta_0^2 \hat{f}\, dx\right] + o\left(n^{-2}\right).$$

Since $\hat{\theta} - \theta_0 = t_n$, Theorem 7 follows. □

Acknowledgments. Andrew Barron, a referee and an Associate Editor discovered and helped correct numerous errors in the original manuscript and suggested significant improvements in presentation.

REFERENCES

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

AMARI, S. (1982). Differential geometry of curved exponential families: curvatures and information loss. *Ann. Statist.* **10** 357–387.

BERNARDO, J. M. (1979). Reference posterior densities for Bayesian inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 113–147.

BERGER, J. O. and BERNARDO, J. M. (1992). On the development of the reference prior method. In *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 35–60. Clarendon Press, Oxford.

BROWN, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–903.

BROWN, L. D. (1979). A heuristic method for determining admissibility of estimators with applications. *Ann. Statist.* **7** 960–994.

CLARKE, B. and BARRON, A. (1994). Jeffreys prior is asymptotically least favourable under entropy risk. *J. Statist. Plann. Inference* **41** 37–60.

GHOSH, J. K. (1994). *Higher Order Asymptotics*. IMS, Hayward, CA.

GHOSH, J. K. and SUBRAMANYAM, K. (1974). Second-order efficiency of maximum likelihood estimators. *Sankhyā Ser. A* **36** 325–358.

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

HARTIGAN, J. A. (1965). The asymptotically unbiased density. *Ann. Math. Statist.* **36** 1137–1152.

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

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

JEFFREYS, H. (1961). *Theory of Probability*. Oxford Univ. Press.

JOHNSON, R. A. (1967). An asymptotic expansion for posterior distributions. *Ann. Math. Statist.* **38** 1899–1907.

KOMAKI, F. (1996). On asymptotic properties of predictive distributions. *Biometrika* **83** 229–313.

LEVIT, B. YA. (1982). Minimax estimation and positive solutions of elliptic equations. *Theory Probab. Appl.* **27** 563–586.

LEVIT, B. YA. (1983). Second-order availability and positive solutions of the Schrödinger equation. *Lecture Notes in Math.* **1021** 372–385. Springer, Berlin.

LEVIT, B. YA. (1985). Second-order asymptotic optimality and positive solutions of Schrödinger's equation. *Theory Probab. Appl.* **30** 333–363.

MCCULLAGH, P. (1987). *Tensor Methods in Statistics*. Chapman Hall, London.

PERKS, W. (1947). Some observations on inverse probability, including a new indifference rule. *J. Inst. Actuaries* **73** 285–334.

PFANZAGL, J. and WEFELMEYER, W. (1978). A third-order optimal property of the maximum likelihood estimator. *J. Multivariate Anal.* **8** 1–29.

RAO, C. R. (1961). Asymptotic efficiency and limiting information. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 531–546. Univ. California Press, Berkeley.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 197–206. Univ. California Press, Berkeley.

STRASSEN, H. (1977). Asymptotic expansions for Bayes procedures. In *Recent Developments in Statistics* (Barra, J. L., eds.) 9–35. North-Holland, Amsterdam.

TAYLOR, M. E. (1997). *Partial Differential Equations III. Nonlinear Equations*. Springer, New York.

WELCH, B. L. and PEERS, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. Ser. B* **25** 318–329.

DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06520