TI 2011-130/4
Tinbergen Institute Discussion Paper

Asymptotically Informative Prior for Bayesian Analysis

Ao Yuan[1]
Jan G. De Gooijer[2]

[1] *Statistical Genetics and Bioinformatics Unit, National Human Genome Center, Howard University, Washington DC, USA;*
[2] *Department of Quantitative Economics, Faculty of Economics and Business, University of Amsterdam, and Tinbergen Institute, The Netherlands.*

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Asymptotically Informative Prior for Bayesian Analysis

Ao Yuan[1] and Jan G. De Gooijer[2]

[1] Statistical Genetics and Bioinformatics Unit
National Human Genome Center, Howard University
Washington DC, USA
e-mail: ayuan@howard.edu

[2] Department of Quantitative Economics and Tinbergen Institute
University of Amsterdam
Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
e-mail: j.g.degooijer@uva.nl

Abstract

In classical Bayesian inference the prior is treated as fixed, it is asymptotically negligible, thus any information contained in the prior is ignored from the asymptotic first order result. However, in practice often an informative prior is summarized from previous similar or the same kind of studies, which contains non-negligible information for the current study. Here, different from traditional Bayesian point of view, we treat such prior to be non-fixed. In particular, we give the data sizes used in previous studies for the prior the same status as the size of the current dataset, viewing both sample sizes as increasing to infinity in the asymptotic study. Thus the prior is asymptotically non-negligible, and its original effects are ressumed under this view. Consequently, Bayesian inference using such prior is more efficient, as it should be, than that regarded under the existing setting. We study some basic properties of Bayesian estimators using such priors under convex losses and the 0–1 loss, and illustrate the method by an example via simulation.

AMS 2000 subject classification: 62C10, 62C12

Key words and phrases: Asymptotically informative prior, asymptotic efficiency, Bayes estimator, information bound, maximum likelihood estimator.

1 Introduction

In classical Bayesian inference, the prior is treated as fixed. Although it is known that if we have a 'good' informative prior, the Bayes estimate has small sample advantage over the frequentist MLE or over Bayes estimate with a non-informative prior, its role is ignored as the data sample size increases, since the prior is asymptotically negligible, in the first order, by the existing Bayesian theory. Although all admissible procedures to a decision problem, including the MLE, can be formulated as Bayesian, or limit of Bayesian procedures (Wald, 1950), in practice the main stream statistical tool is still frequentist. Efron (2005) summarized the main reasons for this as the ease of use, modeling, computation and objectivity of the latter. Especially, when one does not have 'good' prior information, often various forms of non-informative priors (Jeffreys, 1961; Bernardo, 1979) or objective priors (Welch and Peers, 1963; Mukerjee and Ghosh, 1997) are used in Bayesian inference, to avoid possible misleading small sample effects of using a 'bad' subjective prior. But then the motivation of doing a Bayesian analysis is not clear, paying the price for more modeling and computational complexity with no apparent advantage. In our opinion, we use a method with increased complexity only if it has some advantage over other methods. Indeed, such cases for Bayesian modeling do exists, such as with an informative prior, it can have small sample advantage. It potentially can also have asymptotic advantage, but unfortunately such effects are ignored [I don't understand this; perhaps words are missing; should it "have been ignored" or "will be ignored", and should "only" be removed from the text] under the existing point of view, since the prior is treated as fixed thus vanished in the asymptotic results. We intend to give a different view on this case, and justify that an informative prior can have non-negligible asymptotic effects. Suppose there are a number of $k = k(m)$ studies of the same problem by different investigators in the past, resulted in k estimates of the same parameter θ, each study has a sample size of $m_j = m_j(m)$ $(j = 1, \ldots, k)$. These results can be summarized into a prior density q_m for the parameter, apparently this prior is very informative. Now we have a current dataset of size n, and want to perform a Bayesian analysis using the prior q_m, as in many cases we have no access to the previous original data. In classical Bayes theory q_m is treated as fixed, while asymptotic results with respect to n are used in the analysis, and all the information contained in the prior is vanished asymptotically. In practice, often the data size n may be in the hundreds or thousands and treated as 'valid' for the asymptotics, while the data size m in previous studies may also be in the hundreds or thousands but is treated as fixed, thus its effects is ignored by the asymptotic results. This is inappropriate for

q_m and the information contained in it, when m is relatively large. Here, different from traditional Bayesian philosophy, we treat such prior to be non-fixed. In particular, we give the data sizes in the previous studies for the prior the same status as that of the current data, viewing both sample sizes increase without bound in the asymptotic study. Thus the prior is asymptotically non-negligible. Intuitively, when q_m contains increasingly accurate information for θ as m tends to infinity at certain rates along with n, the asymptotic distribution of the posterior will concentrate on the true parameter value at a rate faster than that under the classical Bayes theory, and consequently, the inference is more efficient than that regarded under the classical Bayesian setting or the frequentist method based on the likelihood alone. In other words, the efficiency of Bayesian analysis with such informative prior is underevaluated by the classical Bayesian ideology. Here we give a new point of view for informative priors and attempt to recover their role in Bayesian asymptotics: they are not only useful for small sample size, they are asymptotically informative. We study Bayes estimator of parameters using such priors under convex losses and 0–1 loss.

In Section 2, we introduce the relevant notations and describe our point of view. The asymptotic results are studied in Section 3, and Section 4 illustrates its use with a simulation study and compare the results with those of the frequentist MLE. Relevant technical proofs are put in the Appendix.

2 The method

Let $X^n = (X_1, \ldots, X_n)$ be the observed data, our interest is the estimation of a parameter $\theta \in \Theta \subset R^d$, with a density function $f(\cdot|\theta)$ and we have a prior density $q_m(\cdot)$ for θ, with respect to some common dominating measure. In practice, $q_m(\cdot)$ is constructed using existing inference results based on datasets generated from the parameter(s), but not on the current observed data X^n which is also generated from the parameter(s). Often we have no access to the data sets in the previous studies but that of the prior $q_m(\cdot)$. Let $w(\cdot,\cdot)$ be the loss function. Denote $f(X^n|\theta) = \prod_{i=1}^n f(X_i|\theta)$, $h_m(\theta) = \log q_m(\theta)$, $l(\theta|X^n) = \sum_{i=1}^n \log f(X_i|\theta)$ as the log-likelihood, and $L(\theta|X^n) = l(\theta|X^n) + h_m(\theta)$ the *adjusted* log-likelihood using the informative prior $q_m(\cdot)$.

For $k = 0, 1, 2$, let $l^{(k)}(\theta|X^n)$ be the array of k-th partial derivatives of $l(\theta|X^n)$ with respect to θ, similarly for the notations $L^{(k)}(\theta|X^n)$, $h_m^{(k)}(\theta)$ and $f^{(k)}(x|\theta)$ $(k = 0, 1, 2)$. Let $I(\theta)$ be the Fisher information for θ under $f(\cdot|\theta)$, θ_0 be the true parameter generating the data, and denote $\xrightarrow{D}$ for convergence in distribution.

Definition. $q_m(\theta)$ is an *asymptotically informative prior* (AIP), if $h_m^{(2)}(\cdot)$ exists for all m, and as

$m \to \infty$,

$$\frac{1}{m}h_m^{(1)}(\theta_0) \xrightarrow{a.s.} 0, \qquad m^{-1/2}h_m^{(1)}(\theta_0) \xrightarrow{D} N(0, J(\theta_0)), \qquad \text{and} \qquad \frac{1}{m}h_m^{(2)}(\theta) \xrightarrow{a.s.} -J(\theta)$$

for some $d \times d$ matrix $J(\theta)$ which is non-negative definite and componentwise continuous on some compact set.

Remark 1. i) In the above we defined AIP in terms of $q_m(\cdot)$ and m. In some cases, the AIP constructed from existing independent parameter estimates by a general density estimator may not be explicitly associated with some integer m, as the one given in Section 4.1. In this case we can simply modify the above definition as: let $h(\theta) = \log q(\theta)$. $q(\cdot)$ is an AIP, if

$$\frac{1}{n}h^{(1)}(\theta_0) \xrightarrow{a.s.} 0, \qquad n^{-1/2}h^{(1)}(\theta_0) \xrightarrow{D} N(0, cJ(\theta_0)), \qquad \text{and} \qquad \frac{1}{n}h^{(2)}(\theta) \xrightarrow{a.s.} -cJ(\theta)$$

for some $0 \leq c < \infty$ and some $d \times d$ matrix $J(\theta)$ which is non-negative definite and componentwise continuous on some compact set. This second definition includes the first one by setting $c = \lim_n m/n$, and including any fixed prior by $c = 0$. But we are mainly interested in the case $c \neq 0$. We keep the first definition as it is more intuitive.

ii) In many cases, $q_m(\cdot)$ can be formulated as a multivariate exponential family: $q_m(\theta) = \exp\{m[\bar{\theta}_m' T(\theta) + B(\theta) + C(\bar{\theta}_m)]\}$ for some known differentiable functions $T(\cdot)$, $B(\cdot)$ and some known function $C(\cdot)$, where $\bar{\theta}_m$ is a consistent estimator of θ_0 constructed from past estimators and is asymptotically normal, i.e. $\sqrt{m}(\bar{\theta}_m - \theta_0) \xrightarrow{D} N(0, J^{-1}(\theta_0))$, with $T(\cdot)$ and $B(\cdot)$ satisfying $T^{(1)}(\theta_0) = J(\theta_0) + o(1)$ and $B^{(1)}(\theta_0) = -\theta_0' T^{(1)}(\theta_0) + o(1)$. Here $\bar{\theta}_m$ can be viewed as a hyperparameter.

For example, if $\bar{\theta}_m$ is a consistent and asymptotical normal estimator of θ_0 constructed from existing results, with asymptotical variance matrix $J^{-1}(\theta_0)$. Then $q_m(\theta) = (2\pi)^{-d/2} m^{d/2} |J(\bar{\theta}_m)|^{1/2}$ $\exp\{-\frac{m}{2}(\theta - \bar{\theta}_m)' J(\bar{\theta}_m)(\theta - \bar{\theta}_m)\}$ is an AIP and an exponential family with $T(\theta) = J(\bar{\theta}_m)\theta$, $B(\theta) = -\theta' J(\bar{\theta}_m)\theta/2$ and $C(\bar{\theta}_m) = -\bar{\theta}_m' J(\bar{\theta}_m)\bar{\theta}_m/2 + \frac{d}{2m}\log\frac{m}{2\pi} + \frac{1}{2m}\log|J(\bar{\theta}_m)|$.

As another example, we have independent estimates $\bar{\theta}_1, \ldots, \bar{\theta}_k$ of θ_0, with sample size $m_1, \ldots, m_k$ respectively, $q_m(\theta)$ be a twice differentiable density estimator based on the $\bar{\theta}_j$'s. When $\min\{m_j : j = 1, \ldots, k\}$ and k is large, each $\bar{\theta}_j = \theta_0 + o(1)$ and $Var(\bar{\theta}_j) = O(1/m_j)$, thus $q_m(\theta)$ will have a mode at $\theta_0 + o(1)$ and is an AIP for some m.

Here $q_m(\cdot)$ differs from the prior in the classical Bayesian setting in that it changes along with m, and the latter can be viewed as a special case of the former in which the rate is zero at which

$q_m(\cdot)$ concentrates toward $\theta_0 \in \Theta$. Bayesian inference is based on the posterior

$$q_m(\theta|X^n) = \frac{f(X^n|\theta)q_m(\theta)}{g_m(X^n)}, \quad \text{with} \quad g_m(X^n) = \int f(X^n|\theta)q_m(\theta)d\theta.$$

3 Main results

Below we study some basic asymptotic behavior of $q_m(\theta|X^n)$ and the corresponding Bayes estimators of θ under some commonly used losses. It is known that for fixed prior, the posterior will asymptotically concentrate on the true parameter (Strasser, 1981) generating the data and the scaled posterior will be asymptotically normal (LeCam, 1958; Walker, 1969). Intuitively, with the fixed prior replaced by the AIP, these properties will be kept but with faster rate. Denote $Q_m(\cdot|X^n)$ for the posterior distribution/measure of the density $q_m(\theta|X^n)$, and P_θ the probability measure corresponding to $f(\cdot|\theta)$. We list the following conditions, in which the first six are used in Strasser (1981).

(A1) The metric space (Θ, d) is separable, where $d(\theta, \eta) = ||P_\theta - P_\eta||$.

(A2) The functions $\{l(\theta|X^n)/n\}_{\theta\in\Theta}, n \in N$, are separable and measurable.

(A3) $f(\cdot|\theta)$, $\theta \in \Theta$, are lower semicontinuous, that is, $\limsup_{n\to\infty} f(\cdot|\theta_n) \le f(\cdot|\theta)$ (a.e.) if $d(\theta_n, \theta) \to 0$.

(A4) For every $\theta, \eta \in \Theta$, there is an open neighborhood $U_{\theta,\eta}$ of η such that $E_\theta(\inf_{\theta'\in U_{\theta,\eta}} l(\theta'|X^n))/n > -\infty$ for at least one n.

(A5) There is a prior distribution Q_0 (with density q_0) such that, for every $\theta \in \Theta$ and $\epsilon > 0$, $Q_0(\eta \in \Theta : E_\theta l(\eta|X) < E_\theta l(\theta|X) + \epsilon) > 0$.

(A6) $\forall\ \theta \in \Theta$, $\exists\ n_\theta$ such that $P_\theta^n(X^n : \int \prod_{i=1}^n f(X_i|\eta)Q_0(d\eta) < \infty) = 1$ if $n \ge n_\theta$.

(A7) $0 \le c = \lim_n m/n < \infty$.

(A8) $I(\cdot)$ is non-singular on some compact set and is continuous in a neighborhood of θ_0.

(A9) $\partial \int f(x|\theta)dx/\partial\theta = \int \partial f(x|\theta)/\partial\theta dx$.

(A10) $w(d, \theta) = w(\| d - \theta \|) \ge 0$, $w(0) = 0$, is non-decreasing in $\|\theta\|$ on Θ, is strictly increasing in $\|\theta\|$ in a neighborhood of 0, and is bounded on Θ.

Theorem 1. *Under(A1)-(A7), for any compact set $M \ni \theta_0$, we have*

$$P_{\theta_0}(\liminf_n Q_m(M|X^n) = 1) = 1.$$

Let $\tilde{q}_m(\cdot|X^n)$ be the posterior density of $\alpha = \sqrt{n}(\theta - \hat{\theta}_n)$ ($\hat{\theta}_n$ is the Bayes estimate of θ under the prior $q_m(\cdot)$ and 0–1 loss, as in Theorem 3), and $\phi(\cdot|B)$ be the density of $N(0, B)$.

Theorem 2. *Under (A1)-(A9), as $n \to \infty$, for finite $[a, b] \subset R^d$, we have*

$$\int_a^b \tilde{q}_m(\alpha|X^n)d\alpha \to \int_a^b \phi(\alpha|(I(\theta_0) + cJ(\theta_0))^{-1})d\alpha. \quad a.s.$$

Remark 2. When $m \to \infty$ but $c = \lim m/n = 0$ (for example $m = \log n$, $m = n^a$ with $0 < a < 1$, etc.), we can still use Theorem 2(ii) to approximate the asymptotic distribution of $\tilde{Q}_m$ as

$$\int_a^b \tilde{q}_m(\alpha|X^n)d\alpha \approx \int_a^b \phi(\alpha|(I(\theta_0) + \frac{m}{n}J(\theta_0))^{-1})d\alpha.$$

We refer to this as a small sample asymptotic result, which still gives better accuracy (smaller variance than the inverse Fisher information) than estimators not using AIP, and is very practical to use (as in practice, often n ranges from tens to thousands, but not infinity in the real sense). This remark also applies to Theorem 3(ii) and Theorem 5.

It is known that under the quadratic, absolute error and 0–1 losses, the Bayes estimator of θ is the posterior mean, median and mode of $q_m(\theta|X^n)$ respectively. Doob (1949) gave very simple conditions for the a.s. consistency of Bayes estimate under the quadratic loss. LeCam (1958) and Bickel and Yahav (1969) studied the consistency and asymptotic normality of Bayes estimates under general losses (not including the 0–1 loss). We will study these corresponding results with the asymptotically informative prior. We also study the case of 0–1 loss, due to its connection to the MLE, its simplicity for computation, and as a discrete loss function, it is not covered by the conditions for many other commonly used losses.

The 0–1 loss. Let, $w(\delta, \theta) = 0$ if $||\delta - \theta|| < \epsilon$ and $= 1$ otherwise, for some small $\epsilon > 0$. The Bayes estimate under this loss is or can be arbitrarily close to the posterior mode for small ϵ. In particular the posterior mode is the limit of the Bayes estimates for $\epsilon \to 0$, and hence we regard it a generalized Bayes estimate under the 0–1 loss, and will be of special interest. Alternatively, we can define the 0–1 loss as $w(\delta, \theta) = 0$ if $\delta = \theta$ and $= 1$ otherwise. Under this loss we define the Bayes estimate $\hat{\theta}_n$ of θ as the posterior mode

$$\hat{\theta}_n = \arg\sup_{\theta \in \Theta} q_m(\theta|X^n) = \arg\sup_{\theta \in \Theta}[f(X^n|\theta)q_m(\theta)] = \arg\sup_{\theta \in \Theta} L(\theta|X^n).$$

Theorem 3. *Assume (A7)-(A9), that there is a convex set A such that $\inf_{\theta\in A}|I(\theta)|>0$, $\theta_0\in A$ and $\hat{\theta}_n\in A$ for all large n. Then under the 0–1 loss, we have*

$$(i) \qquad \hat{\theta}_n \to \theta_0 \quad (a.s.)$$

$$(ii) \qquad \sqrt{n}(\hat{\theta}_n-\theta_0) \xrightarrow{D} N(0,[I(\theta_0)+cJ(\theta_0)]^{-1}).$$

Remark 3. The above asymptotical normality result includes the classical Bayes estimator in which $c=0$. When $c>0$, since $J(\cdot)$ is non-negative definite, $[I(\theta_0)+cJ(\theta_0)]^{-1}\leq I^{-1}(\theta_0)$ in the matrix non-negative definite sense, and $I^{-1}(\theta_0)$ is the asymptotical variance matrix for the classical Bayesian estimator and MLE based on the likelihood only. Thus Bayes estimate with asymptotically informative prior can be more efficient than that regarded by the classical Bayes point of view and than the MLE based only on $l(\cdot|X^n)$.

Let θ_n be the Bayes estimator of θ under loss $w(\cdot,\cdot)$ and the prior $q_m(\cdot)$.

Theorem 4. *Under (A1)-(A7) and (A10), we have*

$$\theta_n \to \theta_0, \qquad (a.s.).$$

In Theorem 3 (ii) we get asymptotic normality of the Bayes estimator with AIP under the 0–1 loss. Next we will have the result for Bayes estimators under the quadratic and absolute error losses, then we get the asymptotic normality results with AIP for the three most commonly used losses. Results with more general losses should be parallel, using the methods in Bickel and Yahav (1969) or in Gusev (1975) for example, we leave them here to avoid unnecessary technicalities.

Theorem 5. *Under (A1)-(A9), and assume $\inf_{\theta\in\Theta}|I(\theta)+cJ(\theta)|>0$, then with the quadratic or absolute error loss, we have*

$$\sqrt{n}(\theta_n-\theta_0) \xrightarrow{D} N(0,[I(\theta_0)+cJ(\theta_0)]^{-1}).$$

4 Numerical illustration

4.1 Preamble

In this section, we present the results of several Monte Carlo experiments to show the reduction in asymptotic variance of the Bayes estimate $\hat{\theta}_n$ of θ vis-à-vis the asymptotic variance of the classical

MLE. The general setup is as follows. Suppose that the data $X_1, \ldots, X_n$ are i.i.d. distributed from a d-dimensional continuous distribution $f(x|\theta)$ with unknown mean vector θ and known variance matrix. The frequentist MLE of θ is $\hat{\theta} = \arg\sup_{\theta\in\Theta}\{\sum_{i=1}^{n} \log f(x_i|\theta)\}$ with $I^{-1}(\hat{\theta})$ the asymptotic variance matrix of this estimator.

To represent a large number of independent studies of the same problem, assume that there are available k random draws $Y_1, \ldots, Y_{m_j}$ $(j = 1, \ldots, k)$ of sizes $m_1, \ldots, m_k$ from the same distribution as given above. Now, adopting the AIP $q_m(\cdot)$ requires k estimates $\tilde{\theta}_j$ of θ. Note, however, that in the construction of $q_m(\cdot)$ there is no need to know whether the $\tilde{\theta}_j$'s are Bayesian or frequentist MLEs, all that is needed is that they are independent consistent and asymptotical normal estimates of θ. Hence, to make things simpler, for each j the MLE of θ can be computed as follows $\tilde{\theta}_j = \arg\sup_{\theta\in\Theta}\{\sum_{i=1}^{m_j} \log f(y_{ij}|\theta)\}$.

To construct $q_m(\cdot)$ from the $\tilde{\theta}_j$'s adopt the following weighting method. Let $\bar{m} = \sum_{j=1}^{k} m_j/k$, $m = \sum_{j=1}^{k} m_j$, $\bar{\theta} = \sum_{j=1}^{k}(m_j/m)\tilde{\theta}_j$, assume $\tilde{\theta}_j$ has asymptotical variance matrix of $J_j^{-1}(\theta_0)$. Let $J^{-1}(\theta_0) = \sum_{j=1}^{k}(m_j/m)J_j^{-1}(\theta_0)$. Then for large m_j's and k vary along with n, it is reasonable to set $q_m(\cdot) = \phi(\bar{\theta},\ J^{-1}(\bar{\theta})/\bar{m})$. Since the 0–1 loss is used, the Bayes estimator of θ can be viewed as the MLE under the adjusted log-likelihood $L(\theta|X^n)$. In this way the computation is as simple as the MLE, the information in the AIP $q_m(\cdot)$ is used as the prior density, and the interpretation is intuitive. Especially, when $J_j^{-1}(\theta) = I^{-1}(\theta)$, we have $J^{-1}(\theta) = I^{-1}(\theta)$. Given $q_m(\cdot)$ above, $\hat{\theta}_n = \arg\sup_{\theta\in\Theta}\{\sum_{i=1}^{n} \log f(x_i|\theta) + \log q_m(\theta)\}$. The asymptotic variance matrix of this estimator is given in Theorem 3.

Remark 4. When $m_1 \approx \cdots \approx m_k$, $q_m(\cdot)$ can be constructed from the $\tilde{\theta}_j$'s in a more objective way by a commonly used density estimator, such as the kernel estimator, treating the $\tilde{\theta}_j$'s as approximately i.i.d. In this case, although we may not know m, by the relationship $I(\theta_0)+cJ(\theta_0) \approx I(\theta_0) + (m/n)(h^{(2)}(\theta_n)/m)$, the asymptotic variance matrix of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ in Theorems 3 and 5 can be consistently estimated by $(I(\hat{\theta}_n) + n^{-1}h^{(2)}(\hat{\theta}_n))^{-1}$.

4.2 Example

Let $y = (y_1, \ldots, y_n)'$ be an i.i.d. sample from a normal multiple linear regression model where the $d\times 1$ vector of regression coefficients, θ, is unknown and the variance matrix is known. Specifically, y satisfies $y = X\theta+\varepsilon$ where X is an $n\times d$ nonstochastic (design) matrix with rank d and with ith row denoted by $(x_{i1}, \ldots, x_{id})$. Furthermore, ε is an $n\times 1$ random error vector, and we assume that $\varepsilon|X \sim$

Table 1: Averaged (over 10000 replications) ML-estimates and, in parentheses, averaged Bayesian estimates with, in the last two bottom lines, values of the geometric means of the estimated variances corresponding to the above estimates; $n = 1000$.

θ	$d=1$, $\hat{\theta}$ $(\hat{\theta}_n)$	$d=3$, $\hat{\theta}$ $(\hat{\theta}_n)$	$d=5$, $\hat{\theta}$ $(\hat{\theta}_n)$	$d=7$, $\hat{\theta}$ $(\hat{\theta}_n)$	
θ_1	7.9990 (7.9990)	7.9998 (7.9997)	7.9994 (7.9993)	7.9981 (7.9980)	8.000
θ_2		2.4158 (2.4156)	2.4150 (2.4148)	2.4185 (2.4184)	2.415
θ_3		-3.8056 (-3.8053)	-3.8058 (-3.8055)	-3.8095 (-3.8092)	-3.806
θ_4			2.1184 (2.1183)	2.1238 (2.1237)	2.119
θ_5			-0.9961 (-0.9960)	-1.0054 (-1.0054)	-0.997
θ_6				0.5447 (0.5447)	0.545
θ_7				1.0543 (1.0542)	1.069
Geometric means of estimated variances of $\hat{\theta}$ $(\hat{\theta}_n)$					
	1.1147×10^{-2}	7.0728×10^{-3}	3.0094×10^{-3}	1.8698×10^{-3}	
	(1.000×10^{-3})	(1.5905×10^{-3})	(1.7480×10^{-3})	(1.8610×10^{-3})	

$N(0, \sigma_0^2 I_n)$ with σ_0^2 known. The frequentist MLE of θ and σ_0^2 are respectively $\hat{\theta} = (X'X)^{-1}X'y$ and $s^2 = n^{-1}(y - X\hat{\theta})'(y - X\hat{\theta})$. The inverse Fisher information matrix of $\hat{\theta}$ is given by $s^2(X'X)^{-1}$. Proceeding in the same fashion as before, we obtain k $m_j \times 1$ vectors of i.i.d. observations $z_1, \ldots, z_k$ randomly drawn from the multivariate normal distribution specified above. Then the j MLEs of θ and σ_0^2 are respectively $\tilde{\theta}_j = (X_j'X_j)^{-1}X_j'z_j$ and $s_j^2 = (m_j)^{-1}(z_j - X_j\tilde{\theta}_j)'(z_j - X_j\tilde{\theta}_j)$ where X_j is an $m_j \times d$ submatrix of X. Consequently, the AIP $q_m(\theta)$ for θ is a d-dimensional multivariate normal distribution with prior mean vector $\bar{\theta} = \sum_{j=1}^k (m_j/m)\tilde{\theta}_j$ and prior variance matrix $\hat{\Sigma} = \bar{m}^{-1}J^{-1}(\bar{\theta}) = \bar{m}^{-1}\sum_{j=1}^k (m_j/m)(s_j^2/m_j)I_d$. Then it is easy to verify that the posterior distribution $\prod_{i=1}^n f(y_i|\theta) \times q_m(\theta)$ of θ, up to a normalizing constant, is multivariate normally distributed with hyperparameters

$$E(\theta|y, X, \hat{\Sigma}) = \left(\hat{\Sigma}^{-1} + \sigma_0^2(X'X)\right)^{-1}\left(\hat{\Sigma}^{-1}\bar{\theta} + \sigma_0^2(X'X)\hat{\theta}\right),\ Var(\theta|y, X, \hat{\Sigma}) = \left(\hat{\Sigma}^{-1} + \sigma_0^2(X'X)\right)^{-1}.$$

Note that $E(\theta|y, X, \hat{\Sigma}) \equiv \hat{\theta}_n$. Hence, with $h^{(2)}(\theta) = \hat{\Sigma}^{-1}$, the asymptotic variance matrix of $\sqrt{n}(\hat{\theta}_n - \theta)$ can be readily estimated.

For the actual simulations we specialize the above regression model to the case

$$E(y_i|x_{i1}, \ldots, x_{iP}) = \theta_1 + 2\sum_{p=1}^{P}[\theta_{2p}\sin(2p\pi x_{ip}) + \theta_{2p+1}\cos(2p\pi x_{ip})], \quad (i = 1, \ldots, n),$$

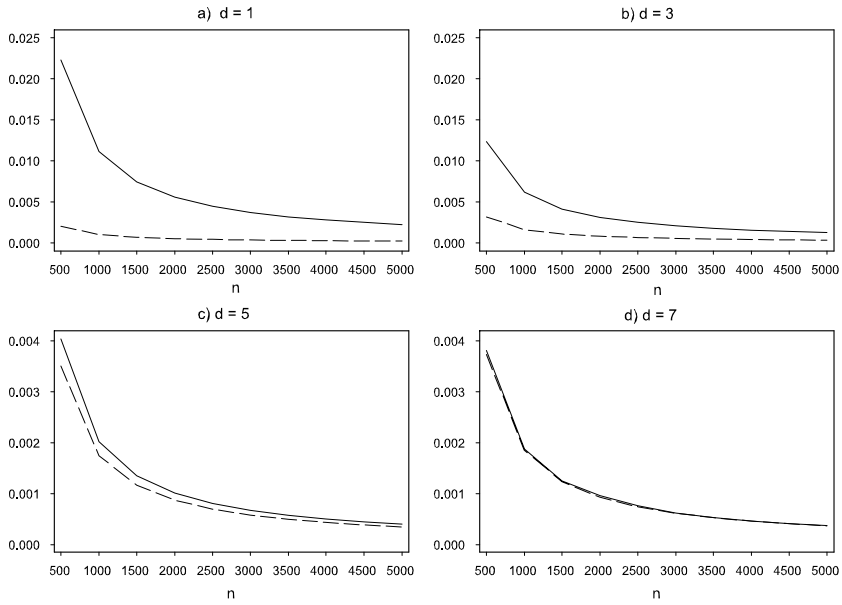

Figure 1. Values of the geometric means of the estimated variances for $\hat{\theta}$ (solid line) and $\hat{\theta}_n$ (medium dashed line), averaged over 10000 replications; $d = 1, 3, 5, 7$ and $n = 1000, 1500, \ldots, 5000$.

where P is a non-negative integer, and $\theta = (\theta_1, \ldots, \theta_d)'$ are real constants with $d = 2P + 1$. We consider the performance of the frequentist MLE and the Bayesian estimator of θ and their associated estimated asymptotic variances for models with successively increasing number of explanatory variables. The value of the true parameter θ_0 is shown in the rightmost column in Table 1. Throughout all simulations the x_{ip}'s are drawn from a uniform (0,1) distribution, with sample sizes $n = 1000, 1500, \ldots, 5000$. Moreover, in all experiments we set $\sigma_0^2 = 1$, $k = 101$, $d = 1, 3, 5, 7$, and $m_j = 100 + (j-1) \times w$ with $w = \lfloor (n - 100)/(k - 1) \rfloor$ and $j = 1, \ldots, k$.

Table 1 shows averaged (over 10000 replications) ML-estimates $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_d)'$ and, in parentheses, averaged Bayesian estimates $\hat{\theta}_n$ for $n = 1000$. The table also shows, on the last two bottom lines, values of the geometric means of the estimated variances corresponding to respectively $\hat{\theta}$ and $\hat{\theta}_n$. We see that both estimators perform very well. However, according to the geometric mean of the estimated variances, the Bayesian estimator $\hat{\theta}_n$ is notably more efficient than the MLE $\hat{\theta}$. This observation is typical for other sample sizes.

Figure 1 shows values of the geometric means of the estimated variances for sample sizes $n = 1000, 1500, \ldots, 5000$. Clearly, the same picture emerges as above. However, as expected, for increasing values of n the difference in efficiency between both estimators diminishes. In addition we see that, for fixed n, the MLE estimator $\hat{\theta}$ has nearly the same, but still higher, asymptotic

efficiency than the Bayesian estimator when d increases. These results are in agreement with theory, i.e. the asymptotic variance of the Bayesian estimator with AIP is always smaller than that of the MLE.

5 Appendix

Proof of Theorem 1. Let M^c be the complement of M, and $\hat{\theta}_m$ be the mode of $q_m(\cdot)$. Similarly as in the proof of Theorem 3, we have $\hat{\theta}_m \to \theta_0$ (a.s.) as $m \to \infty$. Thus for large m, $\hat{\theta}_m \in M$. Then

$$\begin{aligned} Q_m(M|X^n) &= \frac{\int_M f(X^n|\theta)q_m(\theta)d\theta}{\int_M f(X^n|\theta)q_m(\theta)d\theta + \int_{M^c} f(X^n|\theta)q_m(\theta)d\theta} \\ &= \frac{\int_M \frac{f(X^n|\theta)}{f(X^n|\theta_0)}\frac{q_m(\theta)}{q_m(\hat{\theta}_m)}d\theta}{\int_M \frac{f(X^n|\theta)}{f(X^n|\theta_0)}\frac{q_m(\theta)}{q_m(\hat{\theta}_m)}d\theta + \int_{M^c} \frac{f(X^n|\theta)}{f(X^n|\theta_0)}\frac{q_m(\theta)}{q_m(\hat{\theta}_m)}d\theta}. \end{aligned}$$

For any fixed prior q_0 with posterior distribution $Q_0(\cdot|X^n)$, by Theorem 2.5 in Strasser (1981)

$$\liminf_n Q_0(M|X^n) \to 1 \qquad \text{(a.s.)}.$$

This is equivalent to

$$\frac{\int_{M^c} \frac{f(X^n|\theta)}{f(X^n|\theta_0)}q_0(\theta)d\theta}{\int_M \frac{f(X^n|\theta)}{f(X^n|\theta_0)}q_0(\theta)d\theta} \to 0 \quad \text{(a.s.)}. \tag{A.1}$$

We now consider q_m with $m \to \infty$ as $n \to \infty$. Note $h_m^{(1)}(\hat{\theta}_m) = 0$, so by the definition of $h_m(\cdot)$,

$$\begin{aligned} &\frac{q_m(\theta)}{q_m(\hat{\theta}_m)} = \exp(h_m(\theta) - h_m(\hat{\theta}_m)) = \exp(-\frac{m}{2}(\theta - \hat{\theta}_m)'\frac{-h_m^{(2)}(\dot{\theta})}{m}(\theta - \hat{\theta}_m)) \\ &\approx \exp(-\frac{m}{2}(\theta - \hat{\theta}_m)'J(\dot{\theta})(\theta - \hat{\theta}_m)) \geq \exp(-\frac{m}{2}(\theta - \hat{\theta}_m)'J(\bar{\theta}_m)(\theta - \hat{\theta}_m)) \end{aligned}$$

where $\dot{\theta}$ is between θ and $\hat{\theta}_m$. We have, for large m,

$$\begin{aligned} \int_M \frac{f(X^n|\theta)}{f(X^n|\theta_0)}\frac{q_m(\theta)}{q_m(\hat{\theta}_m)}d\theta &\approx \int_M \frac{f(X^n|\theta)}{f(X^n|\theta_0)}\exp(-\frac{m}{2}(\theta - \hat{\theta}_m)'J(\dot{\theta})(\theta - \hat{\theta}_m))d\theta \\ &\geq \int_M \frac{f(X^n|\theta)}{f(X^n|\theta_0)}\exp(-\frac{m}{2}(\theta - \hat{\theta}_m)'J(\underline{\theta})(\theta - \hat{\theta}_m))d\theta \\ &= (\frac{2\pi}{m})^{d/2}|J(\underline{\theta})|^{-1/2}\int_M \frac{f(X^n|\theta)}{f(X^n|\theta_0)}\phi(\theta - \hat{\theta}_m|J^{-1}(\underline{\theta})/m)d\theta, \end{aligned}$$

where

$$\underline{\theta} = \underline{\theta}_m = \arg\inf_{\alpha\in\Theta}\int_M \frac{f(X^n|\theta)}{f(X^n|\theta_0)}\exp(-\frac{m}{2}(\theta - \hat{\theta}_m)'J(\alpha)(\theta - \hat{\theta}_m))d\theta.$$

Similarly,

$$\int_{M^c} \frac{f(X^n|\theta)}{f(X^n|\theta_0)}\frac{q_m(\theta)}{q_m(\hat{\theta}_m)}d\theta \leq (\frac{2\pi}{m})^{d/2}|J(\bar{\theta})|^{-1/2}\int_{M^c} \frac{f(X^n|\theta)}{f(X^n|\theta_0)}\phi(\theta - \hat{\theta}_m|J^{-1}(\bar{\theta})/m)d\theta,$$

where

$$\bar{\theta} = \bar{\theta}_m = \arg\sup_{\alpha\in\Theta} \int_M \frac{f(X^n|\theta)}{f(X^n|\theta_0)} \exp(-\frac{m}{2}(\theta - \hat{\theta}_m)' J(\alpha)(\theta - \hat{\theta}_m))d\theta.$$

Since $\hat{\theta}_m \to \theta_0$ (a.s.), both $f(X^n|\theta)/f(X^n|\theta_0)$ and $\phi(\theta - \hat{\theta}_m|J^{-1}(\bar{\theta})/m)$ become more and more concentrated around θ_0 as n and m increase. Both $\phi(\theta - \hat{\theta}_m|J^{-1}(\bar{\theta})/m)$ and $q_0(\theta)$ are density functions, but the former is more concordant with $f(X^n|\theta)/f(X^n|\theta_0)$ than the latter. Thus, for large n and m,

$$\int_M \frac{f(X^n|\theta)}{f(X^n|\theta_0)} \phi(\theta - \hat{\theta}_m|J^{-1}(\underline{\theta})/m)d\theta \geq \int_M \frac{f(X^n|\theta)}{f(X^n|\theta_0)} q_0(\theta)d\theta$$

and

$$\int_{M^c} \frac{f(X^n|\theta)}{f(X^n|\theta_0)} \phi(\theta - \hat{\theta}_m|J^{-1}(\bar{\theta})/m)d\theta \leq \int_{M^c} \frac{f(X^n|\theta)}{f(X^n|\theta_0)} q_0(\theta)d\theta.$$

These give

$$\begin{aligned}
\frac{\int_{M^c} \frac{f(X^n|\theta)}{f(X^n|\theta_0)} q_m(\theta)d\theta}{\int_M \frac{f(X^n|\theta)}{f(X^n|\theta_0)} q_m(\theta)d\theta} &= \frac{\int_{M^c} \frac{f(X^n|\theta)}{f(X^n|\theta_0)} \frac{q_m(\theta)}{q_m(\hat{\theta}_m)}d\theta}{\int_M \frac{f(X^n|\theta)}{f(X^n|\theta_0)} \frac{q_m(\theta)}{q_m(\hat{\theta}_m)}d\theta} \\
&\leq \frac{(\frac{2\pi}{m})^{d/2}|J(\bar{\theta})|^{-1/2} \int_{M^c} \frac{f(X^n|\theta)}{f(X^n|\theta_0)} \phi(\theta - \hat{\theta}_m|J^{-1}(\bar{\theta})/m)d\theta}{(\frac{2\pi}{m})^{d/2}|J(\underline{\theta})|^{-1/2} \int_M \frac{f(X^n|\theta)}{f(X^n|\theta_0)} \phi(\theta - \hat{\theta}_m|J^{-1}(\underline{\theta})/m)d\theta} \\
&\leq \frac{|J(\underline{\theta})|^{1/2}}{|J(\bar{\theta})|^{1/2}} \frac{\int_{M^c} \frac{f(X^n|\theta)}{f(X^n|\theta_0)} q_0(\theta)d\theta}{\int_M \frac{f(X^n|\theta)}{f(X^n|\theta_0)} q_0(\theta)d\theta} \to 0 \quad \text{(a.s.)}
\end{aligned}$$

by (A.1), and the above is equivalent to $\lim_n \inf Q_m(M|X^n) = 1$ (a.s.).

Proof of Theorem 2. Use the definition of $\hat{\theta}_n$ as given before in Theorem 3, we have

$$\begin{aligned}
Q_m([n^{-1/2}a + \hat{\theta}_n, n^{-1/2}b + \hat{\theta}_n]|X^n) &= \frac{\int_{n^{-1/2}a+\hat{\theta}_n}^{n^{-1/2}b+\hat{\theta}_n} f(X^n|\theta)q_m(\theta)d\theta}{\int_\Theta f(X^n|\theta)q_m(\theta)d\theta} \\
&= \frac{\int_{n^{-1/2}a+\hat{\theta}_n}^{n^{-1/2}b+\hat{\theta}_n} \exp\{L(\theta|X^n) - L(\hat{\theta}_n|X^n)\}d\theta}{\int_\Theta \exp\{L(\theta|X^n) - L(\hat{\theta}_n|X^n)\}d\theta} \\
&= \frac{\int_{n^{-1/2}a+\hat{\theta}_n}^{n^{-1/2}b+\hat{\theta}_n} \exp\{-\frac{n}{2}(\theta - \hat{\theta}_n)' \frac{-L^{(2)}(\dot{\theta}|X^n)}{n}(\theta - \hat{\theta}_n)\}d\theta}{\int_\Theta \exp\{-\frac{n}{2}(\theta - \hat{\theta}_n)' \frac{-L^{(2)}(\dot{\theta}|X^n)}{n}(\theta - \hat{\theta}_n)\}d\theta},
\end{aligned}$$

where $\dot{\theta}$ is between θ and $\hat{\theta}_n$. Since by Theorem 3 (i), $\hat{\theta}_n \to \theta_0$ (a.s.), and a and b are finite, for the $\dot{\theta}$ in the numerator, we have $\dot{\theta} \to \theta_0$ (a.s.). As $-l^{(2)}(\theta_0)/n \to I(\theta_0)$ (a.s.), and by definition of $h_m(\cdot)$ and condition (A7),

$$-\frac{L^{(2)}(\theta_0|X^n)}{n} = -\frac{1}{n} l^{(2)}(\theta_0|X^n) - \frac{m}{n}\frac{h_m^{(2)}(\theta_0)}{m} \to I(\theta_0) + cJ(\theta_0) \quad \text{(a.s.)}.$$

Also, $I(\cdot)$ and $J(\cdot)$ are continuous in a neighborhood of θ_0, we get

$$-\frac{L^{(2)}(\dot{\theta}|X^n)}{n} \to I(\theta_0) + cJ(\theta_0) \quad \text{(a.s.)}.$$

For the denominator, transform the integration with respect to $\alpha = \sqrt{n}(\theta - \hat{\theta}_n)$, we will see that the integration is finite as $|I(\dot{\theta}) + cJ(\dot{\theta})|$ is bounded away from zero. Thus for any $\epsilon > 0$, we can choose $[a', b'] \supset [a, b]$ such that (a.s.)

$$\frac{\int_{[n^{-1/2}a'+\hat{\theta}_n, n^{-1/2}b'+\hat{\theta}_n]^c} \exp\{-\frac{n}{2}(\theta-\hat{\theta}_n)'\frac{-L^{(2)}(\dot{\theta}|X^n)}{n}(\theta-\hat{\theta}_n)\}d\theta}{\int_{[n^{-1/2}a'+\hat{\theta}_n, n^{-1/2}b'+\hat{\theta}_n]} \exp\{-\frac{n}{2}(\theta-\hat{\theta}_n)'\frac{-L^{(2)}(\dot{\theta}|X^n)}{n}(\theta-\hat{\theta}_n)\}d\theta} \leq \epsilon,$$

and similarly for $\theta \in [n^{-1/2}a' + \hat{\theta}_n, n^{-1/2}b' + \hat{\theta}_n]$,

$$-\frac{L^{(2)}(\dot{\theta}|X^n)}{n} \to I(\theta_0) + cJ(\theta_0) \quad \text{(a.s.)}.$$

Thus, since $\epsilon > 0$ is arbitrary, for a' and b' large enough in norm, we have (a.s.),

$$\begin{aligned}
&\frac{\int_{n^{-1/2}a+\hat{\theta}_n}^{n^{-1/2}b+\hat{\theta}_n} \exp\{-\frac{n}{2}(\theta-\hat{\theta}_n)'\frac{-L^{(2)}(\dot{\theta}|X^n)}{n}(\theta-\hat{\theta}_n)\}d\theta}{\int_\Theta \exp\{-\frac{n}{2}(\theta-\hat{\theta}_n)'\frac{-L^{(2)}(\dot{\theta}|X^n)}{n}(\theta-\hat{\theta}_n)\}d\theta} \\
&= \frac{\int_{n^{-1/2}a+\hat{\theta}_n}^{n^{-1/2}b+\hat{\theta}_n} \exp\{-\frac{n}{2}(\theta-\hat{\theta}_n)'\frac{-L^{(2)}(\dot{\theta}|X^n)}{n}(\theta-\hat{\theta}_n)\}d\theta}{\int_{n^{-1/2}a'+\hat{\theta}_n}^{n^{-1/2}b'+\hat{\theta}_n} \exp\{-\frac{n}{2}(\theta-\hat{\theta}_n)'\frac{-L^{(2)}(\dot{\theta}|X^n)}{n}(\theta-\hat{\theta}_n)\}d\theta} + o(1) \\
&= \frac{\int_{n^{-1/2}a+\hat{\theta}_n}^{n^{-1/2}b+\hat{\theta}_n} \exp\{-\frac{n}{2}(\theta-\hat{\theta}_n)'(I(\theta_0)+cJ(\theta_0))(\theta-\hat{\theta}_n)\}d\theta}{\int_{n^{-1/2}a'+\hat{\theta}_n}^{n^{-1/2}b'+\hat{\theta}_n} \exp\{-\frac{n}{2}(\theta-\hat{\theta}_n)'(I(\theta_0)+cJ(\theta_0))(\theta-\hat{\theta}_n)\}d\theta} + o(1) \\
&= \frac{\int_{n^{-1/2}a+\hat{\theta}_n}^{n^{-1/2}b+\hat{\theta}_n} \exp\{-\frac{n}{2}(\theta-\hat{\theta}_n)'(I(\theta_0)+cJ(\theta_0))(\theta-\hat{\theta}_n)\}d\theta}{\int_\Theta \exp\{-\frac{n}{2}(\theta-\hat{\theta}_n)'(I(\theta_0)+cJ(\theta_0))(\theta-\hat{\theta}_n)\}d\theta} + o(1).
\end{aligned}$$

From the above, set $\alpha = \sqrt{n}(\theta - \hat{\theta}_n)$, we get (a.s.)

$$\tilde{Q}_m([a,b]|X^n) = \frac{\int_a^b \exp\{-\frac{1}{2}\alpha'(I(\theta_0)+cJ(\theta_0))\alpha\}d\alpha}{\int_\Theta \exp\{-\frac{1}{2}\alpha'(I(\theta_0)+cJ(\theta_0))\alpha\}d\alpha} + o(1)$$

$$\to \frac{\frac{|I(\theta_0)+cJ(\theta_0)^{1/2}|}{(2\pi)^{d/2}}\int_a^b \exp\{-\frac{1}{2}\alpha'(I(\theta_0)+cJ(\theta_0))\alpha\}d\alpha}{\frac{|I(\theta_0)+cJ(\theta_0)^{1/2}|}{(2\pi)^{d/2}}\int_\Theta \exp\{-\frac{1}{2}\alpha'(I(\theta_0)+cJ(\theta_0))\alpha\}d\alpha} = \int_a^b \phi(\theta|I(\theta_0)+cJ(\theta_0))d\theta.$$

Proof of Theorem 3. (i) By definition of $\hat{\theta}_n$, $L^{(1)}(\hat{\theta}_n|X^n) = 0$, so we have

$$-L^{(1)}(\theta_0|X^n) = L^{(1)}(\hat{\theta}_n|X^n) - L^{(1)}(\theta_0|X^n) = L^{(2)}(\dot{\theta}|X^n)(\hat{\theta}_n - \theta_0),$$

where $\dot{\theta}$ is between $\hat{\theta}_n$ and θ_0. By the given condition, $\dot{\theta} \in A$, and so for large n and m, $-\frac{1}{n}l^{(2)}(\dot{\theta}|X^n) \approx I(\dot{\theta})$ (a.s.), which is non-singular by assumption, and $-\frac{1}{m}h_m^{(2)}(\dot{\theta}) \approx J(\dot{\theta})$. Thus

$$-\frac{1}{n}L^{(2)}(\dot{\theta}|X^n) = -\frac{1}{n}l^{(2)}(\dot{\theta}|X^n) - \frac{m}{n}\frac{1}{m}h_m^{(2)}(\dot{\theta}) = I(\dot{\theta}) + cJ(\dot{\theta}) + o(1), \qquad \text{(a.s.)}$$

i.e., $-\frac{1}{n}L^{(2)}(\dot{\theta}|X^n) \approx I(\dot{\theta}) + cJ(\dot{\theta})$ is non-singular (a.s.) for all large n. Thus, for large n, a.s.

$$\hat{\theta}_n - \theta_0 = (-\frac{1}{n}L^{(2)}(\dot{\theta}|X^n))^{-1}(\frac{1}{n}L^{(1)}(\theta_0|X^n)).$$

Similarly,

$$\frac{1}{n}L^{(1)}(\theta_0|X^n) = \frac{1}{n}l^{(1)}(\theta_0|X^n) + \frac{m}{n}\frac{1}{m}h_m^{(1)}(\theta_0) = \frac{1}{n}l^{(1)}(\theta_0|X^n) + o(1).$$

Since

$$\frac{1}{n}l^{(1)}(\theta_0|X^n) \stackrel{a.s.}{\to} E\left(\frac{f^{(1)}(X|\theta_0)}{f(X|\theta_0)}\right) = \frac{\partial}{\partial\theta_0}\int f(x|\theta_0)dx = 0,$$

we have

$$\hat{\theta}_n - \theta_0 = (-\frac{1}{n}L^{(2)}(\dot{\theta}|X^n))^{-1}(\frac{1}{n}l^{(1)}(\theta_0|X^n) + o(1)) \to 0, \qquad \text{(a.s.)}$$

or $\hat{\theta}_n \to \theta_0$ (a.s.).

(ii). By (i), we get $\dot{\theta} \to \theta_0$ (a.s.). Since $I(\cdot)$ is continuous at θ_0, so $-\frac{1}{n}L^{(2)}(\dot{\theta}|X^n) = -\frac{1}{n}l^{(2)}(\dot{\theta}|X^n) - \frac{m}{n}\frac{1}{m}h_m^{(2)}(\dot{\theta}) \to I(\theta_0) + cJ(\theta_0)$ (a.s.). Also, since $n^{-1/2}l^{(1)}(\theta_0|X^n) \stackrel{D}{\to} N(0, I(\theta_0))$, $m^{-1/2}h_m^{(1)}(\theta_0) \stackrel{D}{\to} N(0, J(\theta_0))$, and $l^{(1)}(\theta_0|X^n)$ and $h_m^{(1)}(\theta_0)$ are independent, so $n^{-1/2}l^{(1)}(\theta_0|X^n) + \sqrt{\frac{m}{n}}m^{-1/2}h_m^{(1)}(\theta_0) \stackrel{D}{\to} N(0, I(\theta_0) + cJ(\theta_0))$, and we get

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = (-n^{-1}L^{(2)}(\dot{\theta}|X^n))^{-1}[n^{-1/2}L^{(1)}(\theta_0|X^n)]$$
$$= (-n^{-1}L^{(2)}(\dot{\theta}|X^n))^{-1}\left(n^{-1/2}l^{(1)}(\theta_0|X^n) + \sqrt{\frac{m}{n}}m^{-1/2}h_m^{(1)}(\theta_0)\right) \stackrel{D}{\to} N(0, [I(\theta_0) + cJ(\theta_0)]^{-1})$$

by Slutzky's theorem.

Proof of Theorem 4. By (A10), for $\epsilon > 0$, there is a $\delta > 0$ such that $w(\theta, \theta_0) \le \epsilon/2$ as long as $\| \theta - \theta_0 \| \le \delta$. Let $M \subset R^d$ be the closed ball with center θ_0 and radius δ. Since θ_n is the Bayes estimator, i.e. $\theta_n = \arg\min_d \int_\Theta w(d, \theta)\pi_m(\theta|X^n)d\theta$, so the Bayes risk

$$\begin{aligned} R_n &= \int_\Theta w(\theta_n, \theta)\pi_m(\theta|X^n)d\theta \le \int_\Theta w(\theta_0, \theta)\pi_m(\theta|X^n)d\theta \\ &= \int_M w(\theta_0, \theta)\pi_m(\theta|X^n)d\theta + \int_{M^c} w(\theta_0, \theta)\pi_m(\theta|X^n)d\theta \le \epsilon/2 + \int_{M^c} w(\theta_0, \theta)\pi_m(\theta|X^n)d\theta. \end{aligned}$$

By (A10), there is a constant $0 < C < \infty$ such that $\sup_{\theta \in M^c} w(\theta_0, \theta) \le C$. Also, by Theorem 1, $\pi_m(M^c|X^n) \to 0$ (a.s.), thus for large n we have $\pi_m(M^c|X^n) \le \epsilon/(2C)$ (a.s.), and

$$\limsup_n \int_{M^c} w(\theta_0, \theta)\pi_m(\theta|X^n)d\theta \le C\pi_m(M^c|X^n) \le \epsilon/2. \qquad \text{(a.s.)}.$$

Since $\epsilon > 0$ is arbitrary, we get $R_n \to 0$ (a.s.).

On the other hand, since $w(\theta) = 0$ and is strictly increasing in $\|\theta\|$ and non-decreasing, if $\theta_n \nrightarrow \theta_0$ (a.s.), then there is a sub-sequence $\theta_{n_k} \stackrel{a.s.}{\to} \theta_1 \neq \theta_0$ ($n_k \to \infty$) and a compact $M \ni \theta_0$, such that

$$\liminf_{n_k} \inf_{\theta \in M} w(\theta_{n_k}, \theta) \geq w(\theta_1, \theta_0)/2 > 0. \quad \text{(a.s.)}.$$

The above argument can be easily understood by drawing a picture (assume θ be 1-dimensional for simplicity). Also, by Theorem 1, $\pi_m(M|X^{n_k}) \to 1$ (a.s.), thus

$$\begin{aligned} \limsup_n R_n &\geq \liminf_{n_k} \int_M w(\theta_{n_k}, \theta)\pi_m(\theta|X^{n_k})d\theta \\ &\geq \liminf_{n_k} \inf_{\theta \in M} w(\theta_{n_k}, \theta)\pi_m(M|X^{n_k}) \geq w(\theta_1, \theta_0)/2 > 0, \quad \text{(a.s.)} \end{aligned}$$

a contradiction. So we must have $\theta_n \to \theta_0$ (a.s.).

Proof of Theorem 5. In view of Theorem 3 (ii), we only need to show, in componentwise sense,

$$\sqrt{n}(\theta_n - \hat{\theta}_n) = o_P(1).$$

We first consider the quadratic loss. Then $w(\delta, \theta) = \sum_{j=1}^d c_j(\delta_j - \theta_j)^2$ for some $0 \leq c_j < \infty$. Let $w^{(1)}(\delta, \theta) = (\partial w(\delta, \theta)/\partial\delta_1, \ldots, \partial w(\delta, \theta)/\partial\delta_d)' = 2\sum_{j=1}^d c_j(\delta_j - \theta_j)$. We only prove the result for the first component, and without loss of generality we assume θ is 1-dimensional and $c_1 = 1$. By definition of θ_n, we have

$$\begin{aligned} 0 &= \int w^{(1)}(\theta_n - \theta)\exp\{L(\theta|X^n)\}d\theta = \int w^{(1)}(\theta_n - \theta)\exp\{L(\theta|X^n) - L(\hat{\theta}_n|X^n)\}d\theta \\ &= \int w^{(1)}(\theta_n - \theta)\exp\{-\frac{n}{2}(\theta - \hat{\theta}_n)'\frac{-L^{(2)}(\dot{\theta})}{n}(\theta - \hat{\theta}_n)\}d\theta. \end{aligned}$$

Let $\alpha = \sqrt{n}(\theta - \hat{\theta}_n)$, and note $w^{(1)}(\theta_n - \theta) = 2(\theta_n - \theta)$, the above is

$$0 = \int [\sqrt{n}(\theta_n - \hat{\theta}_n) - \alpha]\exp\{-\frac{1}{2}\alpha'\frac{-L^{(2)}(\dot{\theta})}{n}\alpha\}d\alpha$$

or

$$\sqrt{n}(\theta_n - \hat{\theta}_n)\int \exp\{-\frac{1}{2}\alpha'\frac{-L^{(2)}(\dot{\theta})}{n}\alpha\}d\alpha = \int \alpha\exp\{-\frac{1}{2}\alpha'\frac{-L^{(2)}(\dot{\theta})}{n}\alpha\}d\alpha.$$

As in the proof of Theorem 3, for each fixed $\dot{\theta}$, $-L^{(2)}(\dot{\theta})/n \to I(\dot{\theta}) + cJ(\dot{\theta})$ (a.s.), and $I(\cdot) + cJ(\cdot) > 0$ by assumption. So there are $\{\eta_n\}$ and $\{\zeta_n\}$ independent of α such that

$$\liminf_n \int \exp\{-\frac{1}{2}\alpha'\frac{-L^{(2)}(\dot{\theta})}{n}\alpha\}d\alpha \geq \liminf_n \int \exp\{-\frac{1}{2}\alpha'(I(\eta_n) + cJ(\eta_n))\alpha\}d\alpha > 0$$

and

$$\int \alpha \exp\{-\frac{1}{2}\alpha' \frac{-L^{(2)}(\dot{\theta})}{n}\alpha\}d\alpha = \int \alpha \exp\{-\frac{1}{2}\alpha'(I(\zeta_n) + cJ(\zeta_n))\alpha\}d\alpha + o(1) = o(1), \quad \text{(a.s.)}.$$

This gives $\sqrt{n}(\theta_n - \hat{\theta}_n) = o_p(1)$.

For the absolute error loss, $\hat{\theta}_n$ is the posterior median, so we have

$$\int_{\theta<\hat{\theta}_n} \exp\{L(\theta|X^n)\}d\theta = \int_{\theta\geq\hat{\theta}_n} \exp\{L(\theta|X^n)\}d\theta,$$

or, similarly as before,

$$\int_{\theta<\hat{\theta}_n} \exp\{-\frac{n}{2}(\theta-\hat{\theta}_n)'\frac{-L^{(2)}(\dot{\theta})}{n}(\theta-\hat{\theta}_n)\}d\theta = \int_{\theta\geq\hat{\theta}_n} \exp\{-\frac{n}{2}(\theta-\hat{\theta}_n)'\frac{-L^{(2)}(\dot{\theta})}{n}(\theta-\hat{\theta}_n)\}d\theta,$$

or

$$\int_{\alpha<\sqrt{n}(\theta_n-\hat{\theta}_n)} \exp\{-\frac{1}{2}\alpha'\frac{-L^{(2)}(\dot{\theta})}{n}\alpha\}d\alpha = \int_{\alpha\geq\sqrt{n}(\theta_n-\hat{\theta}_n)} \exp\{-\frac{1}{2}\alpha'\frac{-L^{(2)}(\dot{\theta})}{n}\alpha\}d\alpha. \qquad (A.2)$$

Let $\Phi(\cdot)$ be the distribution function of the standard normal. For $\epsilon > 0$, there is $0 < M < \infty$ such that

$$\int_{\alpha<-M} \exp\{-\frac{1}{2}\alpha'\frac{-L^{(2)}(\dot{\theta})}{n}\alpha\}d\alpha < \epsilon/4, \qquad \int_{\alpha>M} \exp\{-\frac{1}{2}\alpha'\frac{-L^{(2)}(\dot{\theta})}{n}\alpha\}d\alpha < \epsilon/4,$$

and $\Phi(-M) < \epsilon/4$. Also, for $\alpha \in [-M, \sqrt{n}(\theta_n - \hat{\theta}_n))$, or $\theta \in [-n^{-1/2}M + \hat{\theta}_n, \theta_n)$, we have $\dot{\theta} \in [-n^{-1/2}M + \hat{\theta}_n, \theta_n)$. Since $\hat{\theta}_n \to \theta_0$ (a.s.) by Theorem 3(i) and $\theta_n \to \theta_0$ (a.s.) by Theorem 4, we get $\dot{\theta} \to \theta_0$ (a.s.). This gives, by dominated convergence,

$$\int_{[-M,\sqrt{n}(\theta_n-\hat{\theta}_n))} \exp\{-\frac{1}{2}\alpha'\frac{-L^{(2)}(\dot{\theta})}{n}\alpha\}d\alpha$$

$$= \int_{[-M,\sqrt{n}(\theta_n-\hat{\theta}_n))} \exp\{-\frac{1}{2}\alpha'(I(\theta_0) + cJ(\theta_0))\alpha\}d\alpha + o_P(1) = \Phi(\sqrt{n}(\theta_n - \hat{\theta}_n)) - \Phi(-M) + o_P(1).$$

Similarly,

$$\int_{[\sqrt{n}(\theta_n-\hat{\theta}_n),M)} \exp\{-\frac{1}{2}\alpha'\frac{-L^{(2)}(\dot{\theta})}{n}\alpha\}d\alpha$$

$$= \int_{[\sqrt{n}(\theta_n-\hat{\theta}_n),M)} \exp\{-\frac{1}{2}\alpha'(I(\theta_0) + cJ(\theta_0))\alpha\}d\alpha + o_P(1) = \Phi(M) - \Phi(\sqrt{n}(\theta_n - \hat{\theta}_n)) + o_P(1).$$

Now from (A.2) and the above results for large n we have, since $\Phi(M) \geq 1 - \epsilon$,

$$|1 - 2\Phi(\sqrt{n}(\theta_n - \hat{\theta}_n))| \leq \epsilon.$$

Since $\epsilon > 0$ is arbitrary, the above is possible only if $\Phi(\sqrt{n}(\theta_n - \hat{\theta}_n)) \to 1/2$ or $\sqrt{n}(\theta_n - \hat{\theta}_n) = o_p(1)$.

Acknowledgements. This work is supported in part for Yuan by the National Center for Research Resources at NIH grant 2G12RR003048.

References

Bickel, P.J. and Yahav, J.A. (1969). Some contributions to the asymptotic theory of Bayes solutions. *Z. Wahrscheinlichkeitstheorie view. Geb.* (Current English title: *Probab. Theory Related Fields*) **11**, 257–276.

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

Doob, J.L. (1949). Application of the theory of martingales. *Colloque International Centre Nat. Rech. Sci, Paris*, 22–28.

Efron, B. (2005). Bayesians, frequentists, and the scientists. *J.Am. Statist. Ass.*, **100**, 1–5.

Jeffreys, H. (1961). *Theory of Probability*, 3rd edition, Oxford: Clarendon Press.

Gusev, S.I. (1975). Asymptotic expansions associated with some statistical estimators in the smooth case I. Expansions of random variables. *Theory Probab. Appl.*, **20**, 470–498.

LeCam, L. (1958). Les propriétés asymptotiques des solutions de Bayes. *Publ. Inst. Statist. Univ. Paris*, **7**, 18–35.

Mukerjee, R. and Ghosh, M. (1997). Second-order probability matching priors. *Biometrika*, **84**, 970–975.

Strasser, H. (1981). Consistency of maximum likelihood and Bayes estimates. *Ann. Statist.*, **9**, 1107–1113.

Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York.

Walker, A.M. (1969). On the asymptotic behavior of posterior distributions. *J.R. Statist. Soc.* B, **31**, 80–88.

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