IMS Collections
Borrowing Strength: Theory Powering Applications –
A Festschrift for Lawrence D. Brown
Vol. 6 (2010) 87–98
© Institute of Mathematical Statistics, 2010
DOI: 10.1214/10-IMSCOLL607

High dimensional Bernstein-von Mises: simple examples

Iain M. Johnstone[1,*]

Stanford University

Abstract: In Gaussian sequence models with Gaussian priors, we develop some simple examples to illustrate three perspectives on matching of posterior and frequentist probabilities when the dimension p increases with sample size n: (i) convergence of joint posterior distributions, (ii) behavior of a non-linear functional: squared error loss, and (iii) estimation of linear functionals. The three settings are progressively less demanding in terms of conditions needed for validity of the Bernstein-von Mises theorem.

The Bernstein-von Mises theorem is a formalization of conditions under which Bayesian posterior credible intervals agree approximately with frequentist confidence intervals constructed from likelihood theory. It is traditionally formulated in situations in which the number of parameters p is fixed and the sample size $n \to \infty$. The situation is very different in high dimensional settings in which p is allowed to grow with n. In this primarily expository paper, we use simple Gaussian sequence models to draw some conclusions about when a version of Bernstein-von Mises can hold.

We begin with a somewhat informal statement of the classical theorem. Suppose that $Y_1, \ldots, Y_n$ are i.i.d. observations from a distribution P_θ having density $p_\theta(y)\, d\mu(y)$ where $\theta \in \Theta \subset \mathbb{R}^p$. The log-likelihood for a single observation

$$\ell_\theta = \log p_\theta(y),$$

and, as usual, the score function vector and Fisher information matrix are given by

$$\dot{\ell}_\theta = (\partial/\partial\theta) \log p_\theta(y); \qquad I_\theta = E_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T.$$

Writing $Y^n = (Y_1, \ldots, Y_n)$ for the full data, the log-likelihood

$$L_n(\theta) = \sum_{k=1}^n \ell_\theta(Y_k),$$

and we write $\hat{\theta}_{MLE}$ for a maximizer of $L_n(\theta)$. Classical likelihood theory says that any (nice) estimator satisfies the information bound

$$\mathrm{Var}_\theta\, \hat{\theta} \geq n^{-1} I_\theta^{-1},$$

*Bala Rajaratnam and a referee offered valuable comments on an earlier draft. This work was supported in part by NIH grant RO1 EB 001988. and NSF DMS 0906812.
[1]Department of Statistics, Stanford University, CA 94305
Keywords and phrases: high dimensional inference, Gaussian sequence, linear functional, squared error loss, posterior distribution, frequentist.
AMS 2000 subject classifications: Primary 62E20; secondary 62F15.

in the usual ordering of nonnegative definite matrices, and that the bound is asymptotically attained by the MLE, which is also asymptotically Gaussian:

$$\hat{\theta}_{MLE}|\theta \sim N_p(\theta, n^{-1}I_\theta^{-1}).$$

Now suppose that $\pi(\theta)$ is the density of a prior distribution with respect to Legesgue measure. Then the posterior distribution of θ given Y^n is given by Bayes' rule; we denote it simply by $P_{\theta|Y^n}$.

The Bernstein-von Mises theorem says, informally, that this posterior distribution is, in large samples, approximately normal with mean approximately the MLE, $\hat{\theta}_{MLE}$ and variance matrix approximately $n^{-1}I_{\theta_0}^{-1}$ (here θ_0 is the 'true' value of θ generating the observations $Y_1, \ldots, Y_n$). Using the scalar case for simplicity, and writing $\sigma_n^2 = n^{-1}I_{\theta_0}^{-1}$ and $z_\alpha = \tilde{\Phi}^{-1}(\alpha)$, we have that an approximate $100(1-\alpha)\%$ credible interval for θ would be given by $\hat{\theta}_{MLE} \pm z_{\alpha/2}\hat{\sigma}_n$. This is exactly the same as the frequentist confidence interval based on asymptotic normality of the MLE. Thus in large samples the effect of the prior density π disappears: "the data overwhelms the prior".

A somewhat more formal statement uses the notion of variation distance between probability measures P and Q, and an equivalent expression in terms of the densities $p = dP/d\mu$ and $q = dQ/d\mu$ relative to a dominating measure μ:

$$\|P - Q\| = \max_A |P(A) - Q(A)| = \tfrac{1}{2}\int |p - q|\, d\mu.$$

Suppose that $\pi(\theta)$ is continuous and positive at the 'true' value θ_0, and that $\theta \to P_\theta$ is differentiable in quadratic mean and satisfies a further mild separation condition, then

$$\|P_{\theta|Y^n} - N(\hat{\theta}_{MLE}, n^{-1}I_{\theta_0}^{-1})\| \to 0. \tag{1}$$

in probability under $P_{\theta_0}^n$.

In other words, the variation distance between posterior and the approximating Gaussian distribution is a random variable depending on Y^n, and which converges to zero in probability under repeated draws from P_{θ_0}.

A development of the Bernstein-von Mises theorem as formulated above may be found in [21, §10.2]. A proof due to Bickel is given in [17, §6.8]. Extension from independent to dependent sampling settings are possible, see e.g. [1, 13]. For further references and methods of proof of the classical results, see [12, §1.4 and §1.5].

1. Growing Gaussian location model

In nonparametric and semiparametric settings the situation is very different. Even frequentist consistency of nonparametric Bayesian methods is a difficult issue with a large literature of both positive and negative results (e.g. [12, 11]). One cannot therefore expect Bernstein-von Mises phenomena in any great generality for the full posterior.

In this largely expository paper, we do some simple calculations in symmetric Gaussian sequence models. The Gaussian sequence structure makes possible an elementary set of examples that avoid the technical challenges posed by, and sophistication needed for, posterior Gaussian approximation in high dimensional settings (see references in Section 5). Nevertheless, the Gaussian examples can conveniently illustrate some of the issues related to validity of the Bernstein-von Mises theorem

in high dimensional models. Depending on the frequentist or Bayesian perspective, we assume that $p = p(n)$ grows with n, and one, or both, of

$$\begin{aligned} &(\mathbf{D}) \quad \text{Data:} \quad \bar{Y}|\theta \sim N_p(\theta, \sigma_n^2 I), \qquad \text{and} \\ &(\mathbf{P}) \quad \text{Prior:} \qquad \theta \sim N_p(0, \tau_n^2 I). \end{aligned}$$

The notation $\bar{Y}$ suggests an average $(Y_1 + \cdots + Y_n)/n$ of observations individually of variance σ_0^2, so that in this case $\sigma_n^2 = \sigma_0^2/n$. [If p were held fixed, not depending on n, then σ_n^2 would match with the definition given in the introductory section.] We also allow the prior variance τ_n^2 to depend on the sample size n.

Our goal is to compare the Bayesian posterior distribution $\mathcal{L}(\theta|Y)$ with frequentist distributions, in particular those of the MLE $\mathcal{L}(\hat{\theta}_{MLE}|\theta)$ and of the posterior mean Bayes estimator $\mathcal{L}(\hat{\theta}_B|\theta)$. A key simplification is that since both prior and likelihood are Gaussian, so also is the posterior distribution, and hence all the behavior will be determined by centering and scaling. Thus from standard results, the posterior is given by

$$\begin{aligned} \theta|\bar{Y} = \bar{y} &\sim N_p(w_n y, w_n \sigma_n^2 I), \\ w_n &= \tau_n^2/(\sigma_n^2 + \tau_n^2). \end{aligned} \tag{2}$$

Remarks. 1. The reference to Gaussian sequence models becomes clearer if, as will be helpful later, we write out assumptions (D) and (P) in co-ordinates:

$$\begin{aligned} &(\mathbf{D}_{\text{seq}}) \quad \text{Data:} \quad \bar{y}_k = \theta_k + \sigma_n \epsilon_k, \qquad \text{and} \\ &(\mathbf{P}_{\text{seq}}) \quad \text{Prior:} \quad \theta_k = \tau_n \zeta_k, \end{aligned}$$

with ϵ_k and ζ_k all i.i.d standard Gaussian, for $k = 1, \ldots, p(n)$.

Strictly speaking, the indexing by n of parameters σ_n, τ_n and $p(n)$ creates a sequence of sequence models. However, one can, as needed for almost sure results, think of the infinite sequences $\{(\epsilon_k, \zeta_k), k \in \mathbb{N}\}$ as being drawn from a single common probability space.

2. We also consider the infinite sequence Gaussian white noise model

$$Y_t = \int_0^t f(s)\, ds + \sigma_n W_t, \qquad 0 \le t \le 1 \tag{3}$$

or equivalently, when expressed in any orthonormal basis $\{\varphi_k(t)\}$ for $L_2[0,1]$,

$$y_k = \theta_k + \sigma_n \epsilon_k, \qquad \epsilon_k \overset{\text{ind}}{\sim} N(0,1), \tag{4}$$

where it is assumed that $\sigma_n = \sigma_0/\sqrt{n}$. For some examples, it is helpful to use doubly indexed orthonormal bases $\{\varphi_{jk}(t), k = 1, \ldots, 2^j, j \in \mathbb{N}\}$ such as arise with systems of orthonormal wavelets.

The forthcoming book [14] will have more on estimation in such Gaussian sequence models.

We develop three perspectives on the Bernstein-von Mises phenomenon:

(1) global convergence of the posterior,
(2) behavior of a non-linear functional $\|\theta - \hat{\theta}\|^2$, and of
(3) *linear* functionals Lf, in the white noise model (3)–(4).

We shall see that these situations are progressively "less demanding" in terms of validity of the Bernstein-von Mises phenomenon. Indeed, case (1) requires that $w_n \to 1$ at a sufficiently fast rate, while setting (2) needs only $w_n \to 1$. In case (3), the formulation itself delivers $w_n \to 1$, and covers at least all bounded linear functionals.

2. Global convergence of posterior

The first calculation considers the $p-$dimensional posterior distribution (2) and shows that the convergence in (1) occurs, even in the best possible case that $\theta_0 = 0$, only if the shrinkage factor w_n approaches 1 at a sufficiently fast *rate*.

Proposition 1. *Let $\theta_0 = 0$. The variation distance between posterior distribution $P_{\theta|Y^n}$ and $N(\hat{\theta}_{MLE}, n^{-1}I_{\theta_o}^{-1})$ converges to zero in P_{θ_0}-probability if and only if $\sqrt{p}\sigma_n^2/\tau_n^2 \to 0$, or equivalently, if*

$$w_n = 1 - o(1/\sqrt{p_n}). \tag{5}$$

Proof. We introduce notation $P_{y,n}(d\theta)$ for the posterior distribution of $\theta|Y^n = y$ and $Q_{y,n}(d\theta)$ for the distribution centered at $\hat{\theta}_{MLE} = \bar{y}$. Thus

$$\begin{aligned} P_{y,n}(d\theta) &\leftrightarrow N_p(w_n\bar{y}, \sigma_n^2 w_n I), \\ Q_{y,n}(d\theta) &\leftrightarrow N_p(\quad \bar{y},\ \sigma_n^2 I). \end{aligned} \tag{6}$$

Let $\rho(P,Q) = \int \sqrt{p}\sqrt{q}\, d\mu$ denote the Hellinger affinity between two probability measures P, Q having densities p, q with respect to a common dominating measure μ. We recall an elementary bound [21, p. 212] for variation distance in terms of Hellinger distance and hence Hellinger affinity:

$$2[1 - \rho(P,Q)] \leq \|P - Q\| \leq \sqrt{8}[1 - \rho(P,Q)]^{1/2}, \tag{7}$$

Thus $\|P_{y,n} - Q_{y,n}\| \to 0$ if and only if $\rho(P_{y,n}, Q_{y,n}) \to 1$. We recall also that affinity commutes with products:

$$\rho\Big(\prod P_i, \prod Q_i\Big) = \prod \rho(P_i, Q_i).$$

An elementary calculation shows that

$$\rho^2(N(\theta_1, \sigma_1^2), N(\theta_2, \sigma_2^2)) = \Big(\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}\Big) \exp\Big\{-\frac{(\theta_1 - \theta_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\Big\}. \tag{8}$$

When applied to $P_{y,n}$ and $Q_{y,n}$, we set $\theta_{1i} = w_n\bar{y}_i, \theta_{2i} = \bar{y}_i$ and $\sigma_1^2 = \sigma_n^2 w_n, \sigma_2^2 = \sigma_n^2$ to obtain

$$\rho(P_{y,n}, Q_{y,n}) = \exp\Big\{ -\frac{p}{2}\log\tfrac{1}{2}(w_n^{1/2} + w_n^{-1/2}) - \frac{(1-w_n)^2}{4(1+w_n)}\frac{\|\bar{y}\|^2}{\sigma_n^2}\Big\}. \tag{9}$$

Introduce $r_n = \sigma_n^2/\tau_n^2 = w_n^{-1} - 1$. Suppose first that $pr_n^2 \to 0$. Since $w_n = (1 + r_n)^{-1}$, we have

$$\log\tfrac{1}{2}(w_n^{1/2} + w_n^{-1/2}) = -\tfrac{1}{2}\log(1 + r_n) + \log(1 + \tfrac{1}{2}r_n) \leq 2c_1 r_n^3$$

for $r_n \leq \frac{1}{2}$, say. When $p \to \infty$, we have with probability tending to one that $\|\bar{y}\|^2/\sigma_n^2 < 2p$, and so for $r_n \leq \frac{1}{2}$,

$$\frac{(1-w_n)^2}{4(1+w_n)}\frac{\|\bar{y}\|^2}{\sigma_n^2} \leq \frac{2pr_n^2}{(1+r_n)(1+r_n/2)} \leq c_2 pr_n^2.$$

Consequently, when $pr_n^2 \to 0$,

$$\rho(P_{y,n}, Q_{y,n}) \geq \exp\{-c_1 pr_n^3 - c_2 pr_n^2\} \to 1.$$

Suppose now that pr_n^2 does not approach 0. Again with probability tending to one, $\|\bar{y}\|^2/\sigma_n^2 > p/2$, and since $\frac{1}{2}(w_n^{1/2} + w_n^{-1/2}) > 1$, we have from (9) that

$$-\log \rho(P_{y,n}, Q_{y,n}) > \frac{p(1-w_n)^2}{8(1+w_n)} > c_3 \min\{pr_n^2, p\}$$

which cannot converge to zero if pr_n^2 does not. □

Remark. If $\theta_0 = \theta_{0n} \neq 0$, so that the data mean differs from the prior mean, then the rate condition is replaced by

$$w_n - 1 = o(1/q_n(\theta_{0n})), \qquad q_n(\theta_{0n}) = \sqrt{p_n} + \|\theta_{0n}\|/\sigma_n.$$

Example. We illustrate the result by considering estimation in the Gaussian white noise model (3). When expressed in a suitable orthonormal basis of wavelets, we obtain $y_{jk} \overset{\text{ind}}{\sim} N(\theta_{jk}, \sigma_n^2)$, for $k = 1, \ldots, 2^j$, and $j \in \mathbb{N}$. Pinsker's theorem [18] describes the minimax linear estimator of f, or equivalently of (θ_{jk}), under squared error loss when it is assumed that f has α mean square derivatives and shows that such minimax linear estimators are asymptotically minimax among *all* estimators as $\sigma_n \to 0$.

Pinsker's estimator is necessarily posterior mean Bayes for a corresponding Gaussian prior. The mean square differentiability condition can be equivalently expressed in terms of the coefficients as

$$\sum_{j,k} 2^{2j\alpha}\theta_{jk}^2 \leq C^2,$$

and the corresponding least favorable Gaussian prior puts

$$\theta_{jk} \overset{\text{ind}}{\sim} N(0, \tau_j^2), \qquad \tau_j^2 = \sigma_n^2(\mu_n 2^{-j\alpha} - 1)_+, \tag{10}$$

where $\mu_n = c_{\alpha n}(C/\sigma_n)^{2\alpha/(2\alpha+1)}$. The constant $c_{\alpha n}$ satisfies bounds independent of n, $c_{1\alpha} \leq c_{\alpha n} \leq c_{2\alpha}$, whose precise values are unimportant here–for further details see [14].

We consider the validity of the Bernstein-von Mises phenomenon for the collection of coefficients $\{\theta_{jk}, k = 1, \ldots, 2^j\}$ at a given level $j = j(n)$–possibly fixed, or possibly varying with n.

The prior variances τ_j^2 decrease with j, and vanish above a "critical level" $j_* = j_*(\alpha, C; n)$. Since $j_* \sim (2/(2\alpha+1))\log(C/\sigma_n)$ grows with n, so does the number of parameters $\theta_{j_*,k}$ at the critical level. From (10), we conclude that

$$\tau_{j_*}^2/\sigma_n^2 \leq 2^\alpha - 1,$$

and hence that $w_n \leq 1 - 2^{-\alpha}$ does not approach 1, so that the condition of Proposition 1 fails.

On the other hand, at a *fixed* level j_0, we have $p = 2^{j_0}$ fixed and $\tau_{j_0}^2/\sigma_n^2 = \mu_n 2^{-j_*\alpha} - 1 \to \infty$, so that $\sqrt{p}\sigma_n^2/\tau_{j_0}^2 \to 0$ and so Proposition 1 applies. Thus we may say informally that the Bernstein-von Mises phenomenon holds at a fixed level but fails at the critical level.

3. Behavior of the squared loss

In this section, we pay homage to a remarkable paper by Freedman [6], itself stimulated by Cox [4], which sets out the failure of the Bernstein-von Mises theorem in a simple sequence model of function estimation in Gaussian white noise. To further simplify the calculations, we use the growing Gaussian location model (D), (P), yielding results parallel to, but not identical with, Freedman's. Hence, define

$$T_n(\theta, Y) = \|\theta - \hat{\theta}_B\|^2 = \sum_{k=1}^{p(n)} (\theta_k - \hat{\theta}_k)^2.$$

The posterior distribution of $\theta|Y$ is described by (2); in particular the shrinkage factor $w_n = \tau_n^2/(\sigma_n^2 + \tau_n^2)$ again plays a critical role.

Theorem 2 (Bayesian). *The posterior distribution $\mathcal{L}(T_n|Y)$ is given by*

$$T_n = C_n + \sqrt{D_n} Z_{1n},$$

where

$$C_n = p\sigma_n^2 w_n, \tag{11}$$

$$\sqrt{D_n} = \sqrt{2p}\sigma_n^2 w_n \tag{12}$$

and the random variable Z_{1n} has mean 0, variance 1 and converges in distribution to $N(0,1)$ as $n \to \infty$.

Proof. From (2), the posterior distribution of T_n given Y is $\sigma_n^2 w_n \chi^2_{(p)}$ and in particular it is free of Y. Hence we have the representation

$$T_n = p\sigma_n^2 w_n + \sqrt{2p}\sigma_n^2 w_n Z_{1n},$$

and the theorem follows because $(\chi_p^2 - p)/\sqrt{2p} \Rightarrow N(0,1)$ as $p \to \infty$. □

Turn now to the frequentist perspective, in which θ is a fixed and unknown (sequence of) parameters. We will therefore use the decomposition $y_k = \theta_k + \sigma_n \epsilon_k$, with $\epsilon_k \overset{\text{iid}}{\sim} N(0,1)$, c.f. ($\mathbf{D}_{\text{seq}}$) above. Since $\hat{\theta}_{B,k} = w_n y_k$, we have

$$\theta_k - \hat{\theta}_{B,k} = (1 - w_n)\theta_k - w_n \sigma_n \epsilon_k. \tag{13}$$

Some of the conclusions will be valid only for "most" θ: to formulate this it is useful to give θ a distribution. The natural one to use is ($\mathbf{P}$), despite the possible confusion arising because, for the frequentist, this is *not* an a priori law!

Theorem 3 (Frequentist). *The conditional distribution $\mathcal{L}(T_n|\theta)$ is given by*

$$T_n = C_n + \sqrt{F_n} Z_{2n}(\theta) + \sqrt{G_n(\theta)} Z_{3n}(\theta, \epsilon), \tag{14}$$

where C_n is as in Theorem 2, while $Z_{3n}(\theta, \epsilon)$ has mean 0 and variance 1.

If θ is distributed according to ($\boldsymbol{P}$), then $Z_{2n}(\theta)$ has mean 0, variance 1 and converges in distribution to $N(0,1)$ as $n \to \infty$. In addition, if $w_n \to w = 1 - \cos\omega$,

$$\sqrt{F_n} \sim \sqrt{D_n}\cos\omega,$$

$$\sqrt{G_n(\theta)} \sim \sqrt{D_n}\sin\omega, \tag{15}$$

and

$$Z_{3n}(\theta, \cdot) \Rightarrow N(0,1). \tag{16}$$

Formulas (15) and (16) hold as $n \to \infty$, for almost all θ's generated from ($\boldsymbol{P}$).

Proof. Using (13), and $(1-w_n)^2\tau_n^2 + w_n^2\sigma_n^2 = \sigma_n^2 w_n$, we may write

$$T_n = \sum_k [(1-w_n)\theta_k - w_n\sigma_n\epsilon_k]^2 \tag{17}$$
$$= p\sigma_n^2 w_n + \sqrt{2p}\tau_n^2(1-w_n)^2 \cdot \frac{\sum \theta_k^2 - p\tau_n^2}{\sqrt{2p}\tau_n^2} + R_n(\theta,\epsilon),$$

with

$$R_n(\theta,\epsilon) = -2w_n(1-w_n)\sigma_n \sum \theta_k\epsilon_k + w_n^2\sigma_n^2 \sum(\epsilon_k^2 - 1).$$

This leads immediately to the representation (14) after observing that $\tau_n^2(1-w_n) = \sigma_n^2 w_n$ and setting

$$\sqrt{F_n} = \sqrt{2p}\sigma_n^2 w_n(1-w_n),$$
$$Z_{2n}(\theta) = \sum(\theta_k^2 - \tau_n^2)/\sqrt{2p}\tau_n^2,$$
$$G_n(\theta) = \mathrm{Var}R_n(\theta,\epsilon) = 4w_n^2(1-w_n)^2\sigma_n^2 \sum \theta_k^2 + w_n^4 \cdot 2p\sigma_n^4$$
$$= G_{1n}(\theta) + G_{2n}.$$

Turning to the final assertions, we may rewrite

$$R_n(\theta,\epsilon) = \sqrt{G_{1n}(\theta)}Z_{4n}(\theta) + \sqrt{G_{2n}}Z_{5n},$$

where

$$Z_{4n}(\theta) = \sum \theta_k\epsilon_k / \Big(\sum \theta_k^2\Big)^{1/2}, \qquad Z_{5n} = \Big(\sum \epsilon_k^2 - p\Big)/\sqrt{2p}.$$

Using again $\sigma_n^2 w_n = \tau_n^2(1-w_n)$, we have

$$G_{1n}(\theta) = 2p\sigma_n^4 \cdot w_n^2(2w_n - 2w_n^2) \cdot p^{-1}\sum_1^p (\theta_k/\tau_n)^2.$$

For almost all θ's generated from (**P**), $p^{-1}\sum_1^p(\theta_k/\tau_n)^2 \to 1$, and since $G_n(\theta) = G_{1n}(\theta) + G_{2n}$, (15) follows.

Clearly $Z_{4n}(\theta) \sim N(0,1)$, free of θ, while $Z_{5n} \Rightarrow N(0,1)$ and so (16) follows. □

Remark. The doctrinaire frequentist would not contemplate the joint distribution of (θ, Y) in (**D**, **P**); but anyone else would observe that in that joint distribution, $T_n \sim \sigma_n^2 w_n \chi^2_{(p)}$, as follows easily in two ways, either from the proof of Theorem 2, or from (17).

The Bernstein-von Mises theorem fails if $\lim w_n = w < 1$, as may be seen in Figure 1. For the Bayesian, conditional on Y, $\theta - \hat{\theta}_B$ is a noise vector, and Theorem 2 says that the distribution of $\|\theta - \hat{\theta}_B\|^2$ is approximately normal with mean C_n and standard deviation $\sqrt{D_n}$. For the frequentist, $E[\hat{\theta}_B|\theta] = w_n\theta$ is biased (also asymptotically), and some of $\|\hat{\theta}_B - \theta\|^2$ comes from this bias. As a result, Theorem 3 says that, conditional on θ, $\|\hat{\theta}_B - \theta\|^2$ is approximately normal with mean $C_n + \sqrt{F_n}Z_{2n}(\theta)$ and standard deviation $\sqrt{G_n(\theta)}$. Comparing (12) and (15) shows that the frequentist SD is smaller than the Bayesian SD: $\sqrt{G_n(\theta)} < \sqrt{D_n}$.

Under the assumption (**P**), $\theta_i \overset{\text{iid}}{\sim} N(0,\tau_n^2)$, the 'wobble' in the frequentist mean can be arbitrarily large relative to $\sqrt{D_n}$: from the law of the iterated logarithm, with probability one

$$\liminf Z_{2n}(\theta)/\sqrt{2\log\log p} = 1.$$

By contrast, if $\lim w_n = 1$, then the wobble disappears: $\sqrt{F_n} = o(\sqrt{D_n})$ and the Bayesian SD equals the frequentist SD asymptotically: $\sqrt{G_n(\theta)} \sim \sqrt{D_n}$.

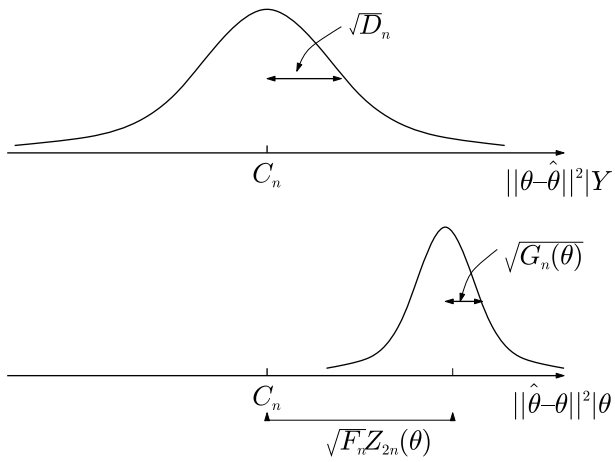

FIG 1. *The top panel, for the Bayesian, has $\theta - \hat{\theta}_B$ as a noise vector, and the posterior distribution $\mathcal{L}(T|Y)$ approximately $N(C_n, D_n)$. The bottom panel, for the frequentist, shows the effect of the bias of $E[\hat{\theta}_B|\theta]$ for θ, with $\mathcal{L}(T|\theta)$ approximately $N(C_n + \sqrt{F_n}Z_{2n}(\theta), G_{2n}(\theta))$.*

4. Linear functionals

We turn now to the least demanding of our three scenarios for the Bernstein-von Mises theorem: the behavior of linear functionals. We change the setting slightly to the infinite sequence Gaussian white noise model (3). We consider linear functionals Lf such as integrals $\int_B f$ or derivatives $f^{(r)}(t_0)$: if f has expansion $f(t) = \sum \theta_k \varphi_k(t)$, then on setting $a_k = L\varphi_k$, we have

$$Lf = \sum \theta_k L\varphi_k = \sum \theta_k a_k.$$

Again, for maximum simplicity, we consider Gaussian priors on the coefficients:

$$\theta_k \overset{\text{ind}}{\sim} N(0, \tau_k^2). \tag{18}$$

In order that $\sum \theta_k^2 < \infty$ with probability 1, it is necessary and sufficient that $\sum \tau_k^2 < \infty$.

Consequently, the posterior laws are Gaussian

$$\theta_k | y_k \overset{\text{ind}}{\sim} N(w_{kn} y_k, w_{kn}\sigma_n^2),$$

again with

$$w_{kn} = \tau_k^2/(\sigma_n^2 + \tau_k^2), \tag{19}$$

so that the posterior mean estimate

$$\widehat{Lf}_n = \sum_k a_k w_{kn} y_k.$$

Centering at posterior mean

For the Bayesian, the posterior distribution

$$Lf|y \sim N(\widehat{Lf}_n, V_{yn}), \qquad V_{yn} = \sigma_n^2 \sum_k a_k^2 w_{kn}$$

while for the frequentist, the conditional distribution

$$\widehat{Lf}_n | f \sim N(E_f \widehat{Lf}_n, V_{Fn}), \qquad V_{Fn} = \sigma_n^2 \sum_k a_k^2 w_{kn}^2.$$

The Bayesian might use $100(1-\alpha)\%$ posterior credible intervals of the form $\widehat{Lf}_n \pm z_{\alpha/2}\sqrt{V_{yn}}$, while the frequentist might employ $100(1-\alpha)\%$ confidence intervals $\widehat{Lf}_n \pm z_{\alpha/2}\sqrt{V_{Fn}}$. This leads us to consider the variance ratio

$$\frac{V_{Fn}}{V_{yn}} = \frac{\sum a_k^2 w_{kn}^2}{\sum a_k^2 w_{kn}} < 1, \tag{20}$$

from which we see that the frequentist intervals are narrower–because the frequentist bias $E_f L\hat{f} - Lf$ is being ignored for now, along with the attendant implications for coverage (but see below).

As sample size $n \to \infty$, the noise variance $\sigma_n^2 \to 0$ and so for a given Gaussian prior (18), the weights (19) converge marginally: $w_{kn} \to 1$ for each fixed k. This alone does not imply convergence of the variance ratio $V_{Fn}/V_{yn} \to 1$, as a later example shows. A sufficient condition is that the linear functional Lf be bounded (as a mapping from $L_2[0,1]$ to $\mathbb{R}$.) This amounts to saying that Lf has the representation $Lf = \int_0^1 a(t) f(t)\, dt$ with $\int a^2(t)\, dt < \infty$, or equivalently, in sequence terms, that $\sum a_k^2 \le \infty$.

Proposition 4. *Let $\mathbb{P}_f^n$ denote the measure corresponding to* (3)*. If $Lf = \int af$ is a bounded linear functional, then the variation distance between Bayesian and frequentist distributions converges to zero:*

$$\|N(\widehat{Lf}_n, V_{yn}) - N(\widehat{Lf}_n, V_{Fn})\| \xrightarrow{\mathbb{P}_f^n} 0. \tag{21}$$

Proof. We again use the Hellinger affinity (7) and apply (8) to the laws $P = N(\widehat{Lf}_n, V_{yn})$ and $Q = N(\widehat{Lf}_n, V_{Fn})$ to obtain

$$\rho^2(P,Q) = \frac{2\sqrt{V_{Fn}/V_{yn}}}{1 + V_{Fn}/V_{yn}}.$$

In view of (7), the merging in (21) occurs if and only if

$$V_{Fn}/V_{yn} \to 1.$$

When $\sum a_k^2 < \infty$, this convergence follows from (20) and the dominated convergence theorem. □

Remarks. 1. Examples of bounded functionals include polynomials $a(t) = \sum_{k=0}^K c_k t^k$ and "regions of interest" $a(t) = I\{t \in B\}$.

2. Examples of unbounded functionals are given by evaluation of a function (or its derivatives) at a point: $Lf = f^{(r)}(t_0)$. We shall see that the variance ratio does not converge to 1, and so the Bernstein-von Mises theorem fails. Indeed, in the Fourier basis

$$\varphi_0(t) \equiv 1, \qquad \begin{cases} \varphi_{2k-1}(t) = \sqrt{2}\sin 2\pi k t, \\ \varphi_{2k}(t) = \sqrt{2}\cos 2\pi k t, \end{cases} \qquad k = 1, 2, \ldots$$

we find that, $a_k = L\varphi_k = d^r\varphi_k(t_0)/dt^r$ and an easy calculation shows that $a_{2k-1}^2 + a_{2k}^2 = 2(2\pi k)^{2r}$. We use a Gaussian prior (18) with $\tau_{2k-1}^2 = \tau_{2k}^2 = k^{-2m}$ and $2m > 2r+1$. It follows from (19) that, after writing V_{1n} and V_{2n} for V_{yn} and V_{Fn} respectively, we have

$$V_{jn} = 2(2\pi)^{2r}\sigma_n^2 \sum_k k^{2r}(1+\sigma_n^2 k^{2m})^{-j}.$$

As $\lambda \to 0$, sums of the form

$$\sum_{k=0}^{\infty} k^p(1+\lambda k^q)^{-r} \sim \kappa\lambda^{-(p+1)/q},$$

with $\kappa = \kappa(p,r;q) = \int_0^\infty v^p(1+v^q)^{-r}\,dv = \Gamma(r-\mu)\Gamma(\mu)/(q\Gamma(r))$ and $\mu = (p+1)/q$. In the present case, with $p = 2r, q = 2m$ and $r = j$, we conclude that

$$\frac{V_{Fn}}{V_{yn}} \to 1 - \frac{2r+1}{2m} < 1.$$

Centering at the MLE

For a bounded linear functional, the MLE $L\hat{f}_M = \sum a_k y_k$ is well defined and unbiased, with mean $E(L\hat{f}_M) = Lf$ and frequentist variance $V_{Mn} = \mathrm{Var}(L\hat{f}_M) = \sigma_n^2 \sum_k a_k^2$. A frequentist might prefer to use $100(1-\alpha)\%$ intervals $L\hat{f}_M \pm z_{\alpha/2}\sqrt{V_{Mn}}$ which will have the correct coverage property. However, extra conditions are required for the Bernstein-von Mises result to hold in this case.

Proposition 5. *Assume that $Lf = \int af$ is a bounded linear functional. Suppose also that the coefficients of $\theta_k = \langle f, \varphi_k\rangle$ of the 'true' f, and the variances τ_k^2 of the Gaussian prior together satisfy $\sum |a_k\theta_k/\tau_k| < \infty$. Then the distance between Bayesian and frequentist distributions*

$$\|N(\widehat{Lf}_n, V_{yn}) - N(L\hat{f}_M, V_{Mn})\| \xrightarrow{\mathbb{P}_f^n} 0. \tag{22}$$

Proof. The argument is a slight elaboration of that used in the previous proposition. We use (7) and $P = N(\widehat{Lf}_n, V_{yn})$ as before, but now $Q = N(L\hat{f}_M, V_{Mn})$ and (8) yields

$$\rho^2(P,Q) = \frac{2\sqrt{V_{yn}V_{Mn}}}{V_{yn}+V_{Mn}} \exp\Big\{-\tfrac{1}{2}\frac{(L\hat{f}_M - \widehat{Lf}_n)^2}{V_{yn}+V_{Mn}}\Big\}.$$

As before $V_{yn}/V_{Mn} = \sum a_k^2 w_{kn}^2 / \sum a_k^2 \to 1$ by dominated convergence. Using this and the expression $V_{Mn} = \sigma_n^2 \sum a_k^2$, and in view of the bounds (7), the conclusion (22) is equivalent to $\sigma_n^{-1}|L\hat{f}_M - \widehat{Lf}_n| \xrightarrow{\mathbb{P}_f^n} 0$. We may write

$$\sigma_n^{-1}(L\hat{f}_M - \widehat{Lf}_n) \stackrel{\mathcal{D}}{=} \sum_k a_k\sigma_n^{-1}(1-w_{kn})\theta_k + \sum a_k(1-w_{kn})\epsilon_k.$$

The stochastic term has mean 0 and variance $\sum a_k^2(1-w_{kn})^2 \to 0$, again by dominated convergence. Thus we may focus on the deterministic term, and note that the merging in (22) occurs if and anly if

$$\sigma_n \sum_k \frac{a_k\theta_k}{\sigma_n^2+\tau_k^2} \to 0.$$

The bound $\sigma_n\tau_k/(\sigma_n^2+\tau_k^2) \le \frac{1}{2}$ along with the dominated convergence theorem then shows that $\sum |a_k\theta_k\tau_k^{-1}| < \infty$ is a sufficient condition for (22) as claimed. $\square$

5. Related work

As remarked earlier, this paper avoids the important Gaussian approximation part of the Bernstein-von Mises phenomenon by focusing on examples with Gaussian likelihoods and priors. A growing literature addresses the approximation challenges; we give a brief listing here, and refer to the books [12, 11] and the survey discussion in [10, §2.7] for more detailed discussion.

Ghosal [7, 8, 9] developed posterior normality results for the full posterior in cases where the dimension of the parameter space increases sufficiently slowly. In each case, the emphasis is on conditions under which a non-Gaussian likelihood and appropriate prior sequence can yield approximate Guassian posteriors. However Ghosal [9, Section 4] specializes his results to our setting (D) with $\sigma_n^2 = 1/n$ and notes that one can choose priors–in general not Gaussian–so that the posterior distribution centered by the MLE is approximately Gaussian if $p^3(\log p)/n \to 0$.

In survival analysis, Bernstein-von Mises theorems for the cumulative hazard function are established by Kim and Lee [16] and for the cumulative hazard and fixed dimensional covariate regression parameter in a proportional hazards model in [15].

Boucheron and Gassiat [2] develop a Bernstein-von Mises theorem for discrete probability distributions of growing dimension, and consider application to functionals such as Shannon and Renyi entropies.

In a semiparametric setting, where a finite dimensional parameter of interest can be separated from an infinite dimensional nuisance parameter, Castillo [3] obtains conditions leading to a Bernstein-von Mises theorem on the parametric part, clarifying an earlier work of Shen [20].

Rivoirard and Rousseau [19] give conditions under which Bernstein-von Mises holds for linear functionals of a nonparametrically specified probability density function.

While this manuscript was in press, we learnt of the paper by DasGupta and Lahiri [5], which does some calculations related to those in our Section 2 as part of their study of the L_1 error estimation of a Gaussian density in high dimensions.

References

[1] Borwanker, J., Kallianpur, G. and Prakasa Rao, B. L. S. (1971). The Bernstein-von Mises theorem for Markov processes. *Ann. Math. Statist.* **42** 1241–1253.

[2] Boucheron, S. and Gassiat, E. (2009). A Bernstein-von Mises theorem for discrete probability distributions. *Electron. J. Stat.* **3** 114–148.

[3] Castillo, I. (2008). A semiparametric Bernstein-von Mises theorem. Submitted.

[4] Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903–923.

[5] Dasgupta, A. and Lahiri, S. N. (2010). Density estimation in high and ultra high dimensions, regularization, and the L_1 asymptotics. In *A Festschrift for William Strawderman* (D. Fourdrinier and É. Marchand and A. Rukhin, eds.). IMS.

[6] Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite dimensional parameters. *Ann. Statist.* **27** 1119–1140.

[7] Ghosal, S. (1997). Normal approximation to the posterior distribution for generalized linear models with many covariates. *Math. Methods Statist.* **6** 332–348.

[8] Ghosal, S. (1999). Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli* **5** 315–331.

[9] Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.* **74** 49–68.

[10] Ghosal, S. (2010). The Dirichlet process, related priors and posterior asymptotics. In *Bayesian Nonparametrics* (N. L. Hjort, C. Holmes, P. Müller and S. G. Walker, eds.), Chapter 2. Cambridge Univ. Press.

[11] Ghosal, S. and van der Vaart, A. (2010). *Theory of Nonparametric Bayesian Inference*. Cambridge Univ. Press. In preparation.

[12] Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics. Springer Series in Statistics*. Springer, New York.

[13] Heyde, C. C. and Johnstone, I. M. (1979). On asymptotic posterior normality for stochastic processes. *J. Roy. Statist. Soc. Ser. B* **41** 184–189.

[14] Johnstone, I. M. (2010). Function estimation and Gaussian sequence models. Book manuscript at www-stat.stanford.edu.

[15] Kim, Y. (2006). The Bernstein-von Mises theorem for the proportional hazard model. *Ann. Statist.* **34** 1678–1700.

[16] Kim, Y. and Lee, J. (2004). A Bernstein-von Mises theorem in the nonparametric right-censoring model. *Ann. Statist.* **32** 1492–1512.

[17] Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed. *Springer Texts in Statistics*. Springer, New York.

[18] Pinsker, M. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems of Information Transmission* **16** 120–133. Originally in Russian in *Problemy Peredatsii Informatsii* **16** 52–68.

[19] Rivoirard, V. and Rousseau, J. (2009). Bernstein von Mises theorem for linear functionals of the density. Submitted.

[20] Shen, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *J. Amer. Statist. Assoc.* **97** 222–235.

[21] van der Vaart, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge.