Technical Report No. 492
Department of Statistics
University of California
Berkeley, CA 94720
1 November 1999

On the Bernstein-von Mises Theorem with Infinite Dimensional Parameters

by David Freedman*

Abstract

If there are many independent, identically distributed observations governed by a smooth, finite-dimensional statistical model, the Bayes estimate and the maximum likelihood estimate will be close. Furthermore, the posterior distribution of the parameter vector around the posterior mean will be close to the distribution of the maximum likelihood estimate around truth. Thus, Bayesian confidence sets have good frequentist coverage properties, and conversely. However, even for the simplest infinite-dimensional models, such results do not hold. The object here is to give some examples.

1. Introduction

With a large sample from a smooth, finite-dimensional statistical model, the Bayes estimate and the maximum likelihood estimate will be close. Furthermore, the posterior distribution of the parameter vector around the posterior mean must be close to the distribution of the maximum likelihood estimate around truth: both are asymptotically normal with mean 0, and both have the same asymptotic covariance matrix. That is the content of the Bernstein-von Mises theorem. Thus, a Bayesian 95%-confidence set must have frequentist coverage of about 95%, and conversely. In particular, Bayesians and frequentists are free to use each others' confidence sets. (Bayesians may view this as an advantage of their approach, since Bayesian confidence sets are relatively easy to obtain by simulation.) However, even for the simplest infinite-dimensional models, the Bernstein-von Mises theorem does not hold (Cox, 1993). The object here is to give some examples, which may help to clarify Cox's arguments.

The sad lesson for inference is this. If frequentist coverage probabilities are wanted in an infinite-dimensional problem, then frequentist coverage probabilities must be computed. Bayesians too need to proceed with caution in the infinite-dimensional case, unless they are convinced of the fine details of their priors. Indeed, the consistency of their estimates and the coverage probability of their confidence sets depend on the details of their priors. I suggest that similar conclusions apply to models with a finite—but large—number of parameters. The data swamp the prior only when the sample size is large relative to the number of parameters.

---

* This paper was presented as part of my Wald Lectures in 1998.

The examples in Cox (1993) involve continuous-time stochastic processes. Basically, there is an unknown smooth function, observed subject to random error at n points; the function is estimated using Bayesian techniques with a Gaussian prior. The examples here involve only sequences of independent normal variables, so that calculations can be done more or less explicitly. (Section 3 below indicates how Cox's examples connect with ours.) The setup is an extension of Lindley-Smith (1973) to infinitely many dimensions, and the model can be stated as follows.

(1) $Y_i = \beta_i + \epsilon_i$ for $i = 1, 2, \ldots$. The ϵ_i are independent, identically distributed normal random variables, with mean 0 and positive variance $\sigma_n^2 \to 0$, while $\sum_i \beta_i^2 < \infty$.

In principle, the variables Y_i and ϵ_i in (1) need another subscript, n. For each n, the data consist of an infinite sequence $\{Y_{n,1}, Y_{n,2}, \ldots\}$ with $Y_{n,i} = \beta_i + \epsilon_{n,i}$. Intuitively, n stands for sample size. In the leading special case, $\{Y_{n,1}, Y_{n,2}, \ldots\}$ equals the mean of n observations on a parameter vector of infinite length. The parameter vector β does not depend on the sample size, but the law of the sampling error $\epsilon_{n,i}$ certainly does. The theorems only involve the distribution of $\{\epsilon_{n,i} : i = 1, 2, \ldots\}$, which are taken to be independent, identically distributed random variables, having mean 0 and variance $\sigma_n^2 \to 0$. The joint distribution of the doubly-infinite array $\{\epsilon_{n,i} : n = 1, 2, \ldots, i = 1, 2, \ldots\}$ does not seem to matter for the results presented here. For finer estimates, however, assumptions would be needed on the doubly-infinite array. The subscript n will usually be omitted in what follows, to ease the notation.

The MLE for β is, of course, Y. We also consider a Bayesian analysis of (1), with the following prior:

(2) The β_i are independent normal variables, with mean 0 and variance τ_i^2, where $\tau_i^2 > 0$ and $\sum_i \tau_i^2 < \infty$. The β's are independent of the ϵ's.

If (2) holds, then $\sum \beta_i^2 < \infty$ almost surely. It is of some importance that there are two variance scales, an "objective" one for the ϵ's and a "subjective" one for the β's. The leading special case has $\sigma_n^2 = 1/n$ and $\tau_i^2 = 1/i^2$, corresponding to the average of n independent observations on one sequence of β_i's, the prior being specified by the choice of τ_i's. Most of the inferential difficulties already appear when i is restricted to the finite—but growing—range $i = 1, \ldots, \sqrt{n}$. Iain Johnstone has suggested a variation on this setup which makes the calculations easier: set $\tau_i^2 = 1/n$ for $i = 1, \ldots, n$; now the prior too depends on n. Further examples with interesting behavior can be obtained by setting $\tau_i^2 = A_n/n$ for $i = 1, \ldots, n$; when A_n grows with n, Johnstone's example seems to have different asymptotics from ours.

Given (1) and (2), the posterior is readily computed, as in Proposition 1. Indeed, from the Bayesian perspective, Y_i has conditional mean β_i and conditional variance σ_n^2 given β; unconditionally, however, Y_i has mean 0 and variance $\sigma_n^2 + \tau_i^2$. Furthermore, $\text{cov}(Y_i, \beta_i) = \tau_i^2$.

Proposition 1. For the Bayesian. Assume (1–2). Given the data Y, the β's are independent and normal. Moreover,

(a) $\hat{\beta}_i = E\{\beta_i | Y\} = w_{ni} Y_i$, with $w_{ni} = \tau_i^2/(\sigma_n^2 + \tau_i^2)$.
(b) $\beta_i - \hat{\beta}_i = (1 - w_{ni})\beta_i - w_{ni}\epsilon_i \perp Y$.
(c) $\text{var}\{\beta_i | Y\} = v_{ni}$, where $v_{ni} = \sigma_n^2 \tau_i^2/(\sigma_n^2 + \tau_i^2) = (1/\sigma_n^2 + 1/\tau_i^2)^{-1}$.

In effect, the proposition defines a regular conditional distribution $Q_Y(d\beta)$ for the parameter vector β given the data Y. If we consider only a finite number of β's, say $\beta_1, \ldots, \beta_k$, the Q_Y-

distribution of $\{\beta_i - \hat{\beta}_i : i = 1, \ldots, k\}$ is asymptotically the same as the frequentist distribution of $\{Y_i - \beta_i : i = 1, \ldots, k\}$, namely, these are k independent normal variables with mean 0 and variance σ_n^2. For the frequentist and the Bayesian, $\hat{\beta}_i - Y_i = o_P(\sigma_n)$. In other words, the difference between the MLE and the Bayes estimate is small compared to the randomness in either. Consequently, the posterior distribution of β around $\hat{\beta}$ is essentially the same as the frequentist distribution of $\hat{\beta}$ around β. That is (a very special case of) the Bernstein-von Mises theorem. For a brief history of this theorem, see Lehmann (1991, p. 482); for its technical details, see LeCam and Yang (1990) or Prakasa Rao (1987).

If we consider the full infinite-dimensional distribution, matters are quite different. To simplify the calculations, we assume

(3a) $\sigma_n^2 = 1/n$,

(3b) $\tau_i^2 \approx A/i^\alpha$ as $i \to \infty$, for $0 < A < \infty$ and $1 < \alpha < \infty$,

where $s_i \approx t_i$ if $s_i/t_i \to 1$ and $s_i \sim t_i$ if s_i/t_i converges to a positive, finite limit. As noted above, condition (3a) obtains if, for instance, the data are obtained by averaging n IID observations on β. The joint distribution in n does not matter here. More general τ_i^2 and σ_n^2 are considered later.

Theorem 1 below gives the Bayesian analysis; Theorem 2, the frequentist. Theorems 3 and 4 draw the implications. There is an ℓ_2 consistency result in Theorem 5. Section 2 has some complements and details. Section 3 makes the connection with stochastic processes. We focus on one infinite-dimensional functional—the square of the ℓ_2 norm. To state Theorem 1, let

$$T_n(\beta, Y) = \|\beta - \hat{\beta}\|^2 = \sum_{i=1}^{\infty} (\beta_i - \hat{\beta}_i)^2. \tag{4}$$

For the frequentist, $\beta \in \ell_2$ by assumption and $T_n < \infty$ a.e. by Proposition 1(b); likewise for the Bayesian. On the other hand, $Y \notin \ell_2$, due to the action of ϵ.

Theorem 1. For the Bayesian. Assume (1)–(3). In particular, β is random. Then $T_n = C_n + \sqrt{D_n} Z_n$, where

$$C_n = \sum_{i=1}^{\infty} v_{ni} \approx n^{-1+1/\alpha} C \text{ with } C = A^{1/\alpha} \int_0^\infty \frac{1}{1+u^\alpha} du,$$

$$D_n = 2 \sum_{i=1}^{\infty} v_{ni}^2 \approx n^{-2+1/\alpha} D \text{ with } D = 2A^{1/\alpha} \int_0^\infty \frac{1}{(1+u^\alpha)^2} du.$$

The v_{ni} are defined in Proposition 1(c); the random variable Z_n has mean 0, variance 1, and converges in law to N(0,1) as $n \to \infty$.

Proof. According to Proposition 1, given Y, T_n is distributed as $\sum_i v_{ni} \xi_{ni}^2$, the ξ_{ni} being for each n independent N(0,1) variables as i varies. (Randomness in T_n is driven by randomness in β.) Thus, $E\{T_n | Y\} = C_n$, which can be estimated as follows:

$$\sum_{i=1}^{\infty} v_{ni} \approx A \sum_{i=1}^{\infty} \frac{1}{An + i^\alpha} \approx A \int_0^\infty \frac{1}{An + x^\alpha} \, dx = A^{1/\alpha} n^{-1+1/\alpha} \int_0^\infty \frac{1}{1+u^\alpha} \, du.$$

To get the last equality, set $x = (An)^{1/\alpha}u$. This argument is heuristic but rigorizable. A variant on the idea is given by Lemma 2 in Section 2; see Remark (iii) there for details.

Moreover, $\text{var}\,\xi_{ni}^2 = 2$, so $\text{var}\{T_n|Y\} = D_n$. This too can be estimated by Lemma 2. Asymptotic normality follows from Lemma 3, with v_{ni} for c_{ni}; the condition that $\max_i v_{ni} = o(\sum_i v_{ni}^2)^{1/2}$ follows from Lemma 2: see Remark (ii) there. QED

Remarks.

(i) Proposition 1 shows that T_n is independent of Y. For the Bayesian, the predictive distribution of T_n coincides with its distribution given the data: the data are needed only to determine $\hat\beta$.

(ii) $\sqrt{D_n} << C_n$ because $n^{-1+(1/2\alpha)} << n^{-1+1/\alpha}$. If for instance $\alpha = 1/2$, then $E\{T_n\} \sim 1/n^{1/2}$ but the randomness in T_n is of order $1/n^{3/4}$. (We write $x_n << y_n$ if $x_n/y_n \to 0$.)

(iii) Fix $\delta > 0$ and $\beta \in \ell_2$. The posterior mass in a δ-ball around $\hat\beta$ tends to 1 as $n \to \infty$, for almost all data sets generated by (1). Indeed, $E\{T_n\} \sim n^{-1+1/\alpha} \to 0$ and $\text{var}\,T_n \sim n^{-2+1/\alpha}$, so

$$P\{|T_n - E(T_n)| > \delta_n\} = O\left(\frac{1}{\delta_n^2}\frac{1}{n^{2-1/\alpha}}\right),$$

which sums if for instance $\delta_n \sim 1/n^{1/\alpha'}$ and $2 - 1/\alpha - 2/\alpha' > 1$. Thus, posterior mass concentrates around $\hat\beta$ in the weak-star topology generated by the ℓ_2 norm, for almost all data generated by the model. Frequentist consistency for the Bayesian will follow—once we show that $\hat\beta$ is consistent, as in Theorem 5.

(iv) A law of the iterated logarithm is available for $T_n - E(T_n)$, as one sees by looking at

$$\sum_{i=I_n}^{J_n} v_{ni}(\xi_{ni}^2 - 1)$$

with I_n, J_n chosen so that $\tau_{I_n}^2 \approx \delta/n$ and $\tau_{J_n}^2 \approx 1/(\delta n)$. This would require some appropriate joint distribution for the ϵ's across n. Compare Cox (1993, pp.913ff).

We pursue now the frequentist analysis of the Bayes estimates. From this perspective, β is an unknown parameter, not subject to random variation. However, some results can be proved only for "most" β—and the natural (if slightly confusing) measure to use is that defined by (2).

Theorem 2. For the frequentist. Assume (1) and (3) but not (2), so β is fixed but unknown. Then

$$T_n(\beta,\epsilon) = C_n + \sqrt{F_n}U_n(\beta) + \sqrt{G_n(\beta)}V_n(\beta,\epsilon), \tag{5}$$

where C_n is as in Theorem 1, while $V_n(\beta,\cdot)$ has mean 0 and variance 1. If β is distributed according to (2), then $U_n(\beta)$ has mean 0, variance 1, and converges in law to N(0,1) as $n \to \infty$. Furthermore,

$$F_n \approx n^{-2+1/\alpha}F, \quad \text{with} \quad F = 2A^{1/\alpha}\int_0^\infty \frac{u^{2\alpha}}{(1+u^\alpha)^4}\,du, \tag{6}$$

(7) $$G_n(\beta) \approx n^{-2+1/\alpha} G, \quad \text{with} \quad G = 2A^{1/\alpha} \int_0^\infty \frac{2u^\alpha + 1}{(1+u^\alpha)^4}\, du,$$

and

(8) $$V_n(\beta, \cdot) \text{ converges in law to N(0,1).}$$

Displays (7) and (8) hold as $n \to \infty$, for almost all β's generated by (2).

This theorem is proved like Theorem 1. Tedious details, along with explicit formulae for F_n, G_n, U_n, and V_n, are postponed to Section 2. The theorem describes the asymptotic behavior of the Bayesian pivot, $T_n = \|\hat{\beta} - \beta\|^2$, from a frequentist perspective. For this purpose, the frequentist agrees to use the same joint distribution for β and Y as the Bayesian. Of course, the Bayesian will compute $\mathcal{L}(T_n|Y)$. The frequentist cannot go that far, but considers $\mathcal{L}(T_n|\beta)$. Among other things, the frequentist has agreed to ignore bad behavior for an exceptional null set of β's—relative to the prior (2). There are results on minimax rates for Bayes estimates, suggesting that β's exist for which the rate of convergence is slower than $n^{-2+1/\alpha}$, so the Bayesian null set may in some other sense be quite large. See Zhao (1997) or Brown, Low, and Zhao (1998); also see Section 4 below, and compare Theorem 3.1 in Cox (1993).

Contrary to experience with the finite-dimensional case, there is a radical difference between the asymptotic behavior of $\mathcal{L}(T_n|Y)$ and the asymptotic behavior of $\mathcal{L}(T_n|\beta)$—even if we ignore the null set of bad β's. Our next main result is Theorem 3, which shows that the Bernstein-von Mises theorem does not apply in the infinite-dimensional context. There will be two reasons.

(i) For the frequentist, the variance of $\hat{\beta}$ is driven by ϵ, that is, by the last term in (5). And this variance is smaller than the Bayes variance. See Theorem 3.

(ii) The middle term in (5) wobbles on the scale of interest, namely, $n^{-1+1/(2\alpha)}$; so the frequentist distribution of T_n is offset from the Bayesian distribution by arbitrarily large amounts. This a consequence of "Bayes bias."

Corollary 1 demonstrates the wobble, the proof being deferred to Section 2.

Corollary 1. Assume (1) and (3). Then

$$\limsup_{n\to\infty} U_n(\beta) = \infty \quad \text{and} \quad \liminf_{n\to\infty} U_n(\beta) = -\infty$$

for almost all β drawn from (2).

Theorem 3. For the frequentist. Assume (1) and (3). The Bayesian posterior is computed from (2), and the frequentist conclusions apply to almost all β drawn from (2). The asymptotic variances D and G are defined in Theorems 1 and 2.

(a) $G < D$. In particular, the asymptotic frequentist variance is smaller than the asymptotic Bayes variance.

(b) There is almost surely a sequence of n's tending to infinity such that the frequentist distribution of T_n is centered to the right of the Bayes distribution by arbitrarily large multiples of $n^{-1+1/(2\alpha)}$, and likewise to the left.

Proof. The inequality in (a) is elementary: $2u^\alpha + 1 < (1 + u^\alpha)^2$. Then use Corollary 1 to prove claim (b). QED

The first part of the theorem—$G < D$—already shows that the conclusions of the Bernstein-von Mises theorem do not hold. More particularly, $F + G = D$. A posteriori, the Bayesian sees β as centered at $\hat{\beta}$, so $\Delta = \hat{\beta} - \beta$ is centered at 0, and $\|\Delta\|^2$ is the squared length of a noise vector. For the frequentist, on the other hand, $\hat{\beta}$ is biased, Δ is not centered at 0, some of $\|\Delta\|^2$ comes from bias and some from randomness. In effect, some Bayesian randomness is reinterpreted as bias. This effect is harder to see in a finite number of dimensions. For results showing that the Bayes bias term matters when rates of convergence are slower than $1/\sqrt{n}$, see Brown and Liu (1993) or Pfanzagl (1998).

The second part of Theorem 3 shows that for certain random times, the posterior distribution of β around $\hat{\beta}$ is nearly orthogonal to the frequentist distribution of $\hat{\beta}$ around β: recall that two probabilities μ and ν are "othogonal" if there is a set A with $\mu(A) = 1$ and $\nu(A) = 0$. This is perhaps a more poignant version of the failure in the conclusions of the Bernstein-von Mises theorem.

We now sharpen the orthogonality result. Consider the random variables

$$W_{ni} = (\beta_i - \hat{\beta}_i)/\sqrt{v_{ni}}\colon\ i = 1, 2, \ldots.$$

Let π_n stand for the Bayesian distribution of $W_{n1}, W_{n2}, \ldots$. This is the posterior distribution, centered and standardized; the randomness is in the parameters, not the data. (From the Bayesian perspective, the W's are independent of the data.) Technically, π_n is a probability on R^∞, the space of sequences of real numbers. Let $\phi_{n,\beta}$ be the frequentist distribution for the same random variables, with the signs reversed:

$$W'_{ni} = (\hat{\beta}_i - \beta_i)/\sqrt{v_{ni}}\colon\ i = 1, 2, \ldots.$$

Now, β is fixed and the randomness is in the data. Again, $\phi_{n,\beta}$ is a probability on R^∞. There is a third distribution to consider: $\psi_n(\beta)$, the law of

$$W''_{ni} = (Y_i - \beta_i)/\sqrt{v_{ni}}\colon\ i = 1, 2, \ldots.$$

For the frequentist, ϕ is the law of the Bayes estimates, centered at the true parameters; ψ is the law of the MLE, also centered at truth. For mathematical convenience, all three laws are standardized using the Bayesian variance; this common standardization cannot affect the orthogonality.

Theorem 4. For the neutral observer. Assume (1). The Bayesian posterior is computed from (2). Condition (3) is not needed.

(a) For the Bayesian, π_n makes the coordinates independent N(0,1) variables.
(b) For the frequentist, $\phi_{n,\beta}$ makes the coordinates independent normal variables. The ith variable has mean $-(1 - w_{ni})\beta_i/\sqrt{v_{ni}}$ and variance w_{ni}.
(c) For the frequentist, $\psi_{n,\beta}$ makes the coordinates independent normal variables. The ith variable has mean 0 and variance $1/w_{ni}$.

(d) For any n and any $\beta,\ \beta^* \in \ell_2$, the probabilities π_n , $\phi_{n,\beta}$ and ψ_{n,β^*} are pairwise orthogonal.

Claims (a), (b) and (c) are immediate from Proposition 1. The proof of (d) is deferred to Section 2, but the idea is simple: although the three probabilities merge on any fixed number of coordinates, the scales are radically different at ∞. The curious centering for ϕ cannot undo the scaling. In the frequentist vision of things, the MLE and the Bayes estimate are radically different. Moreover, the Bayesian a posteriori distribution for the parameters around the Bayes estimate is radically different from the frequentist distribution of the MLE around truth—or the frequentist distribution of the Bayes estimate around truth.

The last result in this section establishes the frequentist consistency of the Bayes estimates: the chance that $\hat{\beta}$ is close to β in ℓ_2 tends to 1 as n gets large. This theorem can be proved for any $\beta \in \ell_2$.

Theorem 5. Assume (1). The Bayes estimate computed from (2) is consistent for all β in ℓ_2, namely, $\|\hat{\beta} - \beta\|^2 \to 0$ in probability. If (3) holds, convergence a.e. will obtain.

Proof. To begin with, by Proposition 1(b),

$$\|\hat{\beta} - \beta\|^2 \le 2\sum_{i=1}^{\infty}(1 - w_{ni})^2\beta_i^2 + 2\sum_i w_{ni}^2\epsilon_i^2. \tag{9}$$

But $w_{ni} \to 1$ as $n \to 0$ for each i. By dominated convergence, the first sum on the right in (9) tends to 0. The expection of the second sum is

$$\sum_{i=1}^{\infty} \frac{\sigma_n^2\tau_i^2}{(\sigma_n^2 + \tau_i^2)^2}\tau_i^2. \tag{10}$$

Again, this goes to zero by dominated convergence: τ_i^2 sums in i, while the coefficients are bounded above by $1/2$ because $ab/(a+b)^2 \le 1/2$. This proves convergence in probability, and we turn to the a.e. result.

The variance of the second sum on the right in (9) is $8q_n$, where

$$q_n = \sigma_n^4 \sum_{i=1}^{\infty} \frac{\tau_i^8}{(\sigma_n^2 + \tau_i^2)^4}. \tag{11}$$

If (3) holds, $q_n \sim 1/n^{2-1/\alpha}$ by Lemma 2 in Section 2, and $\sum q_n < \infty$; convergence a.e. follows from the Borel-Cantelli lemma. QED

2. Complements and details

Lemma 1. Let $1 < \alpha < \infty$, $1 < b < \infty$, and $0 \le c < \infty$. Suppose $\alpha b > c + 1$. Let $f(u) = u^c/(1 + u^\alpha)^b$. Let $h > 0$. Then

$$\lim_{h \to 0} \sum_{i=1}^{\infty} f(ih)h = \int_0^\infty f(u)\,du.$$

Proof. Let L be a large, positive real number. Of course,

$$\lim_{h \to 0} \sum_{i=1}^{L/h} f(ih)h = \int_0^L f(u)\,du$$

and

$$\lim_{L \to \infty} \int_L^\infty f(u)\,du = 0.$$

Abbreviate $\gamma = \alpha b - c > 1$, and let C_γ be a suitable positive constant depending only on γ. We let $h \to 0$ first, and then $L \to \infty$. Since $f(u) < u^{-\gamma}$,

$$\sum_{i=L/h}^{\infty} f(ih)h < h^{1-\gamma} \sum_{i=L/h}^{\infty} 1/i^\gamma < C_\gamma h^{1-\gamma}(h/L)^{\gamma-1} = C_\gamma/L^{\gamma-1},$$

which is small for large L. QED

Lemma 2. Let $1 < \alpha < \infty$, $1 < b < \infty$, and $0 \le c < \infty$. Suppose $\alpha b > c + 1$. Let $\gamma_n \to \infty$. Let $s_i \approx i^\alpha$ and $t_i \approx i^c$. Let $g_n = \gamma_n^{b-(1+c)/\alpha}$. Let i_1 be a positive integer.

(a) $$\lim_{n \to \infty} g_n \sum_{i=i_1}^{\infty} \frac{t_i}{(\gamma_n + s_i)^b} = \lim_{n \to \infty} g_n \sum_{i=i_1}^{\infty} \frac{i^c}{(\gamma_n + i^\alpha)^b} = \int_0^\infty \frac{u^c}{(1+u^\alpha)^b}\,du;$$

(b) $$\max_i \frac{t_i}{(\gamma_n + s_i)^b} \sim \gamma_n^{(c/\alpha)-b}.$$

Proof. For the first equality in (a), an upper bound can be obtained if $s_i \ge (1 - \epsilon)i^\alpha$ and $t_i \le (1 + \epsilon)i^c$ for $i \ge i_1$; likewise for lower bounds. Set $h = \gamma_n^{-1/\alpha}$ and define f as in Lemma 1. By tedious algebra, $g_n i^c/(\gamma_n + i^\alpha)^b = hf(ih)$, and Lemma 1 completes the proof of part (a). For (b), it is easy to verify that f has a maximum on $[0, \infty)$. So the max in (b) is $O(h/g_n)$. But $h/g_n = \gamma_n^{(c/\alpha)-b}$. QED

Remarks.

(i) The γ in the proof of Lemma 1 is unrelated to the γ_n in the statement of Lemma 2.
(ii) The max in i of $t_i/(\gamma_n + s_i)^b$ is smaller than the sum of these terms, by a factor asymptotically of order $\gamma_n^{1/\alpha}$.
(iii) To estimate C_n in Theorem 1, take $\gamma_n = A/\sigma_n^2$, $c = 0$, and $b = 1$; for D_n, take $b = 2$.
(iv) Lemma 1 can of course be generalized, for instance, to functions convex on (x_0, ∞). In our applications $\alpha,\ b > 1$, but all that seems to be needed here is $\alpha,\ b > 0$.

Lemma 3. Let c_{ni} be constants with $0 < c_n^2 = \sum_i c_{ni}^2 < \infty$, and $\max_i |c_{ni}| = o(c_n)$ as $n \to \infty$. Let X_{ni} be random variables which are independent in i for each n and have common distribution for all n and i. Suppose $E\{X_{ni}\} = 0$ and $E\{X_{ni}^2\} = \sigma^2 < \infty$. Then

$$\frac{1}{c_n} \sum_{i=1}^{\infty} c_{ni} X_{ni} \to \mathrm{N}(0, \sigma^2)$$

in law as $n \to \infty$.

Proof. This is immediate from Lindeberg's theorem. QED

Remark. It is enough that the X_{ni} are uniformly L_2, with constant variance.

Lemma 4. Let U_i be independent N(0,1) variables. Let c_i be real numbers with $c^2 = \sum_i c_i^2 < \infty$. Let $\delta > 0$ with $\delta |c_i|/c^2 < 1$ for all i, and let $V = \sum_i c_i (U_i^2 - 1)$. Then

$$P\{V > \delta\} < \exp[-\delta^2/(12c^2)],$$

where $\exp(x) = e^x$. Likewise, $P\{V < -\delta\} < \exp[-\delta^2/(12c^2)]$.

Proof. If U is N(0,1) and $\lambda < .2$, we claim

$$E\{\exp[\lambda(U^2 - 1)]\} = \frac{\exp(-\lambda)}{\sqrt{1 - 2\lambda}} \le \exp(3\lambda^2). \tag{12}$$

The inequality is strict except at $\lambda = 0$. To prove the inequality, square both sides and take logs: we have to prove $f \ge 0$, with $f(\lambda) = \log(1 - 2\lambda) + 2\lambda + 6\lambda^2$. Now $f(0) = f'(0) = 0$. And

$$f''(\lambda) = 12 - \frac{4}{(1 - 2\lambda)^2} > 0$$

provided $\lambda < \left(1 - \sqrt{1/3}\right)/2 = .21\ldots$. In this range, f' is strictly increasing and f is strictly convex, decreasing for $\lambda < 0$ and increasing for $\lambda > 0$. So $f(\lambda) > 0$ except at $\lambda = 0$, proving the inequality in (12).

Next,

$$E\{\exp(\theta V)\} = \prod E\{\exp[\theta c_i (U_i^2 - 1)]\} < \exp(3\theta^2 c^2), \tag{13}$$

provided $\theta c_i < .2$ for all i. Now

$$P\{V > \delta\} < \exp(3\theta^2 c^2 - \theta\delta)$$

by Chebychev's inequality. Choose $\theta = \delta/(6c^2)$, which satisfies the condition $\theta c_i < .2$ by assumption. The bound on $P\{V < -\delta\}$ follows, on changing the signs of all the c_i. QED

Proof of Theorem 2. To ease notation, we consider $\sigma_n^2 = 1/n$ and $\tau_i^2 = 1/i^\alpha$, so $w_{ni} = n/(n+i^\alpha)$ and $v_{ni} = 1/(n+i^\alpha)$ from Proposition 1. Modifications of the proof for $\tau_i^2 = A/i^\alpha$ are obvious, and then $\tau_i^2 \approx A/i^\alpha$ is quite easy. Recall the definition (4) of T_n. By Proposition 1,

$$\begin{aligned} T_n &= \sum_{i=1}^{\infty}(1-w_{ni})^2\beta_i^2 - 2\sum_{i=1}^{\infty}(1-w_{ni})w_{ni}\beta_i\epsilon_i + \sum_{i=1}^{\infty} w_{ni}^2\epsilon_i^2 \\ &= C_n + Q_n(\beta) + R_n(\beta,\epsilon) \end{aligned}$$

where

$$\begin{aligned} C_n &= \sum_{i=1}^{\infty}(1-w_{ni})^2\tau_i^2 + \sigma_n^2\sum_{i=1}^{\infty} w_{ni}^2 \;= \sum_{i=1}^{\infty}\frac{1}{(n+i^\alpha)} \\ Q_n(\beta) &= \sum_{i=1}^{\infty}(1-w_{ni})^2(\beta_i^2-\tau_i^2) \\ R_n(\beta,\epsilon) &= \sum_{i=1}^{\infty} -2w_{ni}(1-w_{ni})\beta_i\epsilon_i + w_{ni}^2(\epsilon_i^2-\sigma_n^2). \end{aligned}$$

Remark. In the finite-dimensional case, $w_{ni} \approx 1$ for n large; from either the Bayesian or the frequentist perspective, only the ϵ^2-terms in $R_n(\beta,\epsilon)$ contribute to the asymptotic variance of T_n. In the infinite case, Q_n matters, and so do the $\beta\epsilon$ terms. That is the novelty.

By Proposition 1, C_n matches the lead term in Theorem 1. We turn now to $Q_n(\beta)$. Let π stand for the prior distribution on β, as defined in (2). Clearly,

$$Q_n(\cdot) = \sum_{i=1}^{\infty}\frac{i^\alpha}{(n+i^\alpha)^2}(\zeta_i^2-1), \tag{14}$$

where $\zeta_i = \zeta_i(\beta) = \sqrt{i^\alpha}\beta_i$. The ζ_i are independent N(0,1) variables relative to π. In particular, $E_\pi\{Q_n(\cdot)\} = 0$ and

$$\mathrm{var}_\pi\, Q(\cdot) = 2\sum_i \frac{i^{2\alpha}}{(n+i^\alpha)^4}, \tag{15}$$

which is (by definition) the F_n in Theorem 2. Lemma 2 can be used to estimate F_n, proving (6). The $U_n(\beta)$ in Theorem 2 is defined as $Q_n(\beta)/\sqrt{F_n}$. Of course, U_n has mean 0 and variance 1

relative to π. Asymptotic normality follows from Lemma 3. The condition that $\max_i |c_{ni}| = o(c_n)$ follows, as before, from Lemma 2.

Next, we need to consider R_n. Here, the computation is more intricate. We begin with the sum of the $\beta\epsilon$ terms. Fix any $\beta \in \ell_2$. As before, let $\zeta_i(\beta) = \sqrt{i^\alpha}\beta_i$. Also let

$$k_{ni}(\beta) = -\frac{2\sqrt{ni^\alpha}}{(n+i^\alpha)^2}\zeta_i(\beta). \tag{16}$$

For motivation, $-2w_{ni}(1-w_{ni})\beta_i\epsilon_i = [-2\sqrt{ni^\alpha}/(n+i^\alpha)^2]\zeta_i(\beta)\xi_{ni}$, where $\xi_{ni} = \sqrt{n}\epsilon_i$ is N(0,1), so k_{ni} is the coefficient of an N(0,1) variable in the expansion of R_n. In any event,

$$K_{ni}^2 = E_\pi\{k_{ni}(\cdot)^2\} = \frac{4ni^\alpha}{(n+i^\alpha)^4}. \tag{17}$$

By Lemma 2,

$$K_n^2 = \sum_{i=1}^\infty K_{ni}^2 = \sum_{i=1}^\infty \frac{4ni^\alpha}{(n+i^\alpha)^4} \approx \frac{1}{n^{2-1/\alpha}}\int_0^\infty \frac{4u^\alpha}{(1+u^\alpha)^4}\,du. \tag{18}$$

We claim that

$$\Delta_n(\beta) = \Big[\sum_{i=1}^\infty k_{ni}(\beta)^2\Big] - K_n^2 = o\big(1/n^{2-1/\alpha}\big) \text{ as } n \to \infty, \tag{19}$$

for almost all β drawn from (2). Indeed,

$$\Delta_n = \sum_{i=1}^\infty K_{ni}^2(\zeta_i^2 - 1).$$

Now use Lemma 4, with K_{ni}^2 for c_i. Lemma 2 shows that $\max_i K_{ni}^2 \sim 1/n^2$ and $\sum_i K_{ni}^4 \sim 1/n^{4-1/\alpha}$. Fix $\delta_0 > 0$ but small, and use $\delta_0/n^{2-1/\alpha}$ for the δ of Lemma 4:

$$P\big\{|\Delta_n| > \delta_0/n^{2-1/\alpha}\big\} < 2\exp\big(-\text{const.}\, n^{1/\alpha}\big),$$

which sums in n, proving (19). The condition of the lemma holds if δ_0 is small.

We turn now to the sum of the ϵ^2 terms in R_n. Recall that $\xi_{ni} = \sqrt{n}\epsilon_i$. Let $\ell_{ni} = n/(n+i^\alpha)^2$. Then

$$R_n(\beta,\epsilon) = \sum_{i=1}^\infty \big[k_{ni}(\beta)\xi_{ni} + \ell_{ni}(\xi_{ni}^2 - 1)\big]. \tag{20}$$

Here, $k_{ni}(\beta)$ depends on β while ℓ_{ni} is deterministic. And the two seem to be on different scales. (With more effort, however, the scales can be seen as comparable, for the i's that matter. In any case, the total variances are comparable, as will be seen below.)

Lemma 3 does not apply, and a direct appeal must be made to Lindeberg's theorem. First, however,

$$L_n^2 = \sum_{i=1}^{\infty} \ell_{ni}^2 = \sum_{i=1}^{\infty} \frac{n^2}{(n+i^\alpha)^4} \approx \frac{1}{n^{2-1/\alpha}} \int_0^\infty \frac{1}{(1+u^\alpha)^4}\, du. \tag{21}$$

The $G_n(\beta)$ of Theorem 2 is defined as

$$G_n(\beta) = \sum_{i=1}^{\infty} k_{ni}(\beta)^2 + 2\sum_{i=1}^{\infty} \ell_{ni}^2, \tag{22}$$

the factor of 2 being the variance of $\xi_{ni}^2 - 1$. And the $V_n(\beta, \epsilon)$ of the theorem is

$$R_n(\beta, \epsilon)/\sqrt{G_n(\beta)}.$$

Estimates (18), (19) and (21) give a (painful) verification of (7).

Fix $\delta > 0$ small. Except for a set of β's of measure 0, $|\zeta_i(\beta)| < i^\delta$ for all but finitely many i, and $\Delta_n(\beta) = o\big(1/n^{2-1/\alpha}\big)$ by (19). For the remaining β's, we claim, $V_n(\beta, \cdot)$ converges in law to N(0,1) as $n \to \infty$. This will be demonstrated by verifying the Lindeberg condition. The condition involves estimating a series of integrals of the form

$$\int_{|U+V|>2a} (U+V)^2,$$

where a is a small multiple of the asymptotic standard deviation. Now $|U+V| > 2a$ entails $|U| > a$ or $|V| > a$, and $(U+V)^2 \le 2(U^2+V^2)$, so each integral can be estimated according to the scheme

$$\frac{1}{2}\int_{|U+V|>2a} (U+V)^2 \le \int_{|U|>a} U^2 + \int_{|U|\le a<|V|} U^2 + \int_{|V|>a} V^2 + \int_{|V|\le a<|U|} V^2. \tag{23}$$

To flesh this out, the variance of $R_n(\beta, \cdot)$ is $s_n^2 = G_n(\beta) \sim 1/n^{2-1/\alpha}$. Let $\tilde{\xi}_{ni} = k_{ni}(\beta)\xi_{ni}$ and $\check{\xi}_{ni} = \ell_{ni}(\xi_{ni}^2 - 1)$ in (20). We need to show that

$$\sum_{i=1}^{\infty} \int_{|\tilde{\xi}_{ni}+\check{\xi}_{ni}|>\kappa s_n} (\tilde{\xi}_{ni} + \check{\xi}_{ni})^2 = o(s_n^2),$$

with κ a generic small number. In view of (23), we need only show that

$$S_{\nu,n} = o(s_n^2) \text{ for } \nu = 1, 2, 3, 4,$$

where

$$S_{1,n} = \sum_{i=1}^{\infty} \int_{|\tilde{\xi}_{ni}|>\kappa s_n} \tilde{\xi}_{ni}^2, \qquad S_{2,n} = \sum_{i=1}^{\infty} \int_{|\tilde{\xi}_{ni}|\le\kappa s_n<|\check{\xi}_{ni}|} \tilde{\xi}_{ni}^2,$$

$$S_{3,n} = \sum_{i=1}^{\infty} \int_{|\check{\xi}_{ni}|>\kappa s_n} \check{\xi}_{ni}^2, \qquad S_{4,n} = \sum_{i=1}^{\infty} \int_{|\check{\xi}_{ni}|\le\kappa s_n<|\tilde{\xi}_{ni}|} \check{\xi}_{ni}^2.$$

We begin with $S_{1,n}$. Recall the definition (16) of k_{ni}. For our β's, $\zeta_i(\beta) < i^\delta$ eventually, so

$$\max_i |k_{ni}(\beta)| = O\big(1/n^{1-\delta/\alpha}\big)$$

by Lemma 2. Now $s_n \sim 1/n^{1-1/(2\alpha)}$, and $|\tilde{\xi}_{ni}| = |k_{ni}(\beta)\xi_{ni}| > \kappa s_n$ entails $|\xi_{ni}| > \text{const.}\, n^\gamma$, where $\gamma = (\frac{1}{2} - \delta)/\alpha$, which is positive when $0 < \delta < \frac{1}{2}$. Consequently, the ith term in $S_{1,n}$ is bounded by

$$k_{ni}(\beta)^2 \int_{|\xi_{ni}|>\text{const.}\, n^\gamma} \xi_{ni}^2,$$

from which it is immediate that $S_{1,n} = o(s_n^2)$. Indeed, for our β's, $\sum_i k_{ni}(\beta)^2 \sim 1/n^{2-1/\alpha} \sim s_n^2$.

The argument for $S_{3,n}$ is similar but easier, because $\ell_{ni} = n/(n + i^\alpha)^2 < 1/n$. In $S_{2,n}$, $|\check{\xi}_{ni}| > \kappa s_n$ entails $|\xi_{ni}^2 - 1| > \text{const.}\, n^{1/(2\alpha)}$, hence $|\xi_{ni}| > \text{const.}\, n^{1/(4\alpha)}$. The ith term in $S_{2,n}$ is therefore bounded by

$$k_{ni}^2 \int_{|\xi_{ni}|>\text{const.}\, n^{1/(4\alpha)}} \xi_{ni}^2,$$

from which it is immediate that $S_{2,n} = o(s_n^2)$. The argument for $S_{4,n}$ is similar. We have verified the Lindeberg condition, proving (8) and so the theorem. QED

Remarks.

(i) As shown by Proposition 1(b), the term $\sqrt{F_n}U_n(\beta) = Q_n(\beta)$ in (5) is the squared norm of the Bayes bias, centered at its mean relative to the prior defined by (2). It is this deviation which wobbles on the scale of interest.

(ii) The proof exploits the fact that $\xi^2 - 1$ is a function of ξ. However, the two variables are uncorrelated: asymptotically, the sum of the $\beta\epsilon$ terms in R_n is therefore independent of the sum of the ϵ^2 terms. This would follow from the bivariate form of Lindeberg's theorem. On the other hand, (23) is enough to derive the requisite bivariate form of the theorem from the univariate.

Proof of Corollary 1. The argument starts from (14), where $\zeta_i(\beta) = \sqrt{\tau_i}\beta_i$ is a sequence of IID N(0,1) variables that does not depend on n. The F_n in Theorem 2 is the right hand side of (15), while $U_n(\beta) = Q_n(\beta)/\sqrt{F_n}$. Fix a positive integer j. Now $\{U_n > j \text{ i.o.}\}$ is a tail set relative to the ζ's, by Lemma 2(b). This set has positive probability, by asymptotic normality; hence, the probability is 1. Likewise for $-\infty$. QED

Proof of Theorem 4(d). The two probabilities in a pair are either equivalent or singular, and Kakutani's criterion can be used to decide. See, for instance, Williams (1991). Fix n, and $\beta \in \ell_2$. By Proposition 1, $w_{ni} \to 0$ as $i \to \infty$. If we compare π with ϕ, the frequentist variance for the ith variable is neglible relative to the Bayes variance; for equivalence to obtain, the ratio of the variances would need to tend to 1. Likewise for the other comparisons.

Faster decay rates

Theorem 2 depends on the assumed tail behavior of the prior variances τ_i^2. In particular, if α is large, the wobbly middle term in (5) is relatively small. As it turns out, with faster decay rates,

this middle term is negligible. So the conclusions of the Bernstein-von Mises theorem apply to our quadratic functional even with infinitely many parameters. To simplify the notation, we consider only $\sigma_n^2 = 1/n$ and $\tau_i^2 = e^{-\alpha i}$.

Lemma 5. (a) If $\alpha, b > 0$ then

$$\sum_{i=1}^{\infty} \frac{1}{(n + e^{\alpha i})^b} \approx \frac{\log n}{\alpha n^b}.$$

(b) If $\alpha, b, c > 0$ and $\alpha b > c$ then

$$\sum_{i=1}^{\infty} \frac{e^{ci}}{(n + e^{\alpha i})^b} \sim n^{-b+c/\alpha}.$$

Proof. *Claim (a).* The sum to be estimated is $n^{-b} S_n$, where

$$S_n = \sum_{i=1}^{\infty} \frac{1}{\left[1 + e^{\alpha[i - \log(n^{1/\alpha})]}\right]^b}.$$

Fix L_0, a large positive integer. Let L_n be the integer part of $\log(n^{1/\alpha}) - L_0$. Each term in S_n is bounded above by 1, and each of the first L_n terms is bounded below by $1/[1+\exp(-\alpha L_0)]^b$, which is rather close to 1. The sum of the first L_n terms in S_n is therefore essentially $L_n \approx \log(n^{1/\alpha})$. The sum of the remaining terms is bounded above by

$$\sum_{i=-L_0-1}^{\infty} \frac{1}{(1 + e^{\alpha i})^b} = O(1) = o(\log n).$$

Let $n \to \infty$ and then $L_0 \to \infty$.

Claim (b). Let L_n be the integer part of $\log(n^{1/\alpha})$. The sum to be estimated is $n^{-b+c/\alpha} S_n$, where

$$S_n = \sum_{i=1}^{\infty} \frac{e^{c[i - \log(n^{1/\alpha})]}}{\left[1 + e^{\alpha[i - \log(n^{1/\alpha})]}\right]^b}.$$

An upper bound on S_n is

$$\sum_{i=1}^{\infty} \frac{e^{c[i-L_n]}}{[1 + e^{\alpha(i-L_n-1)}]^b} \to \sum_{j=-\infty}^{\infty} \frac{e^{cj}}{[1 + e^{\alpha(j-1)}]^b}.$$

Similarly, an asymptotic lower bound is

$$\sum_{j=-\infty}^{\infty} \frac{e^{c(j-1)}}{(1+e^{\alpha j})^b}.$$

QED

Theorem 6. Assume (1). Instead of (3), suppose $\sigma_n^2 = 1/n$ and $v_i = e^{-\alpha i}$ where $0 < \alpha < \infty$. The posterior is computed from (2), and frequentist probability statements about β are also made relative to (2). In probability, as $n \to \infty$, the frequentist distribution of $\|\beta - \hat{\beta}\|^2$ merges with the Bayesian distribution.

Proof. Theorem 1 and its proof go through, with $C_n \approx (\log n)/(\alpha n)$ and $D_n \approx (2\log n)/(\alpha n^2)$. Theorem 2 also goes through; now, however, $F_n = O(1/n^2)$ is too small to matter: in the notation of this section, $Q_n(\beta)$ can be ignored. We turn now to $R_n(\beta, \epsilon)$. The sum of the ϵ^2 terms has asymptotic variance $(2\log n)/(\alpha n^2)$, like the Bayesian variance. It remains only to show that the total frequentist variance of the $\beta\epsilon$ terms in R_n is $O(1/n^2)$, in probability, for β chosen from (2). Let $V_n(\beta)$ be the variance of these $\beta\epsilon$ terms. Then

$$V_n(\beta) = \sum_{i=1}^{\infty} \frac{4ne^{\alpha i}}{(n+e^{\alpha i})^4} \zeta_i(\beta)^2,$$

where, as before, $\zeta_i(\beta) = \sqrt{e^{\alpha i}}\beta_i$ are independent N(0,1) variables. We compute the expected value and variance of $V_n(\cdot)$, relative to the probability π on β defined by (2): $E_\pi\{V_n(\cdot)\} = O(1/n^2)$ and

$$\mathrm{var}_\pi\{V_n(\cdot)\} = \sum_{i=1}^{\infty} \frac{16n^2e^{2\alpha i}}{(n+e^{\alpha i})^8} = O(1/n^4),$$

by Lemma 5. QED

Remarks.

(i) The statement of Theorem 6 can be clarified as follows. For $\beta \in \ell_2$, let $\phi_{n,\beta}$ be the frequentist distribution of $[T_n(\beta, Y) - C_n]/\sqrt{D_n}$: the randomness is in Y. Likewise, let π_n be the Bayesian distribution of $[T_n(\beta, Y) - C_n]/\sqrt{D_n}$. Here, the randomness is in β, because T_n is—for the Bayesian—independent of Y. Let ρ metrize the weak-star topology on probabilities in R^1 and let ν be the standard normal distribution. Then $\rho(\phi_{n,\beta}, \nu) \to 0$ in probability as $n \to 0$, where "in probability" is relative to the probability on β's defined by (2). Furthermore, $\rho(\pi_n, \nu) \to 0$. Stronger metrics could be used, but that is perhaps not the critical issue here.

(ii) Preliminary calculations suggest that Theorem 6 does not hold a.e.—that is, $\rho(\phi_{n,\beta}, \nu)$ does not converge to 0 for almost all β drawn from (2)—because there are arbitrarily large random n with $V_n \sim \log n/n^2$. In other words, the Bayes bias term shows a limited degree of wobble, almost surely. Indeed, $n^2V_n = \sum_i c_{ni}\zeta_i^2$ where $c_{ni} = O(1)$, with a maximum at $i = L_n$ or $L_n + 1$, where L_n is the integer part

$$L_n = \mathrm{int}\Big[\frac{1}{\alpha}\log\frac{n}{3}\Big].$$

As i moves away from L_n, the c_{ni} decay exponentially. If K is a convenient large positive integer, the i with $|i - L_n| > K$ can be ignored, by Lemma 4. As n increases, there are infinitely many disjoint segments $I_n = \{i : |i - L_n| \le K\}$. The ζ's in these segments are independent, and

$$P\Big\{\sum_{i\in I_n} c_{ni}\zeta_i^2 > C\log n\Big\}$$

is governed by the behavior for $i = L_n$ or $L_n + 1$. Finally, there will almost surely be arbitrarily large i with $\zeta_i^2 \sim \log i$.

(iii) There is another way to salvage Bernstein-von Mises. Suppose (1)–(2)–(3), with $\sigma_n^2 = 1/n$ and $\tau_i^2 = 1/i^2$. Instead of the ℓ_2-norm $\|\beta - \hat\beta\|^2$, consider $\sum_i \|\beta_i - \hat\beta_i\|^2/i^\gamma$. If $\gamma < 1/2$, previous results apply, but Bayesians and frequentists merge when $\gamma \ge 1/2$. Of course, Theorem 4 remains in force: the merging is only for a particular functional.

(iv) The a.e. consistency results—Theorem 5 and Remark (iii) to Theorem 1—depend on the behavior of σ_n^2. For slow rates of convergence to 0, a.e. consistency will depend on the joint distribution of the errors across n. A simple example may illustrate the point: suppose U_n is $N(0, \sigma_n^2)$. If $\sigma_n^2 = 1/n$, then $U_n \to 0$ a.e.—for any joint distribution. On the other hand, suppose $\sigma_1^2 = 2$ and $\sigma_n^2 = 1/\log n$ for $n > 1$. If the U_n are independent, convergence a.e. fails. If $U_n = V_1 + \cdots + V_n$, the V_i being independent $N(0, \tau_i^2)$ variables with $\tau_1^2 + \cdots + \tau_n^2 = \sigma_n^2$, convergence a.e. will hold.

3. Stochastic processes

The lead example in Cox (1993) is once-integrated Brownian motion on the unit interval, which is used as a prior on functions β in the model $y_i = \beta(i/n) + \epsilon_i$, the ϵ being IID $N(0,\sigma^2)$ variables. Eigenvalue expansions of the Karhunen-Loève type transform such problems into discrete problems. We could not find the eigenvalues of integrated Brownian motion in the literature, and give an informal account here—with many thanks to David Brillinger, who showed us all the interesting tricks. On the equivalence between white-noise problems and non-parametric regression or density estimation, see Brown and Low (1996) or Nussbaum (1996).

Let B_s be standard Brownian motion, so $E\{B_s\} = 0$ and $\mathrm{cov}(B_s, B_t) = \min(s, t)$. Once-integrated Brownian motion is $X_t = \int_0^t B_s\,ds$. Plainly, $E\{X_t\} = 0$. Furthermore,

$$\mathrm{var}\, X_t = 2\int_0^t\int_0^u E\{B_u B_v\}\,dv\,du = 2\int_0^t\int_0^u v\,dv\,du = \frac{1}{3}t^3.$$

Then for $0 \le s \le t$,

$$\mathrm{cov}(X_s, X_t) = \frac{1}{3}s^3 + \frac{1}{2}s^2(t - s).$$

In short, if $K(s, t) = \mathrm{cov}(X_s, X_t)$,

$$\begin{aligned} K(s,t) &= \frac{1}{2}s^2t - \frac{1}{6}s^3 \ \text{ for } \ 0 \le s \le t \\ &= \frac{1}{2}st^2 - \frac{1}{6}t^3 \ \text{ for } \ 0 \le t \le s. \end{aligned}$$

Let λ_i be the ith eigenvalue of once-integrated Brownian motion on [0,1], and $\phi_i(t)$ the corresponding eigenfunction. These eigenfunctions are orthonormal in $L^2[0,1]$,

$$\int_0^1 K(s,t)\phi_i(s)\,ds = \lambda_i\phi_i(t) \tag{24}$$

and

$$X_t = \sum_{i=1}^{\infty} \sqrt{\lambda_i} Z_i \phi_i(t),$$

the Z_i being IID N(0,1) variables. Analytically,

$$K(s,t) = \sum_{i=1}^{\infty} \lambda_i \phi_i(s)\phi_i(t).$$

For the existence of eigenfunctions and eigenvalues, and the expansion, see Riesz and Nagy (1955, Section 97).

Our objective is to solve equation (24). Dropping subscripts and rewriting, we get

$$\int_0^t (\frac{1}{2}s^2 t - \frac{1}{6}s^3)\phi(s)\,ds + \int_t^1 (\frac{1}{2}st^2 - \frac{1}{6}t^3)\phi(s)\,ds = \lambda\phi(t). \tag{25}$$

Successive differentiations with respect to t give

$$\int_0^t \frac{1}{2}s^2\phi(s)\,ds + \int_t^1 (st - \frac{1}{2}t^2)\phi(s)\,ds = \lambda\phi'(t) \tag{26}$$

$$\int_t^1 (s-t)\phi(s)\,ds = \lambda\phi''(t) \tag{27}$$

$$-\int_t^1 \phi(s)\,ds = \lambda\phi^{(3)}(t) \tag{28}$$

$$\phi(t) = \lambda\phi^{(4)}(t). \tag{29}$$

The boundary conditions—from (25)–(28)—are

$$\phi(0) = 0, \;\; \phi'(0) = 0, \;\; \phi''(1) = 0, \;\; \phi^{(3)}(1) = 0. \tag{30}$$

The solution to (29) is

$$\phi(t) = A\cos(t/\lambda^{1/4}) + B\sin(t/\lambda^{1/4}) + C\cosh(t/\lambda^{1/4}) + D\sinh(t/\lambda^{1/4}), \tag{31}$$

where the constants A, B, C, D are chosen to satisfy the boundary conditions (30). Since $\phi(0) = 0$, we have $C = -A$; and $\phi'(0) = 0$ entails $D = -B$. The remaining two boundary conditions lead to the following two equations, with $\theta = 1/\lambda^{1/4}$:

$$A(\cos\theta + \cosh\theta) + B(\sin\theta + \sinh\theta) = 0$$
$$A(-\sin\theta + \sinh\theta) + B(\cos\theta + \cosh\theta) = 0.$$

Solve each equation for A/B in terms of θ and equate the results, to see that

$$(\sinh\theta)^2 - (\sin\theta)^2 = (\cos\theta + \cosh\theta)^2,$$

that is,

$$\cos\theta \cosh\theta = -1. \tag{32}$$

Plainly, the roots θ_i of (32) tend to ∞. If θ_i is the ith root, then $\theta_i \doteq (2i-1)\pi/2$ for $i = 1, 2, \ldots$.

Theorem 7. The ith eigenvalue of once-integrated Brownian motion is $\lambda_i = 1/\theta_i^4$, where $\theta_i \approx \pi i$ is the ith root of the transcendental equation (32). The corresponding eigenfunction is given by (31), with

$$A = B, \quad C = -D, \quad A/B = -(\sin\theta + \sinh\theta)/(\cos\theta + \cosh\theta),$$

while B is chosen so the function has norm 1.

Remark. In short, $\tau_i^2 \sim 1/i^4$ in Theorems 1–3 corresponds to integrated Brownian motion. By a more direct calculation, $\tau_i^2 \sim 1/i^2$ corresponds to Brownian motion itself. In this case, of course, everything can be written down explicitly: the ith eigenvalue is $\lambda_i = 1/[(2i-1)\pi/2]^2$, and the corresponding eigenfunction is $\phi_i(t) = A\sin(t/\sqrt{\lambda_i})$.

4. The exceptional null set

We consider the structure of the exceptional null set in Theorem 2. To simplify matters, take

$$\sigma_i^2 = 1/n \quad \text{and} \quad \tau_i^2 = 1/i^2. \tag{33}$$

Recall that $\hat{\beta}_n$ is the Bayes estimate computed from the prior defined by (2). According to Theorem 2, for almost all β drawn from (2), $\|\hat{\beta}_n - \beta\|^2$ is of order $1/n^{1/2}$. There is a wobbly "Bayes bias" term $\sqrt{F_n}U_n(\beta)$, with $F_n \sim n^{-3/2}$. Being standard normal, $U_n(\beta) = O\big(\sqrt{\log n}\,\big)$ a.s. In short, the Bayes bias term is $O\big(\sqrt{\log n}/n^{3/4}\big)$ a.s.; finer results, of course, can be proved. There is also randomness of order $1/n^{3/4}$, due to the ϵ's. The next theorem shows that for a dense set of exceptional β's, $\|\hat{\beta}_n - \beta\|^2$ is of much larger order $1/n^\rho$, with randomness of order $1/\sqrt{n^{1+\rho}}$, where ρ is a small positive number at our disposition.

Theorem 8. Assume (1) and (33). Let $\rho < 1/2$ be any small positive number, and

$$C_0 = \int_0^\infty \frac{u^{3-2\rho}}{(1+u^2)^2}\,du \quad \text{and} \quad C_1 = 4\int_0^\infty \frac{u^{3-2\rho}}{(1+u^2)^4}\,du.$$

Let A be any large positive number and let r be any small positive number. For any $\beta^* \in \ell_2$, there is a parameter vector $\beta \in \ell_2$ with $\|\beta - \beta^*\| < r$ and

(a) $E_\beta\{\|\hat\beta_n - \beta\|^2\} \approx C_0 A/n^\rho$,

(b) $\text{var}_\beta\{\|\hat\beta_n - \beta\|^2\} \approx C_1 A/n^{1+\rho}$.

Sketch of proof. Let i_0 be a large positive integer. Let $\beta_i = \beta_i^*$ for $i < i_0$ while $\beta_i^2 = A/i^{1+2\rho}$ for $i \geq i_0$. By taking i_0 large, we get $\|\beta - \beta^*\|$ to be small. Next we use Proposition 1, with $w_{ni} = n/(n+i^2)$. Let $\hat\beta_{ni}$ be the ith coordinate of the Bayes estimate $\hat\beta_n$, so that

$$(\hat\beta_{ni} - \beta_i)^2 = (1-w_{ni})^2\beta_i^2 - 2(1-w_{ni})w_{ni}\beta_i\epsilon_i + w_{ni}^2\epsilon_i^2 \tag{34}$$

and

$$E_\beta\{\|\hat\beta_n - \beta\|^2\} = \sum_i (1-w_{ni})^2\beta_i^2 + \sum_i w_{ni}^2 E\{\epsilon_i^2\}. \tag{35}$$

Now $E_\beta\{\|\hat\beta_n - \beta\|^2\} = T_0 + T_1 + T_2$, where

$$T_0 = \sum_{i=1}^{i_0-1}(1-w_{ni})^2\beta_i^2 = \sum_{i=1}^{i_0-1}\frac{i^4}{(n+i^2)^2}\beta_i^{*2} = O\big(\frac{1}{n^2}\big), \tag{36}$$

$$T_1 = \sum_{i=i_0}^{\infty}(1-w_{ni})^2\beta_i^2 = \sum_{i=i_0}^{\infty}\frac{Ai^{3-2\rho}}{(n+i^2)^2} \approx \frac{C_0 A}{n^\rho}, \tag{37}$$

$$T_2 = \frac{1}{n}\sum_{i=1}^{\infty} w_{ni}^2 = \frac{1}{n}\sum_{i=1}^{\infty}\frac{n^2}{(n+i^2)^2} = O\big(\frac{1}{\sqrt{n}}\big). \tag{38}$$

This proves claim (a). For (b), there are two random terms in (34), with covariance 0. So

$$\begin{aligned}\text{var}_\beta\{\|\hat\beta_n - \beta\|^2\} &= \sum_{i=1}^{\infty} 4(1-w_{ni})^2 w_{ni}^2\beta_i^2\,\text{var}\,\epsilon_i + \sum_{i=1}^{\infty} w_{ni}^4\,\text{var}\,\epsilon_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{\infty} 4(1-w_{ni})^2 w_{ni}^2\beta_i^2 + \frac{2}{n^2}\sum_{i=1}^{\infty} w_{ni}^4.\end{aligned}$$

As before,

$$\frac{1}{n}\sum_{i=1}^{i_0-1} 4(1-w_{ni})^2 w_{ni}^2\beta_i^2 = O\big(\frac{1}{n^3}\big)$$

is negligible, while

$$\frac{1}{n}\sum_{i=i_0}^{\infty} 4(1-w_{ni})^2 w_{ni}^2 \beta_i^2 \;=\; 4An\sum_{i=i_0}^{\infty} \frac{i^{3-2\rho}}{(n+i^2)^4} \approx \frac{C_1 A}{n^{1+\rho}}$$

and

$$\frac{2}{n^2}\sum_{i=1}^{\infty} w_{ni}^4 = 2n^2 \sum_{i=1}^{\infty} \frac{1}{(n+i^2)^4} = O\Big(\frac{1}{n^{3/2}}\Big)$$

is negligible. We have not checked details, but asymptotic normality must follow. QED

The situation is more manageable if we ignore the variances and consider only

$$\phi_n(\beta) = E_\beta\{\|\hat\beta_n - \beta\|^2\}. \tag{39}$$

As shown in Theorem 5, $\phi_n(\beta) \to 0$ as $n \to \infty$ for any $\beta \in \ell_2$. However, the rate of convergence can be arbitrarily slow for most β's—if "most" is defined in a topological sense. That is the content of the next theorem. The setting is ℓ_2, which is a complete separable metric space. A G_δ is a countable intersection of open sets. If each of these open sets is dense, so is the intersection: that is the "Baire property." Dense G_δ's are the topological analogs of sets of measure 1, and are large "in the sense of category"; see Oxtoby (1980) for discussion.

Theorem 9. Assume (1) and (33). Let $0 < a_n \uparrow \infty$ be a sequence of real numbers that is strictly increasing to ∞, no matter how slowly. Then $\{\beta : \limsup_n a_n\phi_n(\beta) = \infty\}$ is a dense G_δ.

Proof. By (35), ϕ_n is continuous. So

$$F(N, M) = \bigcap_{n=N}^{\infty} \{\, a_n\phi_n \le M \,\}$$

is a closed set of β's. We will show each F to be nowhere dense. If so,

$$\{\,\beta : \limsup_{n\to\infty} a_n\phi_n(\beta) < \infty\,\} = \bigcup_{M=1}^{\infty}\bigcup_{N=1}^{\infty} F(N, M)$$

is a countable union of closed nowhere dense sets, and the argument is done.

To show $F(N, M)$ is nowhere dense, fix $\beta^* \in F(M, N)$. We will approximate β^* by β's with $\lim_n a_n\phi_n(\beta) > M$. Let

$$c_0 = \int_0^1 \frac{4u^3}{(1+u^2)^3}du \tag{40}$$

and $A = M/c_0$. Take $\beta_i = \beta_i^*$ for $i = 1, \ldots, i_0 - 1$ while $\beta_i^2 = A/a_i - A/a_{i+1}$ for all $i \ge i_0$. Abbreviate $c_{n,i} = (1 - w_{ni})^2$. By (35),

$$\phi_n(\beta) > \sum_{i=i_0}^{\infty}(1-w_{ni})^2\beta_i^2 = \sum_{i=i_0}^{\infty} c_{n,i}\beta_i^2 = A\frac{c_{n,i_0}}{a_{i_0}} + A\sum_{i=i_0}^{\infty}\frac{c_{n,i+1}-c_{n,i}}{a_{i+1}}.$$

Let $f(x) = x^4/(n+x^2)^2$, so $c_{n,i} = f(i)$. The dependence of f on n is suppressed in the notation. Now

$$f'(x) = 4nx^3/(n+x^2)^3 \quad \text{and} \quad f''(x) = 12nx^2(n-x^2)/(n+x^2)^4.$$

Thus, f' is increasing on $[0, \sqrt{n}]$ and $f(x+1) - f(x) > f'(x)$ for $0 \le x \le \sqrt{n} - 1$. As a result,

$$\phi_n(\beta) > A \sum_{i=i_0}^{s(n)} \frac{f'(i)}{a_{i+1}} > \frac{A}{a_n} \sum_{i=i_0}^{s(n)} f'(i),$$

where $s(n)$ is the greatest integer that does not exceed $\sqrt{n} - 1$. Whatever i_0 may be,

$$\sum_{i=i_0}^{s(n)} f'(i) \to c_0$$

and $\limsup_n a_n \phi_n(\beta) > c_0 A = M$, as required. QED

The next result characterizes the lim inf. For most β's in the sense of category, the mean squared error of the Bayes estimate—along a suitable subsequence of n's—is $c_1 n^{-1/2}$, where

$$c_1 = \int_0^\infty \frac{1}{(1+x^2)^2}\, dx. \tag{41}$$

Theorem 10. Assume (1) and (33). Define ϕ_n by (39) and c_1 by (41).

(a) $\liminf_n n^{1/2}\phi_n = c_1$ on a dense G_δ.

(b) $\liminf_n n^{1/2}\phi_n(\beta) \ge c_1$ for any $\beta \in \ell_2$.

Proof. Claim (a). The argument is essentially the same as for Theorem 9. Let

$$F(N, M) = \bigcap_{n=N}^{\infty} \{ n^{1/2}\phi_n \ge c_1 + \frac{1}{M} \}.$$

These are closed sets, and

$$\{ \beta : \liminf_{n\to\infty} n^{1/2}\phi_n(\beta) > c_1 \} = \bigcup_{M=1}^{\infty} \bigcup_{N=1}^{\infty} F(N, M).$$

To show that $F(N, M)$ is nowhere dense, we approximate $\beta^* \in F(M, N)$ by β's with

$$\lim_{n\to\infty} n^{1/2}\phi_n(\beta) = c_1.$$

Let $\beta_i = \beta_i^*$ for $i = 1, \ldots, i_0 - 1$ and $\beta_i = 0$ for $i \ge i_0$. We estimate $\phi_n(\beta)$ using equations (35–38). As before, $T_0 = O(1/n^2)$. But now, $T_1 = 0$ so the dominant term is $T_2 \approx n^{-1/2}c_1$. For claim (b), equations (35) and (38) give the lower bound, with all $\beta_i = 0$. QED

To summarize, for most β's in the sense of measure, the mean squared error of the Bayes estimate is $Cn^{-1/2}$ with $C = \int_0^\infty 1/(1+u^2)\,du$. For most β's in the sense of category, the rate (along certain subsequences of n's) is $c_1 n^{-1/2}$, where $c_1 < C$. Along other subsequences, the rate is much slower than $n^{-1/2}$ —as much slower as you please. There are general results on minimax rates of convergence and consistency of Bayes estimates in Zhao (1997) or Brown, Low, and Zhao (1998).

Acknowledgments

Persi Diaconis is a virtual coauthor, and we thank David Brillinger, Larry Brown, Iain Johnstone, Mark Low, Yuval Peres, and Linda Zhao for useful conversations. We also thank our editors and referees.

References

Brown, L. D., Liu, R. (1993). Nonexistence of informative unbiased estimators in singular problems. *Annals of Statistics* 21 1–13.

Brown, L. D., Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Annals of Statistics* 24 2384–98.

Brown, L. D., Low, M. G., and Zhao, L. H. (1998). Superefficiency in nonparametric function estimation. Technical report, Statistics Department, University of Pennsylvania.

Cox, D. (1993). An analysis of Bayesian inference for nonparametric regression. *Annals of Statistics* 21 903–923.

LeCam, Lucien M. and Yang, Grace Lo (1990). *Asymptotics in Statistics: Some Basic Concepts.* New York, Springer-Verlag.

Lehmann, E. (1991). *Theory of Point Estimation.* Pacific Grove, Calif., Wadsworth & Brooks/Cole.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society* 67 1–19.

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and white noise. *Annals of Statistics* 24 2399–2430.

Oxtoby, J. (1980). *Measure and Category.* 2nd ed., Springer.

Pfanzagl, J. (1998). On local uniformity for estimators and confidence limits. Technical report, Statistics Department, University of Cologne.

Prakasa Rao, B. L. S. (1987). *Asymptotic Theory of Statistical Inference.* New York, Wiley.

Riesz, F. and Nagy, B. Sz. (1955). *Functional Analysis.* Translated from the 2d French edition by L. F. Boron. New York, Ungar.

Zhao, L. H. (1997). Bayesian aspects of some nonparametric problems. Technical report, Statistics Department, University of Pennsylvania.

Williams, D. (1991). *Probability with Martingales.* Cambridge University Press.

Zhao, L. H. (1997). Bayesian aspects of some nonparametric problems. Technical report, Statistics Department, University of Pennsylvania.