12

Infinite-Dimensional Bernstein–von Mises Theorem

Although nonparametric theory is mostly concerned with rates of estimation, distributional approximations are possible for statistical procedures that focus on special aspects of a parameter. Within the Bayesian framework these may take the form of a normal approximation to the marginal posterior distribution of a functional of the parameter. We begin this chapter with a review of the classical result in this direction: the Bernstein–von Mises theorem for parametric models. Next we present a result in the same spirit for the posterior distribution based on the Dirichlet process prior, and corresponding strong approximations. After a brief introduction to semiparametric models, we proceed with Bernstein–von Mises theorems for functionals on such models, with strict semiparametric models as a special case. These are illustrated with applications to Gaussian process priors and the Cox proportional hazard model. A discussion of the Bernstein–von Mises theorem in the context of the white noise model illustrates the possibilities and difficulties of extending the theorem to the fully infinite-dimensional setting.

12.1 Introduction

In many statistical experiments with a Euclidean parameter space, good estimators $\hat{\theta}_n$ are asymptotically normally distributed with mean the parameter θ and covariance matrix proportional to the inverse Fisher information. More formally, for r_n^{-1} a rate of convergence, the sequence $r_n(\hat{\theta}_n - \theta)$ tends in distribution to a $\mathrm{Nor}(0, I_\theta^{-1})$-distribution. Under regularity conditions this is true for the maximum likelihood estimator and most Bayes estimators; it is usually proved by approximating the estimator $\hat{\theta}_n$ by an average and next applying the central limit theorem. The Fisher-Cramér-Rao–Le Cam theory designates these estimators as asymptotically efficient, the inverse Fisher information being the minimal attainable asymptotic variance in the local minimax sense.

For a frequentist the randomness in $r_n(\hat{\theta}_n - \theta)$ comes through the observation $X^{(n)}$, hidden in $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$, which is considered drawn from a distribution $P_\theta^{(n)}$ indexed by a fixed parameter θ. To a Bayesian the quantity $r_n(\hat{\theta}_n - \theta)$ has a quite different interpretation. In the Bayesian setup θ is the variable and the randomness is evaluated according to the posterior distribution $\Pi_n(\cdot|\, X^{(n)})$, given a fixed observation $X^{(n)}$. It is remarkable that these two assessments of randomness, with diametrically opposite interpretations, are approximately equal in large samples: by the *Bernstein–von Mises theorem* the posterior distribution of $r_n(\hat{\theta}_n - \theta)$ also tends to a $\mathrm{Nor}(0, I_\theta^{-1})$-distribution, for most observations $X^{(n)}$.

The symmetry between the two statements may be nicely exposed by writing them both in the Bayesian framework, where both θ and $X^{(n)}$ are random:

$$\begin{aligned} r_n(\hat{\theta}_n-\theta)\,|\,\theta=\theta_0 &\rightsquigarrow \mathrm{Nor}_d(0, I_{\theta_0}^{-1}),\\ r_n(\theta-\hat{\theta}_n)\,|\,X^{(n)} &\rightsquigarrow \mathrm{Nor}_d(0, I_{\theta_0}^{-1}). \end{aligned}$$

The first line gives the frequentist statement, which conditions on a given value of the parameter, whereas the left side of the second line refers to the posterior distribution of θ, centered at $\hat{\theta}_n$ and scaled by r_n. In the first case the distribution of $\hat{\theta}_n=\hat{\theta}_n(X^{(n)})$ is evaluated for $X^{(n)}$ following the parameter value θ_0, whereas in the second case $\hat{\theta}_n$ is fixed by conditioning on $X^{(n)}$, but the distribution of the random (posterior) measure on the left side is again evaluated for $X^{(n)}$ following the parameter θ_0.

Besides having conceptual appeal, the Bernstein–von Mises theorem is of great importance for the interpretation of Bayesian credible regions: sets in the parameter space of given posterior probability. From the normal approximation it can be seen that a region of highest posterior probability density will be asymptotically centered at $\hat{\theta}_n$ and spread in such a way that it is equivalent to a frequentist confidence region based on $\hat{\theta}_n$. Consequently, the frequentist confidence level of a Bayesian credible set will be approximately equal to its credibility. This provides a frequentist justification of Bayesian credible regions.

The Bernstein–von Mises theorem extends beyond the setting of i.i.d. observations, to models that are not regular in the parameter (such as uniform distributions) and lead to non-normal limit experiments, and also to parameters of increasing dimension. The mode of approximation by a normal distribution can also be made precise in various ways (e.g. using Kullback-Leibler divergence or higher order approximations). For reference, we state only a version for i.i.d. observations that requires minimal regularity of the model and employs a testing condition similar to the ones used in posterior consistency and convergence rate theorems.

A set $\{p_\theta: \theta\in\Theta\}$ of probability densities with respect to a σ-finite dominating measure ν on a measurable space $(\mathfrak{X}, \mathscr{A})$ indexed by an open subset $\Theta\subset\mathbb{R}^d$ is said to be *differentiable in quadratic mean* at θ if there exists a measurable map $\dot{\ell}_\theta: \mathfrak{X}\to\mathbb{R}$ such that, as $\|h\|\to 0$,

$$\int\Big[\sqrt{p_{\theta+h}}-\sqrt{p_\theta}-\frac{1}{2}h^\top\dot{\ell}_\theta\sqrt{p_\theta}\Big]^2\,d\nu=o(\|h\|^2). \tag{12.1}$$

The function $\dot{\ell}_\theta$ is a version of the *score function* of the model, and its covariance matrix $I_\theta = P_\theta(\dot{\ell}_\theta\dot{\ell}_\theta^\top)$ is the *Fisher information matrix*. For simplicity we assume that this is nonsingular. For $X_1, X_2, \ldots \overset{\text{iid}}{\sim} p_\theta$, set

$$\Delta_{n,\theta}=\frac{1}{\sqrt{n}}\sum_{i=1}^n I_\theta^{-1}\dot{\ell}_\theta(X_i).$$

This sequence is well defined and asymptotically normally distributed as soon as the model is differentiable in quadratic mean.

Theorem 12.1 (Bernstein–von Mises–Le Cam) *If the model $(p_\theta: \theta\in\Theta)$ is differentiable in quadratic mean at θ_0 with nonsingular Fisher information matrix I_{θ_0}, and for any $\epsilon>0$ there exists a sequence of tests ϕ_n based on $(X_1,\ldots,X_n)$ such that*

$$P^n_{\theta_0}\phi_n \to 0, \qquad \sup_{\theta:\|\theta-\theta_0\|\geq\epsilon} P^n_\theta(1-\phi_n) \to 0,$$

then for any prior distribution on θ that possesses a density with respect to the Lebesgue measure in a neighborhood of θ_0 that is bounded away from zero,

$$\mathrm{E}_{\theta_0}\Bigl\|\Pi_n(\theta\colon \sqrt{n}(\theta-\hat\theta_n)\in\cdot\,|\,X_1,\ldots,X_n) - \mathrm{Nor}_d(\Delta_{n,\theta_0}, I^{-1}_{\theta_0})\Bigr\|_{TV} \to 0.$$

For a proof, see van der Vaart (1998), pages 140–143. To translate back to the preceding we note that under regularity conditions the scaled and centered maximum likelihood estimators $\sqrt{n}(\hat\theta_n-\theta)$ will be asymptotically equivalent to the sequence Δ_{n,θ_0} when the observations are independently sampled from P_{θ_0}. Because the total variation norm is invariant under a shift of location and a change of scale, the sequence Δ_{n,θ_0} may be replaced by $\sqrt{n}(\hat\theta_n-\theta_0)$ in that case, and the centering shifted by $\sqrt{n}\theta_0$ and the scale by $\sqrt{n}$, leading to

$$\mathrm{E}_{\theta_0}\Bigl\|\Pi_n(\theta\colon \theta\in\cdot\,|\,X_1,\ldots,X_n) - \mathrm{Nor}_d(\hat\theta_n, (nI_{\theta_0})^{-1})\Bigr\|_{TV} \to 0.$$

This is equivalent to the assertion in the theorem, except for the fact that the good behavior of the maximum likelihood estimator requires additional regularity conditions.

In most of the nonparametric examples in this book, a Bernstein–von Mises theorem is not valid in the same way. This appears to be due at least partly to the fact that a bias-variance trade-off is at the core of these examples, whereas in the (parametric) Bernstein–von Mises theorem the bias is negligible. It has thus been claimed in the literature that the “infinite-dimensional Bernstein–von Mises theorem does not hold.” However, for smoother aspects of the parameter the theorem does hold, and one may say that the theorem holds provided the topology on the parameter space is chosen appropriately (and the prior does not unnecessarily introduce a bias). In the next sections we illustrate this by a Bernstein–von Mises theorem for estimating a measure using the Dirichlet prior, and Bernstein–von Mises theorems for smooth functionals on semiparametric models.

12.2 Dirichlet Process

By Theorem 4.6 the posterior distribution of P in the model

$$P\sim\mathrm{DP}(\alpha), \qquad X_1,\ldots,X_n\,|\,P \overset{\text{iid}}{\sim} P,$$

is the Dirichlet process with base measure $\alpha+n\mathbb{P}_n$, for $\mathbb{P}_n$ the empirical measure of $X_1,\ldots,X_n$. We shall show that this satisfies a Bernstein–von Mises theorem.

The empirical measure $\mathbb{P}_n$ is the (nonparametric) maximum likelihood estimator in this problem, which suggests to use it as the centering measure, which will then concern the conditional distribution of the process $\sqrt{n}(P-\mathbb{P}_n)$, for $P\,|\,X_1,X_2,\ldots\sim\mathrm{DP}(\alpha+\sum_{i=1}^n\delta_{X_i})$. The posterior mean in this problem is $(\alpha+n\mathbb{P}_n)/(|\alpha|+n)$, and differs from the empirical measure $\mathbb{P}_n$ only by $(|\alpha|/(|\alpha|+n))(\bar\alpha-\mathbb{P}_n)$, which is of order $1/n$, if $n\to\infty$ and α remains fixed. This difference is negligible even after scaling by $\sqrt{n}$, whence in the Bernstein–von Mises theorem the posterior distribution can equivalently be centered at its mean.

For a fixed measurable set A the posterior distribution of $P(A)$ is the beta distribution $\mathrm{Be}(\alpha(A) + n\mathbb{P}_n(A), \alpha(A^c) + n\mathbb{P}_n(A^c))$. This is the same posterior distribution as in the reduced model where we observe a sample of indicators $\mathbb{1}\{X_1 \in A\}, \ldots, \mathbb{1}\{X_n \in A\}$ from the Bernoulli distribution $\mathrm{Bin}(1, P(A))$, with a $\mathrm{Be}(\alpha(A), \alpha(A^c))$ prior distribution on the success probability. The inverse Fisher information for a binomial proportion p is equal to $p(1-p)$, while the maximum likelihood estimator is the sample proportion $\mathbb{P}_n(A)$. Hence the posterior distribution of $\sqrt{n}(P(A) - \mathbb{P}_n(A))$ is asymptotically equivalent to a $\mathrm{Nor}(0, P(A)(1 - P(A)))$-distribution, by the parametric Bernstein–von Mises theorem. Alternatively, this may be verified directly from the beta-distribution: a weak approximation to the normal distribution is immediate from the central limit theorem and the delta-method (see Problem 12.1), whereas an approximation in the total variation norm needs some work.

The argument extends to the case of multiple measurable sets $A_1, \ldots, A_k$: the posterior distribution of the vectors $\sqrt{n}(P(A_1) - \mathbb{P}_n(A_1), \ldots, P(A_k) - \mathbb{P}_n(A_k))$ tends to a multivariate normal distribution $\mathrm{Nor}_k(0, \Sigma)$, where Σ is the matrix with elements $P(A_i \cap A_j) - P(A_i)P(A_j)$.

This normal limit distribution is identical (of course) to the limit distribution of the scaled maximum likelihood estimator $\sqrt{n}(\mathbb{P}_n(A_1) - P(A_1), \ldots, \mathbb{P}_n(A_k) - P(A_k))$, the *empirical process* at the sets $A_1, \ldots, A_k$. The empirical process $\sqrt{n}(\mathbb{P}_n - P)$ has been studied in detail, and is known to converge also in stronger ways than in terms of its marginal distributions. In particular, it can be viewed as a map $f \mapsto \sqrt{n}(\mathbb{P}_n f - Pf)$ that attaches to every integrable function f the expected value $n^{-1/2}\sum_{i=1}^n (f(X_i) - Pf)$. A collection of functions $\mathcal{F} \subset \mathbb{L}_2(P)$ is said to be *Donsker* if this map is bounded in $f \in \mathcal{F}$ and converges in distribution to a tight limit in the space $\mathfrak{L}_\infty(\mathcal{F})$ of bounded functions $z\colon \mathcal{F} \to \mathbb{R}$, equipped with the supremum norm $\|z\| = \sup\{|z(f)|\colon f \in \mathcal{F}\}$. (See Appendix F.) The limit process $\mathbb{G}$ is the Gaussian process with mean zero and covariance function $\mathrm{cov}\,(\mathbb{G}(f), \mathbb{G}(g)) = P[fg] - P[f]\,P[g]$, for $f, g \in \mathcal{F}$, known as the *P-Brownian bridge*.

The following theorem shows that the Bernstein–von Mises theorem for the Dirichlet posterior is also valid in this stronger, uniform sense. (For the most general version it is necessary to understand the convergence in distribution in terms of outer expectations and to be precise about the definition of the posterior process. The following theorem refers to the version outlined in the proof, and the conditional convergence is understood in terms of the bounded Lipschitz metric, as in the reference given. If there exists a Borel measurable version of the process, then the theorem applies in particular to this version.)

Theorem 12.2 (Bernstein–von Mises theorem for Dirichlet process) *For any P_0-Donsker class $\mathcal{F}$ of functions with envelope function F such that $(P_0 + \alpha)^*[F^2] < \infty$, the process $\sqrt{n}(P - \mathbb{P}_n)$ with $P \sim \mathrm{DP}(\alpha + n\mathbb{P}_n)$ converges conditionally in distribution given $X_1, X_2, \ldots$ in $\mathfrak{L}_\infty(\mathcal{F})$ to a Brownian bridge process a.s. $[P_0^\infty]$, as $n \to \infty$. The same conclusion is valid if the centering $\mathbb{P}_n$ is replaced by the posterior mean $\tilde{\mathbb{P}}_n = (\alpha + n\mathbb{P}_n)/(|\alpha| + n)$.*

Proof By Proposition G.10, the $\mathrm{DP}(\alpha + n\mathbb{P}_n)$-distribution can be represented as $V_n Q + (1 - V_n)\mathbb{B}_n$, where the variables $Q \sim \mathrm{DP}(\alpha)$, $\mathbb{B}_n \sim \mathrm{DP}(n\mathbb{P}_n)$ and $V_n \sim \mathrm{Be}(|\alpha|, n)$ are independent. Assume further that these three variables are defined on a product probability space, with their independence expressed by being functions of separate coordinates in the product, and that $\mathbb{B}_n$ is defined as $\mathbb{B}_n f = \sum_{i=1}^n W_{n,i} f(X_i)$, where $X_1, X_2, \ldots$ are coordinate projections on further factors of the product probability space, and $(W_{n,1}, \ldots, W_{n,n})$ is a $\mathrm{Dir}(n; 1, \ldots, 1)$-vector independent of the other variables and defined on yet one more factor of the underlying product probability space. By Theorem 3.6.13 in van der Vaart and Wellner (1996), it then follows that $\sqrt{n}(\mathbb{B}_n - \mathbb{P}_n)$ tends conditionally given $X_1, X_2, \ldots$ in distribution to a Brownian bridge process, a.s. $[P_0^\infty]$.

Since $\sqrt{n} V_n \to 0$ in probability, and $f \mapsto Qf$ is a well defined element of $\mathfrak{L}_\infty(\mathcal{F})$, the process $\sqrt{n} V_n Q$ tends to zero in the latter space. By Slutsky's lemma the sum $\sqrt{n} V_n Q + \sqrt{n}(\mathbb{B}_n - \mathbb{P}_n)$ has the same limit as the second term. □

Example 12.3 (Cumulative distribution function) The class $\mathcal{F}$ of functions consisting of the set of indicator functions of cells $(-\infty, t]$ in Euclidean space is Donsker for any underlying measure. The measure generated by it can be identified with the corresponding cumulative distribution function $F(t) = P(-\infty, t]$. Thus the process $t \mapsto \sqrt{n}(F - \mathbb{F}_n)(t)$ converges conditionally in distribution given $X_1, X_2, \ldots$ to a Brownian bridge process, where $\mathbb{F}_n$ is the *empirical distribution function* of $X_1, \ldots, X_n$.

The preceding Bernstein–von Mises theorem may be combined with the continuous mapping theorem or the delta-method to obtain further consequences. The delta-method for random processes is treated in Section 3.9.3 of van der Vaart and Wellner (1996). Here we only note two consequences based on the continuous mapping theorem. If $\psi\colon \mathfrak{L}_\infty(\mathcal{F}) \to \mathbb{R}$ is a continuous map, then this theorem shows that $\psi(\sqrt{n}(P - \mathbb{P}_n)) \rightsquigarrow \psi(W)$, for W the limiting Brownian bridge. If the distribution of $\psi(W)$ can be evaluated analytically, then the asymptotic weak limit of $\psi(\sqrt{n}(P - \mathbb{P}_n))$ is obtained, which may be useful in constructing approximate credible sets. The following corollary describes two such occasions, involving one-sided and two-sided Kolmogorov-Smirnov distances, whose limiting distributions, as distributions of the maximum of Brownian bridges and its absolute value process, respectively, are well known in the literature.

Corollary 12.4 *If* $P \sim \mathrm{DP}(\alpha)$ *on the sample space* $\mathbb{R}$, *and* $X_1, \ldots, X_n \,|\, P \overset{iid}{\sim} P$, *then for any* $\lambda > 0$, *a.s.* $[P_0^\infty]$ *as* $n \to \infty$,

$$\text{(i)}\quad \mathrm{P}\Big(\sqrt{n} \sup_{x \in \mathbb{R}} |F(x) - \tilde{\mathbb{F}}_n(x)| > \lambda \,|\, X_1, \ldots, X_n\Big) \to 2 \sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2\lambda^2};$$

$$\text{(ii)}\quad \mathrm{P}\Big(\sqrt{n} \sup_{x \in \mathbb{R}} (F(x) - \tilde{\mathbb{F}}_n(x)) > \lambda \,|\, X_1, \ldots, X_n\Big) \to e^{-2\lambda^2}.$$

12.2.1 Strong Approximation

The weak convergence $X_n \rightsquigarrow X$ of a sequence of random variables involves the distributions of these variables only, and not the underlying probability space(s) on which they are defined. By the *almost sure representation theorem* there always exist variables X_n^* and X^* with identical distributions that are defined on a single probability space and are such that

$X_n^* \to X^*$, almost surely.[1] As almost sure convergence is a stronger and simpler property than convergence in distribution, this result is sometimes helpful. Strong approximations can be viewed as strengthening this result, by giving a rate to the almost sure convergence.

A *strong approximation* at rate ϵ_n of a given sequence of weakly converging variables $X_n \rightsquigarrow X$ is a pair of two sequences X_n^* and $\tilde{X}_n$ of random elements, all defined on a single underlying probability space, such that $X_n^* =_d X_n$, $\tilde{X}_n =_d X$, for all $n \in \mathbb{N}$, and $d(X_n^*, \tilde{X}_n) = O(\epsilon_n)$ a.s.

Strong approximations are useful for deriving approximations to transformed variables $\psi_n(X_n)$, where the function ψ_n depends on n. In particular, if the functions ψ_n satisfy a Lipschitz condition with suitable Lipschitz constants L_n which can be related to the rate of approximation ϵ_n, then it may be possible to analyze the distribution of $\psi_n(X_n)$ through its strong approximation $\psi_n(X_n^*)$, which is equal in distribution. For instance, this technique is useful for approximating the distribution of a kernel smoother.

A classical strong approximation is the *KMT construction* (after Komlós, Major and Tusnády) to the empirical process. The empirical process $\sqrt{n}(\mathbb{F}_n - F)$ of a random sample of n variables from the cumulative distribution function F on the real line tends in distribution to a Brownian bridge process $B \circ F$, for B a standard Brownian bridge on $[0, 1]$ (a mean zero Gaussian process with covariance kernel $\mathrm{E}[B_s B_t] = s \wedge t - st$). The KMT construction provides a random sample of variables and a sequence of Brownian bridges $B_n =_d B$ on a suitable probability space such that

$$\|\sqrt{n}(\mathbb{F}_n - F) - B_n \circ F\|_\infty = O(n^{-1/2}(\log n)^2), \qquad \text{a.s.}$$

Thus the distance between the empirical process and its “limit” is nearly $n^{-1/2}$. It is known that this rate cannot be improved.

The Brownian bridges B_n can also be related amongst themselves, by tying them to a *Kiefer process*. This is a mean-zero Gaussian process $\mathbb{K} = \{\mathbb{K}(s,t): (s,t) \in [0,1] \times [0,\infty)\}$ with continuous sample paths and covariance kernel $\mathrm{E}[\mathbb{K}(s_1,t_1)\mathbb{K}(s_2,t_2)] = [s_1 \wedge s_2 - s_1 s_2](t_1 \wedge t_2)$. The process $s \mapsto t^{-1/2}\mathbb{K}(s,t)$ is a Brownian bridge, for every $t > 0$, and the Brownian bridges in the KMT theorem can be taken as $B_n = n^{-1/2}\mathbb{K}(\cdot, n)$, for every n, and a suitable Kiefer proces.

The following theorem gives an analogous result for the posterior process corresponding to a sample from the Dirichlet process.

Theorem 12.5 *On a suitable probability space there exist random elements F and $\mathbb{K}$ and $X_1, X_2, \ldots \overset{iid}{\sim} F_0$ such that $F \mid X_1, \ldots, X_n \sim \mathrm{DP}(\alpha + n\sum_{i=1}^n \delta_{X_i})$ for every n, and $\mathbb{K}$ is a Kiefer process independent of $X_1, X_2, \ldots$, and such that*

$$\sup_{x \in \mathbb{R}} \Big| \sqrt{n}(F - \mathbb{F}_n)(x) - \frac{\mathbb{K}(F_0(x), n)}{n^{1/2}} \Big| = O\Big(\frac{(\log n)^{1/2}(\log\log n)^{1/4}}{n^{1/4}}\Big), \qquad \text{a.s.}$$

The theorem remains true if the empirical distribution function $\mathbb{F}_n$ is replaced by the posterior mean $(\alpha + n\mathbb{F}_n)/(|\alpha| + n)$. Furthermore, the choice $\alpha = 0$, for which F follows the Bayesian bootstrap, is allowed. The strong approximation rate is close to $n^{-1/4}$, much

[1] See e.g. van der Vaart and Wellner (1996), Theorem 1.10.3, or 1.10.4 for a stronger version assuming less measurability.

weaker than the rate in the KMT construction for the empirical distribution function. For a proof, which is long as for all strong approximation results, see Lo (1987).

The theorem can be applied to obtain the limiting distribution of a smoothed Dirichlet posterior process. For a given kernel w and $F \sim \mathrm{DP}(\alpha + n\mathbb{F}_n)$, set

$$f_n(x) = \int \frac{1}{h_n} w\Big(\frac{x-\theta}{h_n}\Big)\, dF(\theta), \qquad \hat{f}_n(x) = \frac{1}{nh_n}\sum_{i=1}^n w\Big(\frac{x-X_i}{h_n}\Big).$$

The function $\hat{f}_n$ is the usual kernel density estimator, and serves as the centering in the following theorem. Assume that the kernel w integrates to 1, is symmetric about 0, and absolutely continuous on the convex hull of its support, with $\|w'\|_2 < \infty$ and $\int z^2 w(z)\, dz < \infty$ and $\int_3^\infty z^{3/2}(\log\log z)^{1/2}(|w'(z)| + |w(z)|)\, dz < \infty$.

Theorem 12.6 *If $X_1, X_2, \ldots \overset{iid}{\sim} F_0$ for a distribution F_0 with strictly positive, twice differentiable density f_0 such that f_0, $f_0'/f_0^{1/2}$ and f_0'' are bounded, and $F|\, X_1, \ldots, X_n \sim \mathrm{DP}(\alpha + n\sum_{i=1}^n \delta_{X_i})$ for every n, then, for $h_n = n^{-\delta}$ for some $0 < \delta < \frac{1}{2}$ and $n \to \infty$,*

$$\mathrm{P}\bigg(\sqrt{2\log h_n^{-1}}\Big[\sup_{x\in\mathbb{R}} \frac{|f_n(x) - \hat{f}_n(x)|}{h_n f_0(x)^{1/2}\|w\|_2} - a_n\Big] \le t|\, X_1, X_2, \ldots\bigg) \to e^{-2e^{-t}}, \qquad \text{a.s.},$$

where $a_n = \sqrt{2\log h_n^{-1}} + 2\log\left(\|w'\|_2/\|w\|_2\right)) - \log\pi$.

The choice $\alpha = 0$, leading to the *smoothed Bayesian bootstrap process*, is allowed. Furthermore, the true density f_0 in the denominator may be replaced by the kernel estimator $\hat{f}_n$. The proof can be based on Theorem 12.5 and convergence to the extreme value distribution of the maximum of stationary Gaussian processes.

The quantile process of an absolutely continuous distribution also admits strong approximation by Kiefer processes. Under the conditions that the true density f_0 is differentiable and positive on its domain (a, b), and that the function $F_0(1 - F_0)|f_0'|/f_0^2$ is bounded, there exists a Kiefer process $\mathbb{K}$ and observations $X_1, X_2, \ldots$ on a suitable probability space such that, with $\delta_n = 25n^{-1}\log\log n$,

$$\sup_{\delta_n \le u \le 1-\delta_n} \Big|\sqrt{n} f_0 \circ F_0^{-1}(\mathbb{F}_n^{-1} - F_0^{-1})(u) - \frac{\mathbb{K}(u, n)}{\sqrt{n}}\Big| = O\Big(\frac{(\log n)^{1/2}(\log\log n)^{1/4}}{n^{1/4}}\Big), \text{ a.s.}$$

The restriction to the interval $[\delta_n, 1 - \delta_n]$ removes the left and right tails of the quantile process, where its population counterpart can be large. See Csörgő and Révész (1981), Theorem 6 for a proof.

There is a similar strong approximation to the Dirichlet posterior and the Bayesian bootstrap quantile functions.

Theorem 12.7 *Assume that the true density f_0 is differentiable and nonzero in its domain (a, b), and that the function $F_0(1 - F_0)|f_0'|/f_0^2$ is bounded. Then on a suitable probability space there exist random elements F and $\mathbb{K}$ and $X_1, X_2, \ldots \overset{iid}{\sim} F_0$ such that*

$F\,|\,X_1,\ldots,X_n \sim \mathrm{DP}(\alpha + n\sum_{i=1}^n \delta_{X_i})$ for every n, and $\mathbb{K}$ is a Kiefer process independent of $X_1, X_2, \ldots$, and such that

$$\sup_{\delta_n \le u \le 1-\delta_n} \Big| f_0 \circ F_0^{-1}\sqrt{n}(F^{-1} - \mathbb{F}_n^{-1})(u) - \frac{\mathbb{K}(u,n)}{\sqrt{n}}\Big| = O\Big(\frac{(\log n)^{1/2}(\log\log n)^{1/4}}{n^{1/4}}\Big), \text{ a.s.}$$

The proof of the result is based on the same techniques used to prove the corresponding result for the empirical quantile process and using Theorem 12.5 instead of the KMT theorem; see Gu and Ghosal (2008). For an application to approximating the Bayesian bootstrap distribution of the Receiver Operating Characteristic function, see Problem 12.4.

12.3 Semiparametric Models

A *semiparametric model* in the narrow sense is a set of densities $p_{\theta,\eta}$ parameterized by a pair of a Euclidean parameter θ and an infinite-dimensional *nuisance parameter* η, and the problem of most interest is to estimate θ. The parameterization is typically smooth in θ, and in case the nuisance parameter were known, the difficulty of estimating θ given a random sample of observations would be measured by the *score function* for θ:

$$\dot\ell_{\theta,\eta}(x) = \frac{\partial}{\partial\theta}\log p_{\theta,\eta}(x).$$

In particular, the covariance matrix of this function is the *Fisher information* matrix, whose inverse is a bound on the smallest asymptotic variance attainable by a (regular) estimator. When the nuisance parameter is unknown, estimating θ is a harder problem. The increased difficulty can be measured through the part of the score function for θ that can also be explained through *nuisance scores*. These are defined as the score functions

$$B_{\theta,\eta}b(x) := \frac{\partial}{\partial t}\log p_{\theta,\eta_t}(x)|_{t=0} \tag{12.2}$$

of suitable one-dimensional submodels $t \mapsto p_{\theta,\eta_t}$ with $\eta_0 = \eta$. If η ranges over an infinite-dimensional set, then there typically exist infinitely many submodels along which the derivative as in the display exists, and they can often be naturally identified by “directions” b in which η_t approaches η. The expression $B_{\theta,\eta}b$ on the left side of the display may be considered a notation only, but it was chosen to suggest an operator $B_{\theta,\eta}$ working on a direction b in which η_t approaches η. In many examples this *score operator* can be identified with a derivative of the map $\eta \mapsto \log p_{\theta,\eta}$. The linear span of $\dot\ell_{\theta,\eta}$ and all nuisance scores $B_{\theta,\eta}b$ is known as the *tangent space* of the model at the parameter (θ,η) (or at the density $p_{\theta,\eta}$).

For $\mathrm{Proj}_{\theta,\eta}$ the orthogonal projection in $\mathbb{L}_2(p_{\theta,\eta})$ onto the closed, linear span of all score functions $B_{\theta,\eta}b$ for the nuisance parameters, the *efficient score function* for θ is defined as

$$\tilde\ell_{\theta,\eta}(x) = \dot\ell_{\theta,\eta}(x) - \mathrm{Proj}_{\theta,\eta}\,\dot\ell_{\theta,\eta}(x).$$

The covariance matrix $\tilde I_{\theta,\eta} = P_{\theta,\eta}\tilde\ell_{\theta,\eta}\tilde\ell_{\theta,\eta}^\top$ of the efficient score function is known as the *efficient information* matrix. Being the covariance matrix of a projection, it is smaller than the ordinary information matrix, and hence possesses a larger inverse. This inverse can be shown to give a lower bound for estimators of θ in the situation that η is unknown. In fact, an estimator sequence $\hat\theta_n$ is considered to be asymptotically efficient at (θ, η) in the situation that η is unknown if

$$\sqrt{n}(\hat\theta_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde I_{\theta,\eta}^{-1} \tilde\ell_{\theta,\eta}(X_i) + o_{P_{\theta,\eta}}(1). \tag{12.3}$$

In many situations such efficient estimators have been constructed using estimating equations or variants of maximum likelihood estimators (see e.g. van der Vaart 1998, Chapter 25, for an overview). Posterior means for suitable priors could also play this role, as will be seen in the following.

As the efficient score function plays the same role in semiparametric models as the score function in ordinary parametric models, it is reasonable to expect that the Bernstein–von Mises theorem extends to semiparametric models, with ordinary score and information replaced by efficient score and information. In the semiparametric setting we equip both parameters θ and η with priors and given observations $X_1, \ldots, X_n$ form a posterior distribution for the joint parameter (θ, η) as usual. Since interest is in θ, we study the marginal posterior distribution induced on this parameter, written as $\Pi_n(\theta \in \cdot \,|\, X_1, \ldots, X_n)$. A semiparametric Bernstein–von Mises theorem should now assert that

$$\mathrm{E}_{\theta_0,\eta_0} \Big\| \Pi_n(\theta \in \cdot \,|\, X_1, \ldots, X_n) - \mathrm{Nor}(\hat\theta_n, n^{-1} \tilde I_{\theta_0,\eta_0}^{-1}) \Big\|_{TV} \to 0,$$

where $\hat\theta_n$ satisfies (12.3) at $(\theta, \eta) = (\theta_0, \eta)$. We derive sufficient conditions for this assertion below. A main interest is to identify the properties of priors on the nuisance parameter that lead to this property. In analogy with the parametric situation the prior on θ should wash out as $n \to \infty$, as long as it has a positive and continuous density in a neighborhood of the true value θ_0.

If the scores for the nuisance parameter are given as the range of an operator $B_{\theta,\eta}$, and this range is closed, then the projection operator can be found as $\mathrm{Proj}_{\theta,\eta} = B_{\theta,\eta}(B_{\theta,\eta}^\top B_{\theta,\eta})^{-1} B_{\theta,\eta}^\top$, for $B_{\theta,\eta}^\tau$ the adjoint of $B_{\theta,\eta}$. The projection of the score function $\dot\ell_{\theta,\eta}$ onto the linear span of the nuisance scores then takes the form $B_{\theta,\eta}\tilde b_{\theta,\eta}$, for

$$\tilde b_{\theta,\eta} = (B_{\theta,\eta}^\top B_{\theta,\eta})^{-1} B_{\theta,\eta}^\top \dot\ell_{\theta,\eta}.$$

This is known as the *least favorable direction*, and a corresponding submodel $t \mapsto \eta_t$ as a least-favorable submodel, because the submodel $t \mapsto p_{\theta+t,\eta_t}$ has the smallest information about t (at $t = 0$). In many situations the *information operator* $B_{\theta,\eta}^\top B_{\theta,\eta}$ is not invertible, and the preceding formulas are invalid. However, the projection of the θ-score onto the closed, linear span of the nuisance space always exists, and can be approximated by (linear combinations of) scores for the nuisance parameters.

The phrase "semiparametric estimation" is also attached more generally to estimating a Euclidean-valued functional on an infinite-dimensional model. This invites to consider the more general setup of densities p_η indexed by a single parameter η and a Euclidean parameter of interest $\chi(\eta)$. The models give rise to score functions $B_\eta b$ defined as in (12.2), but with θ removed throughout. The parameter χ is said to be *differentiable* at η if there exists a function $x \mapsto \tilde\kappa_\eta(x)$ such that, for every submodel $t \mapsto \eta_t$ as in (12.2) (which approaches $\eta = \eta_0$ from the "direction" b),

$$\chi(\eta_t) = \chi(\eta) + t P_\eta[\tilde{\kappa}_\eta\,(B_\eta b)] + o(t). \tag{12.4}$$

If this is true, then the quantity $P_\eta[\tilde{\kappa}_\eta\,(B_\eta b)]$ on the right side is the ordinary derivative of the map $t \mapsto \chi(\eta_t)$ at $t = 0$, but it is assumed that this derivative can be written as an inner product between the function $\tilde{\kappa}_\eta$ and the score function $B_\eta b$. (This representation must hold for efficient estimators for $\chi(\eta)$ to exist; see van der Vaart 1991.) The function $\tilde{\kappa}_\eta$ is unique only up to projection onto the closure of the tangent space; the unique projection is known as the *efficient influence function*. Efficient estimators $\hat{\chi}_n$ for $\chi(\eta)$ should satisfy

$$\sqrt{n}(\hat{\chi}_n - \chi(\eta)) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\kappa}_\eta(X_i) + o_{P_\eta}(1).$$

A comparison with (12.3) shows that $\tilde{\kappa}_\eta$ plays the same role as $\tilde{I}_{\theta,\eta}^{-1}\tilde{\ell}_{\theta,\eta}$ in the narrow semiparametric case, and hence the covariance matrix of $\tilde{\kappa}_\eta$ is comparable to the inverse of the efficient information matrix. With these substitutions one expects a Bernstein–von Mises theorem for a posterior distribution of $\chi(\eta)$ of the same type as before, with centering at an efficient estimator $\hat{\chi}_n$ and with the covariance matrix of $\tilde{\kappa}_\eta$ in place of the inverse information matrix:

$$\mathrm{E}_{\eta_0}\Big\|\Pi_n(\chi(\eta)\in\cdot\,|\,X_1,\ldots,X_n) - \mathrm{Nor}\Big(\hat{\chi}_n, n^{-1}P_{\eta_0}\Big(\tilde{\kappa}_{\eta_0}\tilde{\kappa}_{\eta_0}^\top\Big)\Big)\Big\|_{TV} \to 0. \tag{12.5}$$

A direction $\tilde{b}_\eta$ such that $B_\eta\tilde{b}_\eta = \tilde{\kappa}_\eta$ is the *least favorable direction*, in that if the parameter were known to belong to a one-dimensional model in this direction, then the influence function would still be $\tilde{\kappa}_\eta$. Hence statistical inference for χ in the submodel is “as hard” as it is in the full model. (A least-favorable direction may not exist, because the range of B_η need not be closed; the “supremum of the difficulty over the submodels” may not be assumed.)

The strict semiparametric setup can be incorporated in the general setup by replacing the parameter (θ, η) of the former setup by η and defining $\chi(\theta, \eta) = \theta$. It can be shown that the latter functional is differentiable as soon as the efficient influence function $\tilde{\ell}_{\theta,\eta}$ for θ exists, with efficient influence function $\tilde{\kappa}_{\theta,\eta} = \tilde{I}_{\theta,\eta}^{-1}\tilde{\ell}_{\theta,\eta}$ (see van der Vaart 1998, Lemma 25.25). Thus it suffices to work in the general setup.

12.3.1 Functionals

We shall give sufficient conditions for the Bernstein–von Mises theorem (12.5) in the general infinite-dimensional setup. As the conditions are in terms of the likelihood of the full observation, there is no advantage to restricting to the i.i.d. setup. Consider a general sequence of statistical models with observations $X^{(n)}$ following a jointly measurable density $p_\eta^{(n)}$ with respect to a σ-finite measure on some sample space, and a parameter η belonging to a Polish parameter set $\mathcal{H}$. For simplicity we take the functional of interest $\chi : \mathcal{H} \to \mathbb{R}$ real-valued.

The Bernstein–von Mises theorem is based on two approximations, which both involve approximately “least-favorable transformations.” These are defined as arbitrary measurable maps $\eta \mapsto \tilde{\eta}_n(\eta)$ that are one-to-one on each set of the form $\{\eta \in \mathcal{H}_n : \chi(\eta) = \theta\}$, and satisfy the conditions (12.7) and (12.8) below. They should be thought of as approaching η approximately in the direction of a scaled version of the least-favorable direction $\tilde{b}_{\eta_0}$, as explained in Section 12.3.1 below. For instance, if $\mathcal{H}$ is embedded in a linear space, then they often take the form $\tilde{\eta}_n(\eta) = \eta-(\chi(\eta)-\theta_0)\tilde{I}_n\tilde{b}_n$, for $\tilde{b}_n$ approximations to the least-favorable direction $\tilde{b}_{\eta_0}$ at the true parameter value and $\tilde{I}_n$ the efficient information (the inverse of the efficient variance). The transformation (and particular its scaling) will typically be chosen such that the value $\chi(\tilde{\eta}_n(\eta))$ of the functional is close to the constant value $\chi(\eta_0)$, although this is not an explicit part of the requirements; only (12.7) and (12.8) need to be satisfied.

Assume that the statistical models are locally asymptotically normal along the approximately least favorable directions in the following sense: for $\theta_0 = \chi(\eta_0)$,

$$\log \frac{p_\eta^{(n)}}{p_{\tilde{\eta}_n(\eta)}^{(n)}}(X^{(n)}) = \sqrt{n}(\chi(\eta)-\theta_0)\,\tilde{G}_n - \tfrac{1}{2}n\tilde{I}_n|\chi(\eta)-\theta_0|^2 + R_n(\eta), \tag{12.6}$$

where $\tilde{G}_n$ is a tight sequence of random variables, $\tilde{I}_n$ are positive numbers that are bounded away from zero, and $(R_n(\eta)\colon \eta\in\mathcal{H})$ are stochastic processes such that $\rho_{n,1}\to 0$, for

$$\rho_{n,1} = \sup_{\eta\in\mathcal{H}_n} \frac{|R_n(\eta)|}{1+n|\chi(\eta)-\theta_0|^2}. \tag{12.7}$$

Here $\mathcal{H}_n$ are sets in the parameter space on which the posterior distribution concentrates with probability tending to one. The variables $\tilde{G}_n$ and numbers $\tilde{I}_n$ may depend on η_0, but not on η, suggesting that the sets $\mathcal{H}_n$ must shrink to η_0.

A second main condition concerns an invariance of the prior of the nuisance parameter under a shift in the direction of the least-favorable direction. Assume that the conditional distribution $\Pi_{n,\theta}$ of $\tilde{\eta}_n(\eta)$ given $\chi(\eta)=\theta$ under the prior is absolutely continuous relative to this distribution Π_{n,θ_0} at $\theta=\theta_0$, with density $d\Pi_{n,\theta}/d\Pi_{n,\theta_0}$ satisfying $\rho_{n,2}\to 0$, for

$$\rho_{n,2} = \sup_{\eta\in\mathcal{H}_n} \frac{|\log(d\Pi_{n,\chi(\eta)}/d\Pi_{n,\theta_0}(\eta))|}{1+n|\chi(\eta)-\theta_0|^2}. \tag{12.8}$$

Theorem 12.8 (Semiparametric Bernstein–von Mises, functionals) *Suppose that there exist measurable sets $\mathcal{H}_n\subset\mathcal{H}$ and maps $\tilde{\eta}_n\colon\mathcal{H}\to\mathcal{H}$ that are one-to-one and bimeasurable on each set $\mathcal{H}_{n,\theta}:=\{\eta\in\mathcal{H}_n\colon\chi(\eta)=\theta\}$, equal to the identity on $\mathcal{H}_{n,\theta_0}$, for which* (12.7) *and* (12.8) *hold, and such that the sets $\Theta_n:=\{\chi(\eta)\colon\eta\in\mathcal{H}_n\}$ shrink to $\theta_0=\chi(\eta_0)$ with $\sqrt{n}(\Theta_n-\theta_0)\to\mathbb{R}$ and $\Pi_n(\eta\in\mathcal{H}_n|\,X^{(n)})\to 1$ and $\inf_{\theta\in\Theta_n}\Pi_n(\eta\in\tilde{\eta}_n(\mathcal{H}_{n,\theta})|\,X^{(n)},\chi(\eta)=\theta_0)\to 1$. If the induced prior for $\chi(\eta)$ possesses a positive, continuous Lebesgue density in a neighborhood of $\chi(\eta_0)$, then*

$$\mathrm{E}_{\eta_0}\Big\|\Pi_n(\chi(\eta)\in\cdot\,|\,X^{(n)}) - \mathrm{Nor}\Big(\chi(\eta_0)+\tilde{I}_n^{-1}\tilde{G}_n,\, n^{-1}\tilde{I}_n^{-1}\Big)\Big\|_{TV}\to 0.$$

Proof Because the posterior probability that $\eta\in\mathcal{H}_n$ tends to one by assumption, it suffices to establish the normal approximation to $\Pi_n(\chi(\eta)\in\cdot\,|\,X^{(n)},\eta\in\mathcal{H}_n)$. In particular, we may assume that $\chi(\eta)$ belongs (with posterior probability tending to one) to the sets Θ_n. Let $\Pi(d\theta)$ denote the marginal law of $\chi(\eta)$ and let $\Pi_\theta(d\eta)$ denote the conditional law of η given $\chi(\eta)=\theta$, under the prior law of η. By (12.6), for any Borel set B,

$$\Pi_n(\chi(\eta)\in B|\,X^{(n)},\eta\in\mathcal{H}_n) = \frac{\int_{B\cap\Theta_n} e^{\sqrt{n}(\theta-\theta_0)\tilde{G}_n-\frac{1}{2}n\tilde{I}_n|\theta-\theta_0|^2}\,Q_n(\theta)\,\Pi(d\theta)}{\int_{\Theta_n} e^{\sqrt{n}(\theta-\theta_0)\tilde{G}_n-\frac{1}{2}n\tilde{I}_n|\theta-\theta_0|^2}\,Q_n(\theta)\,\Pi(d\theta)},$$

where Q_n is defined as

$$Q_n(\theta) = \int_{\mathcal{H}_n} e^{R_n(\eta)} p^{(n)}_{\tilde{\eta}_n(\eta)}(X^{(n)})\, \Pi_\theta(d\eta).$$

The essential part of the proof is to show that Q_n is asymptotically independent of θ. If it were free of θ for any n, then it could be canceled out from the expression for the marginal posterior distribution, and the remaining expression would be the posterior distribution for a one-dimensional Gaussian location model, and hence satisfy the Bernstein–von Mises theorem. We show below that this is approximately true in that there exist constants $\rho_n \to 0$ such that, for $\theta \in \Theta_n$, with probability tending to one,

$$e^{-\rho_n(1+n|\theta-\theta_0|^2)} \leq \frac{Q_n(\theta)}{Q_n(\theta_0)} \leq e^{\rho_n(1+n|\theta-\theta_0|^2)}. \tag{12.9}$$

Substituting this in the preceding display, and using that the restriction of $\Pi(d\theta)$ to a neighborhood of Θ_n possesses a Lebesgue density that is bounded below and above by positive constants, which can be taken arbitrarily close if Θ_n is sufficiently small, we see that $\Pi_n(\chi(\eta) \in B|\, X^{(n)}, \eta \in \mathcal{H}_n)$ is lower and upper bounded by

$$(1+o(1))e^{\pm\rho_n} \frac{\int_{B\cap\Theta_n} e^{\sqrt{n}(\theta-\theta_0)\tilde{G}_n - \frac{1}{2}n(\tilde{I}_n \mp \rho_n)|\theta-\theta_0|^2}\, d\theta}{\int_{\Theta_n} e^{\sqrt{n}(\theta-\theta_0)\tilde{G}_n - \frac{1}{2}n(\tilde{I}_n \pm \rho_n)|\theta-\theta_0|^2}\, d\theta}.$$

From the assumption that the neighborhoods Θ_n shrink to θ_0 at slower rate than $1/\sqrt{n}$, it can be seen that replacing them by $\mathbb{R}$ multiplies the numerator and denominator by $1+o(1)$ terms. The resulting expression can be seen to be asymptotically equivalent to $\mathrm{Nor}(I_n^{-1}\tilde{G}_n, n^{-1}\tilde{I}_n^{-1})(B-\theta_0)$, uniformly in B, by explicit calculation of the integrals.

It remains to establish (12.9). By (12.7) the term $e^{R_n(\eta)}$ in the definition of $Q_n(\theta)$ can be lower and upper bounded by $e^{\pm\rho_{n,1}(1+n|\theta-\theta_0|^2)}$. Next, by the assumption that $\tilde{\eta}_n$ is one-to-one on the set $\mathcal{H}_{n,\theta}$, we have that $\eta \in \mathcal{H}_n$ and $\chi(\eta) = \theta$ if and only if $\tilde{\eta}_n(\eta) \in \tilde{\eta}_n(\mathcal{H}_{n,\theta})$ and $\chi(\eta) = \theta$. Therefore by making the substitution $\eta \mapsto \tilde{\eta}_n(\eta)$ in the integral, we see that $Q_n(\theta)$ is lower and upper bounded by $e^{\pm\rho_{n,1}(1+n|\theta-\theta_0|^2)}\tilde{Q}_n(\theta)$, for $\tilde{Q}_n(\theta)$ defined by

$$\tilde{Q}_n(\theta) = \int_{\tilde{\eta}_n(\mathcal{H}_{n,\theta})} p^{(n)}_\eta(X^{(n)})\, \Pi_{n,\theta}(d\eta),$$

where $\Pi_{n,\theta}$ is the law of $\tilde{\eta}_n(\eta)$ if $\eta \sim \Pi_\theta$. This is the same law as in the statement of the theorem, so that we can use assumption (12.8) to change $\Pi_{n,\theta}$ to $\Pi_{n,\theta_0} = \Pi_{\theta_0}$, at the cost of inserting a further multiplicative factor $e^{\pm\rho_{n,2}(1+n|\theta-\theta_0|^2)}$ in the lower and upper bounds. The resulting expression is the numerator of the posterior probability

$$\Pi_n(\eta \in \tilde{\eta}_n(\mathcal{H}_{n,\theta})|\, X^{(n)}, \chi(\eta) = \theta_0) = \frac{\int_{\tilde{\eta}_n(\mathcal{H}_{n,\theta})} p^{(n)}_\eta(X^{(n)})\, \Pi_{\theta_0}(d\eta)}{\int p^{(n)}_\eta(X^{(n)})\, \Pi_{\theta_0}(d\eta)}.$$

Since this posterior probability tends to 1 by assumption, it is between $e^{-\rho_{n,3}}$ and 1 for some $\rho_{n,3} \to 0$, uniformly in $\theta \in \Theta_n$. In other words, with probability tending to one, $\tilde{Q}_n(\theta)/\tilde{Q}_n(\theta_0)$ is bounded below and above by $e^{\pm\rho_{n,2}(1+n|\theta-\theta_0|^2)}e^{\pm\rho_{n,3}}$.

This concludes the proof of claim (12.9), with $\rho_n = \sum_i \rho_{n,i}$. □

Although this is not included in the condition, for a true Bernstein–von Mises theorem the variables $\tilde{G}_n$ ought to be asymptotically normally distributed with mean zero and variance (the limit of) $\tilde{I}_n$. This is typically the case (and almost implied by the LAN expansion (12.6)).

The sieves $\mathcal{H}_n$ in the theorem are meant to make the expansions and approximations (12.7) and (12.8) possible, which favors small sieves, but must asymptotically contain posterior mass one. Typically they would be neighborhoods that shrink to the true parameter η_0 at the rate of contraction of the posterior distribution. Then an application of the theorem would be preceded by a derivation of a rate of contraction of the posterior distribution, possibly by the methods of Chapter 8.

In the case of i.i.d. observations, the LAN expansion (12.7) is implied by the pair of approximations, with $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_{\eta_0})$ the empirical process of the observations and $\tilde{G}_n = \mathbb{G}_n(\tilde{\ell}_n)$: for any $\hat{\eta} \in \mathcal{H}_n$ and $\hat{\theta} = \chi(\hat{\eta})$,

$$\mathbb{G}_n\bigl[\log p_{\hat{\eta}} - \log p_{\tilde{\eta}_n(\hat{\eta})} - (\hat{\theta} - \theta_0)\tilde{\ell}_{\eta_0}\bigr] = o_P(\hat{\theta} - \theta_0),$$
$$P_{\eta_0}\bigl[\log p_{\hat{\eta}} - \log p_{\tilde{\eta}_n(\hat{\eta})}\bigr] = -\tfrac{1}{2}\tilde{I}_{\eta_0}(\hat{\theta} - \theta_0)^2(1 + o_P(1)) + o_P(|\hat{\theta} - \theta_0|n^{-1/2}).$$

The first can be verified by showing that the functions within square brackets divided by $\hat{\theta} - \theta_0$ are with probability tending to one contained in a Donsker class, and have second moments tending to zero. Typically the function $\tilde{\ell}_\eta$ will be the efficient score function (the efficient influence function $\tilde{\kappa}_\eta$ divided by its variance) at $\eta = \eta_0$. The second condition is ostensibly a Taylor expansion to the second order. However, because the expectation is relative to the true parameter η_0 and the difference on the left is a difference between perturbations of $\hat{\eta}$, the linear term in the expansion does not necessarily vanish, but takes the form $(\hat{\theta} - \theta_0)P_{\eta_0}\tilde{\ell}_{\hat{\eta}}$. Then the fulfilment of the condition seems to demand that $P_{\eta_0}\tilde{\ell}_{\hat{\eta}} = o_P(n^{-1/2})$. In general this requires a rate for the convergence of $\tilde{\ell}_{\hat{\eta}}$ to the efficient influence function. This is similar to the "no-bias" condition found in the analysis of semiparametric maximum likelihood estimators (see Murphy and van der Vaart 2000 or equation (25.75) and the ensuing discussion in van der Vaart 1998). Analogous reasoning suggests that the bias is quadratic in the discrepancy and is negligible if the rate of contraction is at least $o_P(n^{-1/4})$. If not, then the prior may create a bias and the Bernstein–von Mises theorem may not hold.

Full LAN Expansion

The conditions of the preceding theorem can alternatively be phrased in terms of a LAN expansion of the full likelihood. This is more involved, but makes it evident that for the condition to be satisfied the transformations $\eta \mapsto \tilde{\eta}_n(\eta)$ in (12.6) must be approximately in the least favorable direction and $\tilde{I}_n$ the efficient information.

Suppose that the linear space $\mathbb{B}_0$ that encompasses the parameter set $\mathcal{H}_n$ is a Hilbert space, with inner product $\langle\cdot,\cdot\rangle_0$ and norm $\|\cdot\|_0$, and make two assumptions:

(i) There exist elements $\tilde{b}_n \in \mathbb{B}_0$ such that

$$\sup_{\substack{\eta\in\mathcal{H}_n\\ \theta=\chi(\eta),\theta_0=\chi(\eta_0)}} \frac{n\Bigl|\bigl\langle \eta - \eta_0 - (\theta - \theta_0)\tilde{b}_n, (\theta - \theta_0)\tilde{b}_n\bigr\rangle_0\Bigr|}{1 + n|\theta - \theta_0|^2} \to 0. \qquad (12.10)$$

(ii) The full statistical model is LAN in the sense that

$$\log \frac{p_\eta^{(n)}}{p_{\eta_0}^{(n)}}(X^{(n)}) = \sqrt{n}\,\mathbb{G}_n(\eta-\eta_0) - \tfrac{1}{2}n\|\eta-\eta_0\|_0^2 + R_n(\eta),$$

for linear stochastic processes $(\mathbb{G}_n(b)\colon b\in\mathbb{B}_0)$, and stochastic processes $(R_n(\eta)\colon \eta\in\mathcal{H})$ such that

$$\sup_{\substack{\eta\in\mathcal{H}_n\\ \theta=\chi(\eta),\theta_0=\chi(\eta_0)}} \frac{\bigl|R_n(\eta)-R_n(\eta-(\theta-\theta_0)\tilde{b}_n)\bigr|}{1+n|\theta-\theta_0|^2} \to 0. \tag{12.11}$$

Assumptions (i)–(ii) can be seen to imply (12.6), with the transformations $\tilde{\eta}_n(\theta)=\eta-(\chi(\eta)-\theta_0)\tilde{b}_n$, the variables $\tilde{G}_n=\mathbb{G}_n(\tilde{b}_n)$ and $\tilde{I}_n=\|\tilde{b}_n\|_0^2$, and the remainder $R_n(\eta)$ of (12.6) equal to the sum of twice the numerator of (12.10) and the present $R_n(\eta)-R_n(\tilde{\eta}_n(\eta))$. That condition (12.11) is placed on the latter difference of two remainders makes the two conditions in fact almost equivalent, apart from the introduction of the Hilbert space structure and the linearity of the transformation and the process $\mathbb{G}_n$ in (12.10)–(12.11).

Relation (12.10) is certainly satisfied if the inner product in its numerator vanishes, for all η. This is the case if $(\theta-\theta_0)\tilde{b}_n$ is the orthogonal projection of $\eta-\eta_0$ onto the one-dimensional linear space spanned by $\tilde{b}_n$, i.e. $\theta-\theta_0=\langle\eta-\eta_0,\tilde{b}_n\rangle_0/\tilde{I}_n$. We shall give a heuristic argument that this implies that $\tilde{b}_n$ must be close to a least favorable direction.

The Hilbert space $(\mathbb{B}_0,\langle\cdot,\cdot\rangle_0)$ would typically derive from the information structure of the statistical model at the true parameter value η_0. In the general notation of the introduction to this section,

$$\langle a,b\rangle_0 = P_{\eta_0}\bigl[(B_{\eta_0}a)(B_{\eta_0}b)\bigr].$$

If the efficient influence function can be written in the form $\tilde{\kappa}_{\eta_0}=B_{\eta_0}\tilde{b}_{\eta_0}$, where by definition $\tilde{b}_{\eta_0}$ is a least favorable direction, then the differentiability (12.4) of the functional χ at η_0 gives, for a path η_t that approaches η_0 in the "direction" b,

$$\chi(\eta_t)=\chi(\eta_0)+tP_{\eta_0}[\tilde{\kappa}_{\eta_0}(B_{\eta_0}b)]+o(t)=\chi(\eta_0)+t\langle\tilde{b}_{\eta_0},b\rangle_0+o(t).$$

Thus $\tilde{b}_{\eta_0}$ is a "derivative" of χ at η_0 relative to the information inner product $\langle\cdot,\cdot\rangle_0$, and informally we have that $\chi(\eta)-\chi(\eta_0)\doteq\langle\tilde{b}_{\eta_0},\eta-\eta_0\rangle_0$. If this were true exactly, and $\eta-\eta_0-(\chi(\eta)-\theta_0)\tilde{b}_n$ would be orthogonal to $\tilde{b}_n$, for every η, then it can be seen that $\tilde{b}_n=\tilde{b}_{\eta_0}/\|\tilde{b}_{\eta_0}\|_0^2$ (see Problem 12.8).

If an exact least favorable direction exists, then assumption (12.10) can also be formulated in terms of the distance between $\tilde{b}_n$ and this direction. This gives the alternative condition:

(i′) there exists $\tilde{b}_0\in\mathbb{B}_0$ such that $\langle\eta-\eta_0-(\theta-\theta_0)\tilde{b}_0,\tilde{b}_0\rangle_0=0$ and

$$\sqrt{n}\|\tilde{b}_n-\tilde{b}_0\|_0\sup_{\eta\in\mathcal{H}_n}\|\eta-\eta_0\|_0\to 0,\qquad\text{and}\qquad \|\tilde{b}_n-\tilde{b}_0\|_0\to 0. \tag{12.12}$$

That this condition is stronger than (12.10) can be seen by comparing the numerator of (12.10) with its zero value when $\tilde{b}_n$ is replaced by $\tilde{b}_0$ and using the inequality $\bigl|\langle a,b\rangle_0-\langle a',b'\rangle_0\bigr|\le\|a-a'\|_0\|b\|_0+\|b-b'\|_0\|a'\|_0$, for any directions a,b. The function $\tilde{b}_0$ will be a scaled version $\tilde{b}_0=\tilde{b}_{\eta_0}/\|\tilde{b}_{\eta_0}\|_0^2$ of the least favorable direction $\tilde{b}_{\eta_0}$ encountered previously.

12.3.2 Strict Semiparametric Model

A semiparametric model in the narrow sense is indexed by a partitioned parameter (θ, η), and interest is in the functional $\chi(\theta, \eta) = \theta$. Score functions take the form $a\dot\ell_{\theta,\eta} + B_{\theta,\eta}b$, and are also indexed by pairs (a, b) of "directions." A least favorable direction, if it exists, is one that gives the efficient score function $\tilde\ell_{\theta,\eta}$ as score function, and is given by a direction of the form $(1, -\tilde b_{\theta,\eta})$, yielding the score $\dot\ell_{\theta,\eta} - B_{\theta,\eta}\tilde b_{\theta,\eta}$. For a linear parameter space this motivates to choose the transformation of η of the form $\eta + (\theta - \theta_0)\tilde b_{\theta_0,\eta_0}$, but it may be profitable to replace $\tilde b_{\theta_0,\eta_0}$ by an approximation. (The plus sign arises because in the likelihood ratio (12.13) the transformation on η is inserted in the denominator, whereas θ is perturbed in the numerator.)

We choose the least favorable transformation as in (12.7) of a parameter (θ, η) to take the functional value θ_0, i.e. we use a transformation of the form $(\theta, \eta) \mapsto (\theta_0, \tilde\eta_n(\theta, \eta))$. Condition (12.7) then requires that there exists a tight sequence of variables $\tilde G_n$ and positive numbers $\tilde I_n$ so that the remainder

$$\log \frac{p^{(n)}_{\theta,\eta}}{p^{(n)}_{\theta_0,\tilde\eta_n(\theta,\eta)}}(X^{(n)}) = \sqrt{n}(\theta - \theta_0)\tilde G_n - \tfrac{1}{2}n\tilde I_n|\theta - \theta_0|^2 + R_n(\theta, \eta)$$

to the quadratic expansion of the log likelihood satisfies

$$\sup_{\substack{\theta\in\Theta_n\\ \eta\in\mathcal{H}_n}} \frac{R_n(\theta, \eta)}{1 + n|\theta - \theta_0|^2} \to 0. \tag{12.13}$$

The variables $\tilde G_n$ and numbers $\tilde I_n$ may depend on (θ_0, η_0), but not on (θ, η). The numbers $\tilde I_n$ must be bounded away from zero.

It is natural to choose the two parameters θ and η independent under the prior. Then the measure $\Pi_{n,\theta}$ in (12.8), the distribution of $(\theta_0, \tilde\eta_n(\theta, \eta))$ given that $\chi(\theta, \eta) = \theta$,[2] is a product of the Dirac measure at θ_0 times the law of $\tilde\eta_n(\theta, \eta)$ under the prior on η, for fixed θ. The Dirac factors cancel and the condition reduces to exactly the same condition, but with $\Pi_{n,\theta}$ the distribution of $\tilde\eta_n(\theta, \eta)$. Assume that $\rho_{n,2} \to 0$, for

$$\rho_{n,2} = \sup_{\substack{\theta\in\Theta_n\\ \eta\in\mathcal{H}_n}} \frac{|\log(d\Pi_{n,\theta}/d\Pi_{n,\theta_0}(\eta))|}{1 + n|\theta - \theta_0|^2}. \tag{12.14}$$

In these conditions Θ_n and $\mathcal{H}_n$ may be arbitrary measurable sets of parameters, such that the posterior distribution for (θ, η) concentrates on $\Theta_n \times \mathcal{H}_n$ with probability tending to one. The sets Θ_n are also assumed to shrink at θ_0 at a rate such that $\sqrt{n}(\Theta_n - \theta_0)$ increases to $\mathbb{R}$.

Theorem 12.8 specializes to this setup in the following form. Let $\tilde\eta_n(\theta, \mathcal{H}_n)$ be the set of all parameters $\tilde\eta_n(\theta, \eta)$ as η ranges over $\mathcal{H}_n$.

Theorem 12.9 (Semiparametric Bernstein–von Mises) *Suppose that there exist measurable sets Θ_n and $\mathcal{H}_n$ for every θ a one-to-one bimeasurable map $\eta \mapsto \tilde\eta_n(\theta, \eta)$ for which* (12.13) *and* (12.14) *hold, for $\Pi_{n,\theta}$ the prior distribution of $\tilde\eta_n(\theta, \eta)$ given fixed θ, and such*

[2] The notation is awkward; the last appearance of θ is a fixed value, the first two appearances in this assertion refer to the prior random variable θ.

that $\Pi_n(\theta \in \Theta_n, \eta \in \mathcal{H}_n |\, X^{(n)}) \to 1$ *and* $\inf_{\theta \in \Theta_n} \Pi_n(\eta \in \tilde{\eta}_n(\theta, \mathcal{H}_n) |\, X^{(n)}, \theta = \theta_0) \to 1$. *If the prior for* θ *possesses a continuous, positive Lebesgue density in a neighborhood of* θ_0*, then*

$$\mathrm{E}_{\theta_0,\eta_0} \Big\| \Pi_n(\theta \in \cdot \,|\, X^{(n)}) - \mathrm{Nor}(\theta_0 + \tilde{I}_n^{-1}\tilde{G}_n, n^{-1}\tilde{I}_n^{-1}) \Big\|_{TV} \to 0.$$

Example 12.10 (Adaptive model) A strict semiparametric model is called *adaptive* if there is no loss of information in not knowing the nuisance parameter. In such a model the efficient score function $\tilde{\ell}_{\theta,\eta}$ coincides with the ordinary score function $\dot{\ell}_{\theta,\eta}$ and the least-favorable direction $\tilde{b}_{\theta,\eta}$ in the nuisance parameter space is equal to zero.

For $\tilde{\eta}_n(\theta, \eta) = \eta$ condition (12.14) is trivially satisfied, and (12.13) simplifies into an ordinary LAN condition on the parametric models $\theta \mapsto p_{\theta,\eta}$, uniformly in $\eta \in \mathcal{H}_n$, where the centering variables $\tilde{G}_n$ may *not* depend on η. The latter suggests that the sets $\mathcal{H}_n$ must shrink to a point as $n \to \infty$.

Example 12.11 (Gaussian prior) A shifted Gaussian prior is absolutely continuous relative to the original Gaussian process if and only if the shift belongs to its reproducing kernel Hilbert space. In that case the Cameron-Martin theorem (see Lemma I.20) gives an explicit description of the density. This shows where to choose the approximate least favorable directions $\tilde{b}_n$ and enables to verify (12.14) for a Gaussian prior on η.

If η is a centered Gaussian random element in a separable Banach space and $\tilde{b}_n$ is in its RKHS $\mathbb{H}$, then the law $\Pi_{n,\theta}$ of $\eta + (\theta - \theta_0)\tilde{b}_n$ has log-density

$$\log \frac{d\Pi_{n,\theta}}{d\Pi_{n,\theta_0}}(\eta) = (\theta - \theta_0) U(\tilde{b}_n, \eta) - \tfrac{1}{2}|\theta - \theta_0|^2 \|\tilde{b}_n\|_{\mathbb{H}}^2,$$

where $U(\tilde{b}_n, \eta)$ is a centered Gaussian random variable with variance $\|\tilde{b}_n\|_{\mathbb{H}}^2$, under $\Pi_{n,\theta_0} = \Pi$. If we set $\mathcal{H}_n = \{\eta \in \mathcal{H}\colon |U(\tilde{b}_n, \eta)| \leq 2\sqrt{n}\epsilon_n \|\tilde{b}_n\|_{\mathbb{H}}\}$, then

$$\sup_{\substack{\eta \in \mathcal{H}_n \\ \theta = \chi(\eta)}} \frac{|\log(d\Pi_{n,\theta}/d\Pi_{n,\theta_0}(\eta))|}{1 + n|\theta - \theta_0|^2} \leq 2\epsilon_n \|\tilde{b}_n\|_{\mathbb{H}} + \frac{1}{n}\|\tilde{b}_n\|_{\mathbb{H}}^2.$$

Thus (12.14) is satisfied provided $\epsilon_n \|\tilde{b}_n\|_{\mathbb{H}} \to 0$, for some $\epsilon_n \gg n^{-1/2}$. To ensure that the sieve $\mathcal{H}_n$ has posterior probability tending to one, we choose ϵ_n large enough, so that the prior mass of $\mathcal{H}_n^c$ is small enough. Since $U(\tilde{b}_n, \eta)$ is Gaussian, the tail bound for the normal distribution gives $\Pi(\mathcal{H}_n^c) \leq e^{-2n\epsilon_n^2}/(\sqrt{n}\epsilon_n)$. Then Theorem 8.20 shows that the posterior mass of $\mathcal{H}_n$ tends to one if

$$\Pi\Big((\theta, \eta)\colon K(p_{\theta_0,\eta_0}^{(n)}; p_{\theta,\eta}^{(n)}) \leq n\epsilon_n^2, V_{2,0}(p_{\theta_0,\eta_0}^{(n)}; p_{\theta,\eta}^{(n)}) \leq n\epsilon_n^2\Big) \geq e^{-n\epsilon_n^2}.$$

Often this will be satisfied for ϵ_n the rate of contraction for the full densities, which in turn will typically be determined by the concentration function of the Gaussian prior at η_0.

In the situation that there exists an exact least-favorable direction $\tilde{b}_0$, the resulting rate $\epsilon_n \|\tilde{b}_n\|_{\mathbb{H}} \to 0$ must be traded to the rate of approximation of $\tilde{b}_n$ to $\tilde{b}_0$, as measured in (12.12). For a Gaussian prior this rate can be bounded by the concentration function $\varphi_{\tilde{b}_0}(\cdot)$ given in (11.11) relative to the information norm: for any δ_n there exists $\tilde{b}_n$ with $\|\tilde{b}_n - \tilde{b}_0\|_0 \leq \delta_n$ and $\|\tilde{b}_n\|_{\mathbb{H}}^2 \leq \varphi_{\tilde{b}_0}(\delta_n)$. Then the pair of conditions (12.12) and (12.14) are satisfied if, for

ϵ_n as in the preceding display and $\epsilon_{n,0}$ the rate of posterior concentration relative to the information metric $\|\cdot\|_0$, and some δ_n,

$$\epsilon_n^2 \varphi_{\tilde{b}_0}(\delta_n) \to 0 \qquad \text{and} \qquad \sqrt{n}\delta_n \epsilon_{n,0} \to 0. \tag{12.15}$$

In nice cases the rates ϵ_n and $\epsilon_{n,0}$ will agree, and be given by the rate equation $\varphi_{\eta_0}(\epsilon_n) \asymp n\epsilon_n^2$. This leads to the condition that there exists a solution $\delta_n \downarrow 0$ to

$$\varphi_{\eta_0}(\epsilon_n) \asymp n\epsilon_n^2 \quad \text{and} \quad \epsilon_n^2 \varphi_{\tilde{b}_0}(\delta_n) \to 0 \quad \text{and} \quad \delta_n^2 \varphi_{\eta_0}(\epsilon_n) \to 0.$$

The two moduli in this display will typically depend on the "regularity" of the functions η_0 and b_0. Here the true parameter η_0 is a given function of a given "regularity," but the least-favorable direction $\tilde{b}_0$ depends on the surrounding parameters in the full "model," as defined implicitly by the prior. This makes the final message opaque in general. Fine properties of the prior and model may matter.

12.3.3 Cox Proportional Hazard Model

In the Cox model under random right censoring (also see Section 13.6) we observe a random sample of observations distributed as (T, Δ, Z), for $T = X \wedge C$ the minimum of a survival time X and a censoring time C, an indicator $\Delta = \mathbb{1}\{X \leq C\}$ that records if the observation is censored $(\Delta = 0)$ or not, and a covariate variable Z. For simplicity we assume that the latter variable is univariate and takes its values in a compact interval in the real line. The survival and censoring times X and C are assumed to be conditionally independent given Z, and the Cox model postulates the conditional hazard function of X given Z to take the form

$$h(x|\, Z) = e^{\theta Z} h(x),$$

for an unknown *baseline hazard* function h, and an unknown real-valued parameter θ. (Section 13.1 gives a brief introduction to survival analysis, including the definition (13.2) of a hazard function.)

For H the cumulative hazard function corresponding to h, a density of (T, Δ, Z) relative to a suitable dominating measure is given by

$$\Bigl(e^{\theta z} h(t) e^{-e^{\theta z} H(t)} \bar{F}_{C|Z}(t-|\, z)\Bigr)^{\delta} \Bigl(e^{-e^{\theta z} H(t)} f_{C|Z}(t|\, z)\Bigr)^{1-\delta} p_Z(z). \tag{12.16}$$

The terms involving the conditional distribution of C given Z and the marginal distribution of Z factor out of this likelihood seen as a function of (θ, h), and can be ignored for inference on (θ, h) if these parameters are chosen a priori independent. Thus we equip only θ and h with priors, which we shall choose independent also. We shall further assume that the censoring distribution is supported on a compact interval $[0, \tau]$, with a positive atom at its right end point τ. (The atom simplifies the technical arguments, but is natural too: it corresponds to ending the study at the finite time τ.) Then the hazard functions h can be restricted to $[0, \tau]$, and the prior on h will be a prior on integrable functions $h\colon [0, \tau] \to [0, \infty)$.

The score function for θ takes the form

$$\dot{\ell}_{\theta,h}(t, \delta, z) = \delta z - z e^{\theta z} H(t).$$

For any measurable function $b: [0,\tau] \to \mathbb{R}$, the path defined by $h_u = he^{ub}$ defines a submodel passing through h at $u = 0$. Its score function at $u = 0$ takes the form

$$B_{\theta,h}b(t,\delta,z) = \delta b(t) - e^{\theta z} \int_{[0,t]} b\,dH.$$

This can be seen to be a mapping $B_{\theta,h}: \mathbb{L}_2(h) \to \mathbb{L}_2(p_{\theta,h})$, and hence it has an adjoint mapping $B_{\theta,h}^\top: \mathbb{L}_2(p_{\theta,h}) \to \mathbb{L}_2(h)$. For continuous H it can be shown that (for details on this and subsequent formulas, see e.g. van der Vaart 1998, Section 25.12.1)

$$B_{\theta,h}^\top B_{\theta,h} h(t) = h(t)\, M_{0,\theta,h}(t), \qquad B_{\theta,h}^\top \dot\ell_{\theta,h}(t) = M_{1,\theta,h}(t),$$

where

$$M_{k,\theta,h}(t) = \mathrm{E}_{\theta,h} \mathbb{1}_{T \geq t} Z^k e^{\theta Z}, \qquad k = 0, 1, \ldots$$

Thus the information operator $B_{\theta,h}^\top B_{\theta,h}$ is simply a multiplication by the function $M_{0,\theta,h}$. If this function is bounded away from zero on its domain, then the operator is invertible, with the inverse being division by the same function. By the general semiparametric theory in the introduction of the section a least-favorable direction is given by

$$\tilde b_{\theta,h}(t) = (B_{\theta,h}^\top B_{\theta,h})^{-1} B_{\theta,h}^\top \dot\ell_{\theta,h}(t) = \frac{M_{1,\theta,h}}{M_{0,\theta,h}}(t).$$

The efficient score function takes the form $\tilde\ell_{\theta,h} = \dot\ell_{\theta,h} - B_{\theta,h}\tilde b_{\theta,h}$, and is explicitly given as

$$\tilde\ell_{\theta,h}(t,\delta,z) = \delta\Big(z - \frac{M_{1,\theta,h}}{M_{0,\theta,h}}(t)\Big) - e^{\theta z} \int_{[0,t]} \Big(z - \frac{M_{1,\theta,h}}{M_{0,\theta,h}}(s)\Big)\, dH(s).$$

The square information norm at parameter (θ, h) is

$$\begin{aligned} \|(a,b)\|_{\theta,h}^2 &= a^2 I_{\theta,h} + 2a\langle \dot\ell_{\theta,h}, B_{\theta,h}b\rangle_{\theta,h} + \langle B_{\theta,h}b, B_{\theta,h}b\rangle_{\theta,h} \qquad (12.17)\\ &= \int_{[0,\tau]} (a^2 M_{2,\theta,h}(t) + 2a\,b(t) M_{1,\theta,h}(t) + b^2(t) M_{0,\theta,h}(t))\, dH(t). \end{aligned}$$

Finally, the efficient information for θ can be computed as

$$\tilde I_{\theta,h} = \mathrm{E}_Z \int \Big(Z - \frac{M_{1,\theta,h}}{M_{0,\theta,h}}(t)\Big)^2 \bar F_{C|Z}(t-|\,Z)\, e^{\theta Z} e^{-e^{\theta Z} H(t)}\, dH(t).$$

This is positive as soon as Z is not almost surely equal to a function of T.

We consider a Gaussian prior on the function $\eta = \log h: [0,\tau] \to \mathbb{R}$. Let $\varphi_w(\epsilon)$ be its concentration function (11.11) at w relative to the uniform norm on functions on $[0,\tau]$.

Theorem 12.12 *Suppose that* $\mathrm{P}(C \geq \tau) = \mathrm{P}(C = \tau) > 0$, *that the conditional distribution of* C *given* Z *admits a continuous strictly positive Lebesgue density on* $[0,\tau)$, *and that* $\mathrm{P}_{\theta_0,h_0}(X \geq \tau) > 0$. *If there exist* ϵ_n *and* δ_n *with* $\varphi_{\eta_0}(\epsilon_n) \asymp n\epsilon_n^2$ *and* $\epsilon_n^2 \varphi_{\tilde b_{\theta_0,\eta_0}}(\delta_n) \to 0$ *and* $n^{3/2}\epsilon_n^4 \to 0$ *and* $\delta_n \epsilon_n \sqrt{n} \to 0$, *then the Bernstein–von Mises theorem for* θ *holds at* (θ_0, h_0).

Proof The theorem is a corollary of Theorem 12.9, applied to the transformations $\tilde\eta_n(\theta,\eta) = \eta + (\theta-\theta_0)\tilde b_n$, where $\tilde b_n$ will be chosen to approximate the least-favorable direction $\tilde b_{\theta_0,\eta_0}$. The proof uses various comparisons between metrics provided in Lemma 12.14 below. We reparameterize the model from (θ, h) to (θ,η), but keep the notation H for the cumulative hazard function corresponding to $h = e^\eta$. We choose Θ_n equal to open intervals that shrink to θ_0 slowly and $\mathcal{H}_n = \{\eta\colon d_H(p_{\theta,\eta}, p_{\theta_0,\eta_0}) < \epsilon_n, \|\eta\|_\infty^2 \lesssim n\epsilon_n^2\}$, where ϵ_n is a big multiple of the rate of contraction of the posterior distribution relative to the Hellinger distance. In view of Lemma 12.14(i)–(iii)+(v) the Hellinger distance is bounded above by the sum of the Euclidean distance on θ and the uniform distance on η, and the Kullback-Leibler neighborhoods contain neighborhoods for this sum-distance. It follows from Theorems 11.20 and 8.9 that an upper bound on ϵ_n is given by the solution to the inequality $\varphi_{\eta_0}(\epsilon_n) \lesssim n\epsilon_n^2$. Furthermore, the prior mass of a Kullback-Leibler type neighborhood of p_{θ_0,η_0} of radius a multiple of ϵ_n is at least $e^{-n\epsilon_n^2}$. By Borell's inequality the prior probability that $\|\eta\|_\infty$ exceeds $C\sqrt{n}\epsilon_n$ is bounded above by $e^{-C_1 n\epsilon_n^2}$, for a constant C_1 that can be made arbitrarily large by choosing C large. Therefore, by Theorem 8.20 the posterior probability that $\|\eta\|_\infty^2 \le Cn\epsilon_n^2$ tends to one. Combined the preceding shows that the posterior mass of $\mathcal{H}_n$ tends to one as well.

By Lemma 12.14(xi) the posterior consistency for the Hellinger distance implies consistency for the supremum norm: $\|H - H_0\|_\infty \to 0$, for any $\eta \in \mathcal{H}_n$.

By the definition of the concentration functions there exists, for every $\delta_n > 0$, elements $\tilde b_n$ such that

$$\|\tilde b_n - \tilde b_{\theta_0,h_0}\|_\infty < \delta_n, \qquad \|\tilde b_n\|_{\mathbb{H}}^2 \le \varphi_{\tilde b_{\theta_0,h_0}}(\delta_n).$$

We use these approximative least-favorable directions. Then condition (12.14) is valid provided $\epsilon_n\|\tilde b_n\|_{\mathbb{H}} \to 0$, as explained in Example 12.11.

The verification of (12.13) is based on establishing the pair of assertions, where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_{\theta_0,\eta_0})$ is the empirical process of the observations (T_i, Δ_i, Z_i): for any random sequences $\hat\theta \in \Theta_n$ and $\hat\eta \in \mathcal{H}_n$,

$$\mathbb{G}_n\big[\log p_{\hat\theta,\hat\eta} - \log p_{\theta_0,\hat\eta+(\hat\theta-\theta_0)\tilde b_n} - (\hat\theta-\theta_0)(\tilde\ell_{\theta_0,\eta_0} - B_{\theta_0,\eta_0}\tilde b_n)\big] = o_P(|\hat\theta-\theta_0|), \quad (12.18)$$

$$P_{\theta_0,\eta_0}\big[\log p_{\hat\theta,\hat\eta} - \log p_{\theta_0,\hat\eta+(\hat\theta-\theta_0)\tilde b_n}\big] = -\tfrac{1}{2}\tilde I_{\theta_0,\eta_0}(\hat\theta-\theta_0)^2(1+o_P(1)) + o_P(|\hat\theta-\theta_0|n^{-1/2}). \quad (12.19)$$

For $\theta_u = \theta_0 + u(\theta-\theta_0)$ we can write the function in square brackets in (12.18), divided by $\theta-\theta_0$ and with $\hat\theta$ and $\hat\eta$ replaced by θ and η, as

$$\begin{aligned}&\int_0^1 \big[(\dot\ell_{\theta_u,\eta} - \dot\ell_{\theta_0,\eta} - (B_{\theta_0,\eta+u(\theta-\theta_0)\tilde b_n} - B_{\theta_0,\eta_0})(\tilde b_n)\big]\,du\\ &\qquad = \int_0^1 \Big[-ze^{\theta_u z}H(t) + z^{\theta_0 z}H_0(t) + e^{\theta_0 z}\int_0^t \tilde b_n\,(e^{u(\theta-\theta_0)\tilde b_n}\,dH - dH_0)\Big]\,du.\end{aligned}$$

The functions $z \mapsto ze^{\theta z}$, $t \mapsto H(t)$ and $t \mapsto \int_0^t \tilde b_n e^{u(\theta-\theta_0)\tilde b_n}\,dH$ are all contained in bounded, universal Donsker classes, if $\|H\|_\infty$ remain bounded. Hence the functions appearing in the integral on the right are contained in a Donsker class by preservation of the Donsker property under Lipschitz transformations (see van der Vaart and Wellner 1996, Chapter 2.10). If $\theta \to \theta_0$, $H \rightsquigarrow H_0$ and $\|\tilde b_n - b_{\theta_0,h_0}\|_\infty$, then the second moments of these functions tend to zero. Therefore (12.18) follows by Lemma 3.3.5 in van der Vaart and Wellner (1996).

By Taylor's theorem with integral remainder, $g(1) = g(0) + g'(0) + \int_0^1 g''(u)(1-u)\,du$, applied to $g(u) = \log(p_{\theta_u,\eta}/p_{\theta_0,\eta+(\theta_u-\theta_0)\tilde b_n})$, we can write the left side of (12.19) with $\hat\theta$ and $\hat\eta$ replaced by θ and η, as

$$(\theta-\theta_0)P_{\theta_0,\eta_0}\big[\dot\ell_{\theta_0,\eta} - B_{\theta_0,\eta}\tilde b_n\big] + |\theta-\theta_0|^2\int_0^1 P_{\theta_0,\eta_0}\Big[-z^2e^{\theta_u z}H(t) + e^{\theta_0 z}\int_0^t \tilde b_n^2 e^{u(\theta-\theta_0)\tilde b_n}\,dH\Big](1-u)\,du.$$

The integral can be seen to converge to $-\int_0^1 \tilde I_{\theta_0,\eta_0}(1-u)\,du = -\frac{1}{2}\tilde I_{\theta_0,\eta_0}$. The proof of (12.19) is complete if we show that the expectation in the linear term is $o(n^{-1/2}+|\theta-\theta_0|)$, uniformly in $\eta\in\mathcal{H}_n$. Because $\tilde\ell_{\theta_0,\eta} = \dot\ell_{\theta_0,\eta} - B_{\theta_0,\eta}b_{\theta_0,h}$ is by its definition orthogonal in $\mathbb{L}_2(p_{\theta_0,\eta})$ to the range of $B_{\theta_0,\eta}$, we can write this expectation as

$$P_{\theta_0,\eta_0}\big[B_{\theta_0,\eta}(\tilde b_{\theta_0,h} - \tilde b_n)\big] + P_{\theta_0,\eta}\tilde\ell_{\theta_0,\eta}\Big[\frac{p_{\theta_0,\eta_0}}{p_{\theta_0,\eta}} - 1 - B_{\theta_0,\eta}(\eta_0-\eta)\Big].$$

Because $\|B_{\theta_0,\eta}b\|_\infty \lesssim \|b\|_\infty$, uniformly in η with uniformly bounded $\|H\|_\infty$, the first term can be seen to be bounded by $\|\tilde b_{\theta_0,h} - \tilde b_n\|_\infty d_H(p_{\theta_0,\eta}, p_{\theta_0,\eta_0})$, where the first factor is bounded by a multiple of $\delta_n + \|\tilde b_{\theta_0,h_0} - \tilde b_{\theta_0,h}\|_\infty$, by Lemma 12.14(x) and the second factor is bounded by a multiple of $|\theta-\theta_0| + d_H(p_{\theta,\eta}, p_{\theta_0,\eta_0})$. Hence this part of the linear term is $o(n^{-1/2}+|\theta-\theta_0|)$ if $\sqrt{n}(\delta_n+\epsilon_n)\epsilon_n \to 0$. The function in square brackets in the second term can be written

$$e^{\delta(\eta_0-\eta)(t) - e^{\theta_0 z}(H_0 - H(t))} - 1 - \delta(\eta_0-\eta)(t) + e^{\theta_0 z}\int_0^t (\eta_0-\eta)\,dH.$$

Applying the inequality $|e^x - 1 - x| \le (e^x \vee 1)x^2$, valid for any $x\in\mathbb{R}$, to the leading exponential and to the exponential in $H_0(t) - H(t) = \int_0^t (e^{\eta_0-\eta} - 1)\,dH$, the expectation of this function can be seen to be bounded by a multiple of $P_{\theta_0,\eta}\big[(e^{\eta_0-\eta}\vee 1)(\eta_0-\eta)^2\big] \lesssim \int(\eta-\eta_0)^2\,d(H+H_0)$. By Lemma 12.14(viii) and (vii)+(ix)+(iii), this is bounded above by $\|\eta-\eta_0\|_\infty^2 d_H^2(p_{\theta_0,\eta}, p_{\theta_0,\eta_0})$. Hence it is $o(n^{-1/2}+|\theta-\theta_0|)$ if $\sqrt{n}\, n\epsilon_n^2\epsilon_n^2 \to 0$. □

Example 12.13 (Riemann-Liouville) The concentration function of the released Riemann-Liouville process with parameter $\alpha > 0$ given in (11.17) at a function $w \in \mathfrak{C}^\beta[0,\tau]$ is proportional to $\epsilon^{-1/\alpha}$ if $\beta \ge \alpha$ and $\epsilon^{-(2\alpha-2\beta+1)/\beta}$ if $\alpha > \beta$ and $\alpha - \beta \notin \mathbb{N} + 1/2$, and otherwise equal to this function times a logarithmic factor. If $h_0 \in \mathfrak{C}^\beta[0,\tau]$, for some $\beta > 0$, and is bounded away from zero, then the requirement $\varphi_{h_0}(\epsilon_n) \lesssim n\epsilon_n^2$ results in the "usual" rate $\epsilon_n = n^{-\alpha\wedge\beta/(2\alpha+1)}$. In order that also $n^{3/2}\epsilon_n^4 \to 0$, it is necessary that $3/2 < \alpha < 4\beta/3 - 1/2$. The remaining two requirements $\epsilon_n^2\varphi_{\tilde b_{\theta_0,h_0}}(\delta_n) \to 0$ and $\delta_n\epsilon_n\sqrt{n}$ can be understood as giving a lower and upper bound for δ_n in terms of ϵ_n and n. If $\tilde b_{\theta_0,h_0} \in \mathfrak{C}^B[0,\tau]$ for some $B > 0$, then the lower bound is smaller than the upper bound if $B \ge \alpha$, or if $1/2 < B \le \alpha$ and $B + \beta > \alpha + 1/2$.

The restriction $\alpha < 4\beta/3 - 1/2$ arises through the “no-bias” condition. A prior on the nuisance parameter h of too large smoothness α relative to the true regularity β of h, may cause a bias in the posterior distribution for the parameter of interest θ.

The additional condition on B can be interpreted as requiring that the least-favorable direction be sufficiently smooth. If the censoring variable C and covariate Z are independent, then the least-favorable direction can be seen to have the same Hölder regularity $B = \beta$ as h_0 and the extra condition on B is automatic.

Lemma 12.14 *For every $\theta, \theta^* \in [-M, M]$, every pair of functions $h = e^{\eta}$, $h^* = e^{\eta^*}$, and $\lesssim$ and $\gtrsim$ denoting inequalities up to multiplicative constants that depend on $\|Z\|_\infty$ and M only, and $\|\eta\|_\infty = \sup_{0 \le t \le \tau} |\eta(t)|$, the following assertions hold:*

(i) $d_H(p_{\theta,h}, p_{\theta^*,h^*}) \le 2(\mathrm{E}Z^2)^{1/2}|\theta - \theta^*| + \|\eta - \eta^*\|_\infty$.

(ii) $K(p_{\theta,h}; p_{\theta^*,h^*}) \lesssim (1 + \|\eta - \eta^*\|_\infty + \|H\|_\infty + \|H^*\|_\infty) d_H^2(p_{\theta,h}, p_{\theta^*,h^*})$.

(iii) $V_2(p_{\theta,h}; p_{\theta^*,h^*}) \lesssim (1 + \|\eta - \eta^*\|_\infty + \|H\|_\infty + \|H^*\|_\infty)^2 d_H^2(p_{\theta,h}, p_{\theta^*,h^*})$.

(iv) $\|\log(p_{\theta^*,h^*}/p_{\theta,h})\|_\infty \lesssim 1 + \|\eta^* - \eta\|_\infty + \|H\|_\infty + \|H^*\|_\infty$.

(v) $\|H - H^*\|_\infty \le \|H\|_\infty \|\eta - \eta^*\|_\infty e^{\|\eta - \eta^*\|_\infty}$.

(vi) $\|(a, b)\|_{\theta,h}^2 \lesssim a^2 + \int b^2\, dH$.

(vii) $\|(a, b)\|_{\theta,h}^2 \gtrsim C(a^2 + \int b^2\, dH)$*, for every h such that $\|H\|_\infty \le K$ and a constant C that depends on K only.*

(viii) $\|b\|_{H,2} \lesssim C\|b\|_\infty d_H(p_{\theta,h}, p_{\theta^*,h^*}) \vee C\|b\|_{H^*,2}$*, for every h and h^* with $\|H\|_\infty \vee \|H^*\|_\infty \le K$ and a constant C that depends on K only.*

(ix) $\|\theta - \theta^*, \eta - \eta^*\|_{\theta,h}^2 \lesssim C\, V_2(p_{\theta,h}; p_{\theta^*,h^*})$*, for every h such that $\|H\|_\infty \le K$, for a constant C that depends on K only.*

(x) $\|\tilde{b}_{\theta,h} - \tilde{b}_{\theta^*,h}\|_\infty \lesssim C d_H(p_{\theta,h}, p_{\theta^*,h})$*, for every h such that $\|H\|_\infty \le K$, for a constant C that depends on K only.*

(xi) *If $d_H(p_{\theta,h}, p_{\theta_0,h_0}) \to 0$, for fixed θ_0 and h_0, then $\theta \to \theta_0$ and $\|H - H_0\|_\infty \to 0$.*

Proof (i). For $a = \theta^* - \theta$ and $b = \eta^* - \eta$ the path $(\theta_u, h_u) := (\theta + ua, e^{\eta + ub})$ is equal to (θ, h) at $u = 0$ and equal to (θ^*, h^*) at $u = 1$. The derivative of the path $u \mapsto p_{\theta_u,h_u}^{1/2}$ is equal to $\frac{1}{2}(a\dot{\ell}_{\theta_u,h_u} + B_{\theta_u,h_u} b)\, p_{\theta_u,h_u}^{1/2}$. Therefore, by the mean value theorem in Hilbert space,

$$\big\| p_{\theta,h}^{1/2} - p_{\theta^*,h^*}^{1/2} \big\|_2 \le \sup_{0 \le u \le 1} \big\| \tfrac{1}{2}(a\dot{\ell}_{\theta_u,h_u} + B_{\theta_u,h_u} b)\, p_{\theta_u,h_u}^{1/2} \big\|_2.$$

The right side is the information norm $\|(a, b)\|_{\theta_u,h_u}$, whose square is given as a sum of three terms in (12.17). Inserting the definitions of the functions $M_{k,\theta,h}$, we can write the square of $\|(a, b)\|_{\theta,h}$ as

$$\mathrm{E}_Z \int (a^2 Z^2 + 2ab(t)Z + b^2(t)) \bar{F}_{C|Z}(t|\, Z)\, e^{\theta Z} e^{-e^{\theta Z} H(t)}\, dH(t). \qquad (12.20)$$

Inequality (i) follows by bounding b and $\bar{F}_{C|Z}(t|\, Z)$ by $\|b\|_\infty$ and $\mathbb{1}\{t \le \tau\}$, and next evaluating the integral as $\int_0^\sigma e^{-u}\, du \le 1$, for $\sigma = e^{\theta Z} H(\tau)$.

(ii), (iii) and (iv) The log-likelihood ratio satisfies, for any $\theta, \theta^*, h = e^{\eta}, h^* = e^{\eta^*}$,

$$\log \frac{p_{\theta^*,h^*}}{p_{\theta,h}}(t, \delta, z) = \delta((\theta^* - \theta)z + (\eta^* - \eta)(t)) + H(t)e^{\theta z} - H^*(t)e^{\theta^* z}.$$

This is bounded as in (iv). Next (ii) and (iii) follow by the general Lemma B.2.

(v). Since $|e^x - e^y| \le e^{x\vee y}|x - y|$, and $x \vee y \le x + |x - y|$, for every $x, y \in \mathbb{R}$, $|H(t) - H^*(t)| \le \int_0^t e^{\eta(s)} e^{|\eta(s)-\eta^*(s)|}|\eta(s) - \eta^*(s)|\, ds$. The result follows.

(vi) and (vii). These are clear from (12.20).

(viii). By the inequality $g - g^* \le 2\sqrt{g}(\sqrt{g} - \sqrt{g^*})$, for any $g, g^* \ge 0$,

$$\begin{aligned} &\mathrm{E}_Z \int b^2\, (e^{-e^{\theta Z}H} e^{\theta Z}\, dH - e^{-e^{\theta^* Z}H^*} e^{\theta^* Z}\, dH^*) \\ &\quad\le \|b\|_\infty \mathrm{E} \int \Bigl[|b|\, 2e^{-e^{\theta Z}H/2} e^{\theta Z/2}\sqrt{h}(e^{-e^{\theta Z}H/2} e^{\theta Z/2}\sqrt{h} - e^{-e^{\theta^* Z}H^*/2} e^{\theta^* Z/2}\sqrt{h^*})\Bigr]\, d\nu \\ &\quad\lesssim \|b\|_\infty \|b\|_{H,2} d_H(p_{\theta,h}, p_{\theta^*,h^*}). \end{aligned}$$

The factors $e^{-e^{\theta Z}H}$ and $e^{-e^{\theta^* Z}H^*}$ are bounded away from zero if $\|H\|_\infty$ and $\|H^*\|_\infty$ are bounded above. We conclude that $\int b^2\, dH$ is bounded by a multiple of the maximum of $\int b^2\, dH^*$ and the right side of the display. This is equivalent to assertion (viii).

(ix). We have

$$\begin{aligned} V_2(p_{\theta,h}, p_{\theta^*,h^*}) &= \mathrm{E} \int (H(t)e^{\theta Z} - H^*(t)e^{\theta^* Z})^2 e^{-e^{\theta Z}H(t)}\, dF_{C|Z}(t|\, Z) \\ &\quad + \mathrm{E}\int \bigl[(\theta^* - \theta)Z + (\eta^* - \eta)(t) + H(t)e^{\theta Z} - H^*(t)e^{\theta^* Z}\bigr]^2 \bar{F}_{C|Z}(t|\, Z) e^{-e^{\theta Z}H(t)} e^{\theta Z}\, dH(t). \end{aligned}$$

The two integrals are bounded below by constants, depending on h and the distribution of (C, Z) only, times the integrals obtained by replacing the measures given by $e^{-e^{\theta Z}H(t)}\, dF_{C|Z}(t|\, Z)$ and $\bar{F}_{C|Z}(t|\, Z)e^{-e^{\theta Z}H(t)}e^{\theta Z}\, dH(t)$ by the Lebesgue measure on $[0, \tau]$. Since $\|g + g^*\|^2 + \|g^*\|^2 \ge \frac{1}{2}\|g^*\|^2$, for any g and g^*, it follows that $V_2(p_{\theta,h}, p_{\theta^*,h^*})$ is bounded below by a multiple of $\mathrm{E}\int_0^\tau ((\theta^* - \theta)Z + (\eta^* - \eta)(t))^2\, dt$. We combine this with (vi).

(x). The functions $M_{k,\theta,h}$ can be seen to be bounded away from zero and infinity uniformly in bounded $\|H\|_\infty$, and $\|M_{k,\theta,h} - M_{k,\theta^*,h}\|_\infty \lesssim d_{TV}(P_{\theta,h}, P_{\theta^*,h})$.

(xi). Because the $\mathbb{L}_1$-distance is bounded by twice the Hellinger distance,

$$\mathrm{E}_Z \int |e^{\theta Z}h(t)e^{-e^{\theta Z}H(t)} - e^{\theta_0 Z}h_0(t)e^{-e^{\theta_0 Z}H_0(t)}|\, \bar{F}_{C|Z}(t|\, Z)\, dt \to 0.$$

The integral $\int_0^t e^{\theta z}h(s)e^{-e^{\theta z}H(s)}\, ds$ can be explicitly evaluated as $\exp(-e^{\theta z}H(t))$, and the survival function $\bar{F}_{C|Z}(t|\, z)$ is bounded away from zero on $[0, \tau]$, for all z in a set $\mathfrak{Z}$ of positive measure under the law of Z. We conclude, that for all $A \subset \mathfrak{Z}$ and all $t \in [0, \tau]$,

$$\mathrm{E}\bigl[\mathbb{1}_A(Z)e^{-e^{\theta Z}H(t)}\bigr] \to \mathrm{E}\bigl[\mathbb{1}_A(Z)e^{-e^{\theta_0 Z}H_0(t)}\bigr].$$

Because $e^{\theta_0 Z}H_0(t)$ is uniformly bounded, the right side is contained in a compact interval in $(0, \infty)$. Since $e^{-e^{\theta Z}H(t)} \le e^{-MH(t)}$, for some $M > 0$, it follows that $H(t)$ must be contained in a bounded interval, uniformly in $t \in [0, \tau]$. Thus we can apply Helly's theorem to select from any sequence of (θ, H) with $\|p_{\theta,h} - p_{\theta_0,h_0}\|_1 \to 0$ a subsequence (θ_m, H_m) with $\theta_m \to \tilde{\theta}$ and $H_m \rightsquigarrow \tilde{H}$, for some $\tilde{\theta}$ and $\tilde{H}$. Then the left side of the preceding display at (θ_m, H_m) converges to its value at $(\tilde{\theta}, \tilde{H})$, for all continuity points $t \in [0, \tau]$ of $\tilde{H}$, and hence

$\mathrm{E}\big[\mathbb{1}_A(Z)e^{-e^{\tilde\theta Z}\tilde H(t)}\big]=\mathrm{E}\big[\mathbb{1}_A(Z)e^{-e^{\theta_0 Z}H_0(t)}\big]$, for all such t and all A as before. This implies that $e^{\tilde\theta z}\tilde H(t)=e^{\theta_0 z}H_0(t)$, for all such t and almost all $z\in\mathfrak{Z}$. Since Z is nondegenerate on $\mathfrak{Z}$, this is possible only if $\tilde\theta=\theta_0$. Then also $\tilde H=H_0$, and the convergence $H_m\rightsquigarrow H_0$ is uniform on every interval $[0,t]$ with $t<\tau$, by the continuity of H_0. We can extend to the full interval $[0,\tau]$ by also considering the second, "censored" part of the sum that defines $\|p_{\theta,h}-p_{\theta_0,h_0}\|_1$, where we use that $F_{C|Z}(\{\tau\}|\,z)>0$, for z in a set of positive probability. □

12.4 White Noise Model

In the white noise model, discussed previously in Example 8.6 and Section 8.3.4, we observe a sequence $X^{(n)}=(X_{n,1},X_{n,2},\ldots)$ of independent variables $X_{n,i}\overset{\text{ind}}{\sim}\mathrm{Nor}(\theta_i,n^{-1})$. A conjugate analysis in this simple problem gives insight into the infinite-dimensional Bernstein–von Mises theorem.

The parameter $\theta=(\theta_1,\theta_2,\ldots)$ belongs to $\mathbb{R}^\infty$, but it is often restricted to a small subset of sequences with $\theta_i\to0$ as $i\to\infty$. This reflects the fact that the coordinates θ_i typically arise as the Fourier coefficients of a function, in which case they minimally satisfy that $\theta\in\ell_2$, and often are further restricted to belong to a smoothness class, such as the Sobolev space $\mathfrak{W}^\beta$ of all sequences with $\|\theta\|^2_{2,2,\beta}:=\sum_{i=1}^\infty i^{2\beta}\theta_i^2<\infty$, for some $\beta>0$.

A Gaussian prior of the type $\theta_i\overset{\text{ind}}{\sim}\mathrm{Nor}(0,\tau_i^2)$ is conjugate, and leads to the posterior distribution

$$\theta_i\,|\,X^{(n)}\overset{\text{ind}}{\sim}\mathrm{Nor}\Big(\frac{\tau_i^2X_{n,i}}{n^{-1}+\tau_i^2},\frac{n^{-1}\tau_i^2}{n^{-1}+\tau_i^2}\Big).\tag{12.21}$$

Any prior variances $\tau_i^2>0$ lead to a proper prior with full support on $\mathbb{R}^\infty$. Prior variances τ_i^2 that decrease to 0 as $i\to\infty$ are natural to model parameters θ with decreasing coordinates. For instance, the choice $\tau_i^2=i^{-1-2\alpha}$, for some fixed $\alpha>0$, leads to a prior Π with $\Pi(\mathfrak{W}^\beta)=1$ whenever $\beta<\alpha$, and $\Pi(\mathfrak{W}^\beta)=0$ if $\beta\geq\alpha$. In Example 8.6 these priors were seen to lead to the posterior contraction rate $n^{-(\alpha\wedge\beta)/(2\alpha+1)}$ if the true parameter θ belongs to $\mathfrak{W}^\beta$, with the (embarrassing) finding that the fastest rate is obtained by the prior with $\alpha=\beta$, which gives prior mass zero to $\mathfrak{W}^\beta$.

In this section we study the Bernstein–von Mises phenomenon, first for the full posterior distribution, and next for the induced marginal posterior distributions of linear functionals of the parameter. We assume throughout that $\tau_i>0$, for every i.

12.4.1 Full Parameter

In the white noise model with an infinite-dimensional parameter set, such as the Sobolev space $\mathfrak{W}^\beta$, there is no notion of an efficient asymptotically normal estimator, and a maximum likelihood estimator fails to exist. This makes the formulation of a Bernstein–von Mises theorem not obvious, in particular regarding the centering variables. We might take a general Bernstein–von Mises theorem to make the assertion that, for given centering variables $\hat\theta=\hat\theta(X^{(n)})$, the frequentist distribution of $\hat\theta-\theta\,|\,\theta$ and the Bayesian distribution of $\theta-\hat\theta\,|\,X^{(n)}$ approach each other as $n\to\infty$, for a given true θ that governs the distribution of $\hat\theta$ in the first case and the distribution of the conditioning variable $X^{(n)}$ in the Bayesian case. In the finite-dimensional case choosing $\hat{\theta}$ the maximum likelihood estimator is typical, but the posterior mean is also possible. If presently we choose $\hat{\theta}$ equal to the posterior mean, then the two relevant distributions, scaled by $\sqrt{n}$ for convenience, are given by

$$\sqrt{n}(\mathrm{E}(\theta|\, X^{(n)}) - \theta)|\,\theta \quad \sim \bigotimes_{i=1}^{\infty} \mathrm{Nor}\Big(-\frac{n^{-1/2}\theta_i}{n^{-1}+\tau_i^2}, \frac{\tau_i^4}{(n^{-1}+\tau_i^2)^2}\Big), \tag{12.22}$$

$$\sqrt{n}(\theta - \mathrm{E}(\theta|\, X^{(n)}))|\, X^{(n)} \sim \bigotimes_{i=1}^{\infty} \mathrm{Nor}\Big(0, \frac{\tau_i^2}{n^{-1}+\tau_i^2}\Big). \tag{12.23}$$

These two distributions differ both in mean (unless $\theta = 0$) and variance, where the mean of the Bayesian distribution (12.23) vanishes by our choice of centering statistic. In general these differences do not disappear as $n \to \infty$. In fact, these infinite product measures (on $\mathbb{R}^\infty$) are orthogonal (and hence at maximum distance) unless their variances are sufficiently close, which is possible only if τ_i^2 tends to infinity fast enough.

Theorem 12.15 (Bernstein–von Mises) *For any $\theta \in \mathbb{R}^\infty$ the following assertions are equivalent:*

(i) *The total variation distance between the measures on the right sides of* (12.22) *and* (12.23) *tends to* 0 *as $n \to \infty$.*
(ii) $\sum_{i=1}^\infty (1/\tau_i^4) < \infty$ *and* $\sum_{i=1}^\infty (\theta_i^2/\tau_i^4) < \infty$.

If these conditions hold, then both measures as in (i) tend to $\otimes_{i=1}^\infty \mathrm{Nor}(0, 1)$.

Proof Convergence in total variation distance to zero is equivalent to convergence in the Hellinger distance to zero, and hence equivalent to the Hellinger affinities between the measures converging to 1. As the measures are product measures, this translates into the convergence

$$\prod_{i=1}^\infty \rho_{1/2}\Big(\mathrm{Nor}\Big(0, \frac{\tau_i^2}{n^{-1}+\tau_i^2}\Big), \mathrm{Nor}\Big(-\frac{n^{-1/2}\theta_i}{n^{-1}+\tau_i^2}, \frac{\tau_i^4}{(n^{-1}+\tau_i^2)^2}\Big)\Big) \to 1.$$

This can be shown to be equivalent to the conditions under (ii). The final statement follows similarly by computing affinities. □

Condition (ii) of the theorem requires that the prior variances τ_i^2 tend to infinity (at a rate), which goes in the opposite direction of the smoothing priors with $\tau_i^2 \downarrow 0$ that were argued to be natural. Thus the theorem may be safely remembered as saying that a Bernstein–von Mises theorem for the full parameter holds only for unrealistic priors, at least in the sense of (i) of the theorem. It may also be noted that the centering $\hat{\theta}$ hardly influences this conclusion, because even if $\theta = 0$, when both distributions in (12.22) and (12.23) are centered at zero, it is still required that $\tau_i^2 \to \infty$.

This negative conclusion can be attributed to the infinite-dimensional setting, where "optimal" estimation of a parameter typically involves a bias-variance trade-off, with the bias controlled through a priori assumptions. More precisely, the conclusion can be linked to the use of the total variation distance, which measures the infinitely many dimensions equally. In the next subsection we shall see that the difficulties may disappear if the full posterior is projected onto finite-dimensional marginals. Alternatively, it is also possible to keep the infinite-dimensional formulation, but down-weight the increasing dimensions. To set this up, consider a positive weight sequence $(w_i) \in \ell_1$, and the Hilbert space $H_w = \{\theta \in \mathbb{R}^\infty \colon \sum_i w_i\theta_i^2 < \infty\}$, with inner product

$$\langle \theta, \theta' \rangle_w = \sum_{i=1}^{\infty} w_i \theta_i \theta_i'.$$

As necessarily $w_i \to 0$, the space H_w is bigger than ℓ_2 and it has a weaker norm. The space is big enough to carry the law $\otimes_{i=1}^{\infty}\mathrm{Nor}(0,1)$ of an infinite sequence $(Z_1, Z_2, \ldots)$ of independent standard normal variables (as $\sum_i w_i Z_i^2 < \infty$ a.s.), which arises as the limit in the preceding theorem. The measures on the right sides of (12.22) and (12.23) also concentrate on H_w, and they both converge to the latter law with respect to the weak topology, under mild conditions.

Theorem 12.16 (Bernstein–von Mises, weak) *Fix $w \in \ell_1$. For every θ with $\sum_i w_i\theta_i^2/\tau_i^2 < \infty$, the measures on the right sides of* (12.22) *and* (12.23) *tend to* $\otimes_{i=1}^{\infty}\mathrm{Nor}(0,1)$ *relative to the weak topology on the space of Borel probability measures on H_w. In particular, this is true for almost every θ from* $\otimes_{i=1}^{\infty}\mathrm{Nor}(0,\tau_i^2)$.

Proof The distribution on the right of (12.22) can be represented as the distribution of the vector Z_n with coordinates $Z_{n,i} = -n^{-1/2}\theta_i/(n^{-1}+\tau_i^2) + Z_i\tau_i^2/(n^{-1}+\tau_i^2)$, for $Z = (Z_1, Z_2, \ldots)$ a sequence of independent standard normal variables. Under the stated conditions $\|Z_n - Z\|_w \to 0$ almost surely, by the dominated convergence theorem, whence $Z_n \rightsquigarrow Z$. Convergence of the distributions on the right side of (12.23) follows similarly. □

Thus an infinite-dimensional Bernstein–von Mises theorem does hold provided the mode of approximation is chosen weak enough.

12.4.2 Linear Functionals

The posterior distribution of a real-valued functional $L\theta$ of the parameter $\theta = (\theta_1, \theta_2, \ldots)$ is the marginal distribution $\Pi_n(\theta\colon L\theta \in \cdot \mid X^{(n)})$ of the full posterior distribution of θ, described in (12.21). A continuous, linear functional $L\colon \ell_2 \to \mathbb{R}$ can be represented by an element $l \in \ell_2$ in the form $L\theta = \sum_i l_i\theta_i$, for $\theta \in \ell_2$. If the prior is concentrated on ℓ_2 (i.e. $\sum_i \tau_i^2 < \infty$), then it is immediate from (12.21) that

$$L\theta \mid X^{(n)} \sim \mathrm{Nor}\Big(\sum_i \frac{l_i\tau_i^2 X_{n,i}}{n^{-1}+\tau_i^2}, \sum_i \frac{l_i^2 n^{-1}\tau_i^2}{n^{-1}+\tau_i^2}\Big). \tag{12.24}$$

This formula is true more generally for *measurable linear functionals relative to the prior* $\otimes_i\mathrm{Nor}(0,\tau_i^2)$. These are defined as Borel measurable maps $L\colon \mathbb{R}^\infty \to \mathbb{R}$ that are linear on a measurable linear subspace of probability one under the prior. For T the coordinatewise multiplication $\theta \mapsto (\tau_1\theta_1, \tau_2\theta_2, \ldots)$ by the prior standard deviations, the map $LT\colon \ell_2 \to \mathbb{R}$ is then automatically continuous and linear, whence representable by some element $m \in \ell_2$, and formula (12.24) is true with $l_i = m_i/\tau_i$. (This satisfies $\sum_i \tau_i^2 l_i^2 < \infty$, but need not be contained in ℓ_2; see Knapik et al. 2011, Proposition 3.2, and Skorohod 1974, pages 25–27, for details.)

The marginal normality (12.24) of the posterior distribution provides one aspect of a Bernstein–von Mises theorem for $L\theta$. However, the theorem ought to include a relationship between the frequentist and Bayesian distributions. For $\widehat{L\theta} := \mathrm{E}(L\theta|\, X^{(n)})$ the posterior mean we say that in the current setting the Bernstein–von Mises theorem *holds* at θ if the total variation distance between the laws of the following variables tends to zero

$$\sqrt{n}(\widehat{L\theta} - L\theta)|\,\theta \sim \mathrm{Nor}\Big(-\sum_i \frac{l_i n^{-1/2}\theta_i}{n^{-1}+\tau_i^2}, \sum_i \frac{l_i^2\tau_i^4}{(n^{-1}+\tau_i^2)^2}\Big),$$

$$\sqrt{n}(L\theta - \widehat{L\theta})|\, X^{(n)} \sim \mathrm{Nor}\Big(0, \sum_i \frac{l_i^2\tau_i^2}{n^{-1}+\tau_i^2}\Big).$$

Here the first law is the frequentist law of the centered posterior mean, considered under θ-probability, while the second is the posterior distribution of $L\theta$, centered at (posterior) mean zero. The scaling by $\sqrt{n}$ is for convenience. Both distributions are Gaussian, and hence their total variation distance tends to zero if the quotient of their two variances tends to 1 and the square difference between their means is negligible relative to this variance. This depends on the combination of l, the prior variances, and θ. The following theorem gives some cases of interest.

Theorem 12.17 (Bernstein–von Mises, functionals) *Let l represent L as indicated.*

(i) *If $l \in \mathfrak{W}^q$ for some $q \geq 0$, and $\tau_i^2 = i^{-2r}$ for some $r > 0$, and $\beta + q > r$, then the Bernstein–von Mises theorem holds at any $\theta \in \mathfrak{W}^\beta$.*

(ii) *If $l \in \ell_1$, then the Bernstein–von Mises theorem holds at almost every θ sampled from $\otimes_{i=1}^\infty \mathrm{Nor}(0, \tau_i^2)$, for any τ_i.*

(ii) *If $l \in \ell_2$, then the Bernstein–von Mises theorem holds in probability if θ is sampled from $\otimes_{i=1}^\infty \mathrm{Nor}(0, \tau_i^2)$, for any τ_i.*

(iv) *If $l_i \asymp i^{-q-1/2}S(i)$ for a slowly varying function S, and $\tau_i^2 = i^{-2r}$ for some $r > 0$, then the Bernstein–von Mises theorem holds at some θ only if $q \geq 0$, in which case $n \mapsto n\,\mathrm{var}(\theta|\, X^{(n)})$ is slowly varying.*

Proof As indicated before the statement of the theorem the Bernstein–von Mises theorem holds if the frequentist variance of the posterior mean is asymptotically equivalent to the variance of the posterior distribution and the square bias of the posterior mean as an estimator of $L\theta$ vanishes relative to this common variance. These quantities are given in the display and hence we are lead to the pair of conditions

$$\Big|\sum_{i=1}^\infty \frac{l_i n^{-1/2}\theta_i}{n^{-1}+\tau_i^2}\Big|^2 \ll \sum_{i=1}^\infty \frac{l_i^2\tau_i^2}{n^{-1}+\tau_i^2}, \quad \text{and} \quad \sum_{i=1}^\infty \frac{l_i^2\tau_i^4}{(n^{-1}+\tau_i^2)^2} \sim \sum_{i=1}^\infty \frac{l_i^2\tau_i^2}{n^{-1}+\tau_i^2}.$$

If $l \in \ell_2$, then the last two series, which provide the variances, tend to $\|l\|_2^2$, by the dominated convergence theorem, whence the variances behave as they should, under every of the assumptions (i)–(iii).

If $\theta \in \mathfrak{W}^\beta$, then the leftmost expression in the display, the square bias, can be bounded above by $\{\sum_i l_i^2 n^{-1} i^{-2\beta}/(n^{-1}+\tau_i^2)^2\}\|\theta\|_{2,2,\beta}^2$, by the Cauchy-Schwarz inequality. For $\tau_i^2 = i^{-2r}$, this is seen to be of order $nn^{-(\beta+q)/r\wedge 2}$ by Lemma K.7, and hence $o(1)$ if $\beta + q > r$.

If θ is sampled from the prior, then the mean of the bias series $\sum_i l_i n^{-1/2}\theta_i/(n^{-1} + \tau_i^2)$ vanishes and its variance is equal to $\sum_i l_i^2 n^{-1}\tau_i^2/(n^{-1} + \tau_i^2)^2$. If $l \in \ell_2$, then the latter is $o(1)$, by the dominated convergence theorem, as $n^{-1}\tau_i^2/(n^{-1} + \tau_i^2)^2$ is bounded by 1 and tends to zero as $n \to \infty$, for every i. This gives the Bernstein–von Mises theorem in probability (iii). If $l \in \ell_1$, then we bound the absolute value of the bias series above by $\sum_i |l_i n^{-1/2}\tau_i Z_i|/(n^{-1} + \tau_i^2)$, for $Z_i = \theta_i/\tau_i$. The terms of this series are bounded above by the terms $|l_i Z_i|$ of the converging series $\sum_i |l_i Z_i|$ and tend to zero for every i, as $n \to \infty$, surely. Thus $\sum_i |l_i n^{-1/2}\tau_i Z_i|/(n^{-1} + \tau_i^2) \to 0$, almost surely, by the dominated convergence theorem.

For the proof of (iv) we first note that the frequentist variance (given by the third series in the display) is always smaller than the posterior variance, as the terms of the two series differ by factors $\tau_i^2/(n^{-1} + \tau_i^2) \leq 1$. For i such that $n\tau_i^2 \leq c$ for some $c > 0$, the extra factor is strictly bounded away from 1. This implies that the quotient of the two series can converge to 1 only if

$$\sum_{i:n\tau_i^2\leq c} \frac{l_i^2\tau_i^2}{n^{-1}+\tau_i^2} \ll \sum_{i=1}^{\infty} \frac{l_i^2\tau_i^2}{n^{-1}+\tau_i^2}.$$

For l satisfying the assumption in (iv) and $\tau_i^2 = i^{-2r}$, the last assertion of Lemma K.8 gives that $(2r + 2q)/(2r) \geq 1$, i.e. $q \geq 0$. The same lemma then gives that the series on the right is a slowly varying function of n (namely $\sum_{i\leq n^{1/(2r)}} S^2(i)/i$ if $q = 0$ and converging to a constant if $q > 0$). □

Parts (ii) and (iii) are comforting to the Bayesian who really believes his prior: the approximation holds for a large class of functionals at almost every true parameter deemed possible by the prior. Part (iv) suggests that the condition that $l \in \ell_2$ cannot be much relaxed (although the theorem may hold for l with $q = 0$ that are "almost but not really in ℓ_2"). Perhaps the most interesting part is (i), which shows the interplay between the regularity β of the true parameter, the smoothness q of the functional, and the regularity of the prior. If $r = \alpha + 1/2$, then the condition becomes $\beta + q > \alpha + 1/2$ and can be interpreted as saying that the prior smoothness α should not exceed the smoothness of the true parameter plus the smoothness of the functional minus 1/2.

It is shown in the preceding proof that under conditions (i)–(iii) the posterior variance and the variance of the posterior mean are of order n^{-1}. The rate of posterior contraction of the marginal distribution of $L\theta$ is then the "parametric rate" $n^{-1/2}$. Under condition (iv) this is true, possibly up to a slowly varying factor. The fast rate of estimation in these cases is possible by the assumed smoothness of the functional L and parameter θ. It is known that for a functional that is "regular" of (exact) order $q < 0$, the minimax rate over balls in $\mathfrak{W}^\beta$ is equal to $n^{-(\beta+q)/(2\beta)} \gg n^{-1/2}$. The following proposition shows that this is the posterior contraction rate for a prior with variances $\tau_i^2 = i^{-1-2\alpha}$ if $\alpha = \beta - 1/2$. (See Knapik et al. 2011 for further discussion and a proof.) This is in contrast with the "parametric case": there any prior that is not overly smooth (precisely: $\alpha < \beta + q - 1/2$) performs well, whereas in the "nonparametric case" only a single prior smoothness is optimal.

Proposition 12.18 (Contraction rate) *If $l \in \mathfrak{W}^q$ and $\theta_0 \in \mathfrak{W}^\beta$ for some $q \geq -\beta$ and $\beta > 0$, then the marginal posterior contraction rate relative to the prior with $\tau_i^2 = i^{-1-2\alpha}$ is given by $\epsilon_n = n^{-(\beta\wedge(\alpha+1/2)+q)/(2\alpha+1)} \vee n^{-1/2}$: for any $M_n \to \infty$,*

$$\mathrm{E}_{\theta_0}\Pi_n(\theta\colon |L\theta - L\theta_0| \geq M_n\epsilon_n |\, X^{(n)}) \to 0.$$

12.5 Historical Notes

Versions of Theorem 12.2 were proved by Lo (1983) and Lo (1986). He treated the classical Donsker classes consisting of indicator functions, and proved tightness by verifying Markov properties and increment bounds; see Problems L.3 and L.4. For the classical Donsker classes Theorem 12.2 is also a special case of Corollary 13.23, which allows more general priors and censored data. A thorough treatment of the theory of strong approximation is given in Csörgő and Révész (1981). The original references Komlós et al. (1975); Csörgő and Révész (1975) were improvements of work by Strassen and Kiefer. Theorems 12.5 and 12.6 are Theorems 2.1 and Proposition 5.7 of Lo (1987). The proof of the latter result is similar to the results in the non-Bayesian setting due to Bickel and Rosenblatt (1973). The main results of Section 12.3 on the semiparametric Bernstein–von Mises theorem are based on Castillo (2012b). We have extended his main result from the strict semiparametric case to functionals, allow general priors, and have formulated the theorem in terms of an LAN expansion of a least favorable submodel, somewhat parallel to the treatment of maximum (or profile) likelihood in Murphy and van der Vaart (2000). The original formulation in Castillo (2012b) used a full LAN expansion as in Section 12.3.1, and was restricted to Gaussian priors. The notation and general discussion of semiparametric models (including the Cox model) are taken from Chapter 25 of van der Vaart (1998). The formulas for the information in the Cox model go back to Begun et al. (1983). Castillo discovered the role of the shift in the prior on the nuisance parameter in the least-favorable direction, a property that is underemphasized in Bickel and Kleijn (2012). For further results see Rivoirard and Rousseau (2012) and Castillo and Rousseau (2015). Castillo (2014) used nonparametric Bernstein–von Mises theorem results to derive posterior contraction rates with respect to the stronger uniform norm for the Gaussian white noise model and density estimation using wavelet series. Theorem 12.15 is Lemma 2 of Leahu (2011) and Theorem 12.16 is a slight extension of his Theorem 1. Leahu (2011) clarified and generalized previous work by Cox (1993) and Freedman (1999). Other versions of the weak infinite-dimensional Bernstein–von Mises Theorem 12.16 were recently developed by Castillo and Nickl (2013), Castillo and Nickl (2014) and Ray (2014). The main results in Section 12.4.2 are based on Knapik et al. (2011), who also consider inverse problems and scaling of the priors, and applied the results to study the coverage of credible sets. The almost sure part (ii) of Theorem 12.17 is due to Leahu (2011).

Problems

12.1 Show that $\sqrt{a_n + b_n}(V_n - \mathrm{E}(V_n)) \rightsquigarrow \mathrm{Nor}(0, \lambda(1-\lambda))$, whenever $V_n \sim \mathrm{Be}(a_n, b_n)$, and $a_n, b_n \to \infty$ in such a way that $a_n/(a_n + b_n) \to \lambda$. Use this to give a direct proof of the Bernstein–von Mises theorem for $P(A)$ based on observing a random sample from the Dirichlet process. [Hint: use a representation by gamma variables, and the delta-method.]

12.2 (James 2008) Let $P \sim \mathrm{DP}(\alpha)$ and let $\mathcal{F}$ be a Donsker class of functions with square integrable envelope with respect to α. Show that $\sqrt{|\alpha|}(P - \bar{\alpha}) \rightsquigarrow \mathbb{G}$ in $\mathfrak{L}_\infty(\mathcal{F})$, as $|\alpha| \to \infty$, where $\mathbb{G}$ is an $\bar{\alpha}$-Brownian bridge indexed by $\mathcal{F}$.

12.3 (Kiefer process) Given a Kiefer process $\mathbb{K}$, show that:

(a) $s \mapsto \mathbb{K}(s, n+1) - \mathbb{K}(s, n)$ are independent Brownian bridges, for $n \in \mathbb{N}$.
(b) $t \mapsto \mathbb{K}(s,t)/\sqrt{s(1-s)}$ is a standard Brownian motion, for any $0 < s < 1$.
(c) $\mathbb{K}$ is equal in distribution to the process $W(s,t) - sW(1,t)$, for W a two-parameter standard Wiener process (which has covariance kernel $\min(s_1, s_2)\min(t_1, t_2)$).

12.4 (Gu and Ghosal 2008) The receiver operating characteristic (ROC) function of two cumulative distribution functions F and G is defined by $R(t) = \bar{G}(\bar{F}^{-1}(t))$, for $t \in [0,1]$, and the area under the curve (AUC) is $A = \int_0^1 R(t)\,dt$. The ROC is invariant under continuous, strictly increasing transformations.The empirical ROC $\mathbb{R}_{m,n}$ and AUC $\mathbb{A}_{m,n}$ of two independent samples $X_1, \ldots, X_m \overset{\mathrm{iid}}{\sim} F$ and $Y_1, \ldots, Y_n \overset{\mathrm{iid}}{\sim} G$ are obtained by replacing F and G by their empirical versions. Induce the Bayesian bootstrap distribution of R from those of F and G. Assume that the true distributions F_0 and G_0 possess continuous densities f_0 and g_0, and that the functions $F_0\bar{F}_0|f_0'|/f_0^2$ and $G_0\bar{G}_0|g_0'|/g_0^2$ are bounded. Show that there exist independent Kiefer processes K_1 and K_2 such that a.s. as $m/(m+n) \to \lambda \in (0,1)$ and $\alpha_m \gg m^{-1/4}$, the Bayesian bootstrap distribution of R can be represented as

$$R(t) = \mathbb{R}_{m,n}(t) + \frac{1}{m}R_0'(t)K_1(t,m) + \frac{1}{n}K_2(R_0(t), n) + O\Big(\frac{(\log m)^{1/2}(\log\log m)^{1/4}}{m^{3/4}\alpha_m}\Big),$$

uniformly on $t \in (\alpha_m^*, 1 - \alpha_m^*)$, where $\alpha_m^* = \alpha_m + m^{-1/2}\sqrt{\log\log m}$. Furthermore, show that the AUC functional satisfies the Bernstein–von Mises theorem:

$$\sqrt{m+n}(A - \mathbb{A}_{m,n})|\, X_1, \ldots, X_m, Y_1, \ldots, Y_n \rightsquigarrow \mathrm{Nor}(0, \sigma^2), \qquad \text{a.s.},$$

as $m, n \to \infty$, where

$$\sigma^2 = \int_0^1 \int_0^1 \Big[\frac{1}{\lambda}R_0'(s)R_0'(t)[s \wedge t - st] + \frac{1}{(1-\lambda)}[R_0(s) \wedge R_0(t) - R_0(s)R_0(t)]\Big]\,dt\,ds,$$

and the ranges of the integrals in the definitions of A and $\mathbb{A}_{m,n}$ are taken $(\alpha_m^*, 1 - \alpha_m^*)$ rather than $(0,1)$.

12.5 (Castillo 2012b) Consider the Gaussian white noise model $dX_t^{(n)}(t) = f(t - \theta) + n^{-1/2}dB_t$, for $t \in [-\frac{1}{2}, \frac{1}{2}]$, where the signal $f \in \mathbb{L}_2[-\frac{1}{2}, \frac{1}{2}]$, symmetric about zero, 1-periodic and satisfies $\int_0^1 f(t)\,dt = 0$. Let $f_k = \sqrt{2}\int f(t)\cos(2\pi kt)\,dt$, $k \in \mathbb{N}$, be the

Fourier coefficients. Assume that the true signal f_0 satisfies $\sum_{k=1}^{\infty} k^{2\beta} f_k^2 < \infty$. Let the prior density on θ be positive and continuous, and let f be given a prior through $f_k \overset{\text{ind}}{\sim} \text{Nor}(0, k^{-2\alpha-1})$. Show that the Bernstein–von Mises theorem for θ holds if $\beta > 3/2$ and $1 + \sqrt{3}/2 < \alpha < (3\beta - 2)/(4 - 2\beta)$.

If the prior for f is modified to $f_k \overset{\text{ind}}{\sim} \text{Nor}(0, k^{-2\alpha-1})$ for $k \le \lfloor n^{1/(1+2\alpha)} \rfloor$ and $f_k = 0$ otherwise, show that the Bernstein–von Mises theorem for θ holds if $\beta > \frac{3}{2}$ and $3/2 < \alpha < (3\beta - 2)/(4 - 2\beta)$.

12.6 (Castillo 2012b) Consider a pair of functional regression models $Y_i \overset{\text{ind}}{\sim} \text{Nor}(f(i/n), 1)$ and $Z_i \overset{\text{ind}}{\sim} \text{Nor}(f(i/n - \theta), 1)$, $i = 1, \ldots, n$, θ lies in a compact interval $[-c_0, c_0] \subset (-1/2, 1/2)$, and f is square integrable and 1-periodic. Assume that the true f_0 has complex Fourier coefficients $\hat{f}_0(k) = \int e^{-2\pi i k t} f(t)\, dt$ satisfying $\hat{f}_0(0) = 0$, $\hat{f}_0(1) \neq 0$, $\sum_{k\in\mathbb{Z}} k|\hat{f}_0(k)| < \infty$ and $\sum_{k\in\mathbb{Z}} k^{2\beta}|\hat{f}_0(k)|^2 < \infty$. Put a prior density on θ which is positive and continuous, and on f, define a prior by the relations $f = Z_0 + \sum_{j=1}^{k_n} \big[Z_j \cos(2\pi j t) + W_j \sin(2\pi j t)\big]$, $Z_0 \sim \text{Nor}(0, 1)$, $Z_j, W_j \overset{\text{ind}}{\sim} \text{Nor}(0, j^{-(1+2\alpha)})$, $j = 1, \ldots, k_n := \lfloor n^{1/(1+2\alpha)} \rfloor$. Show that the Bernstein–von Mises theorem holds for θ if $\beta > 3/2$ and $3/2 < \alpha < 2\beta - 3/2$.

12.7 (Castillo 2012a) Consider two independent Gaussian white noise models $dX(t) = f(t)\, dt + n^{-1/2}\, dB_1(t)$ and $dY(t) = f(t - \theta)\, dt + n^{-1/2}\, dB_2(t)$, $t \in [0, 1]$, where f is an unknown periodic function and θ an unknown location parameter taking values in some interval $[-\tau, \tau]$, $0 < \tau < 1/2$. Let θ be given the uniform prior and for f consider the following two priors Π_1 and Π_2 under which the distributions of f can be represented respectively as

$$f(x) = \sqrt{2}\sum_{k=1}^{\infty}\big[(2k)^{-\alpha-1/2}\xi_{2k}\cos(2\pi kx) + (2k)^{-\alpha-1/2}\xi_{2k}\sin(2\pi kx)\big],$$

$$f(x) = \sqrt{2}\sum_{k=1}^{\infty}\big[(2k)^{-\alpha-1/2}\xi_{2k}\cos(2\pi kx) + (2k+1)^{-\alpha-1/2}\xi_{2k}\sin(2\pi kx)\big],$$

where $\xi_1, \xi_2, \ldots \overset{\text{iid}}{\sim} \text{Nor}(0, 1)$. Assume that the true value of θ is 0 and the true value f_0 of f is given by

$$f_0(x) = \sqrt{2}\sum_{k=1}^{\infty}\big[(2k)^{-\beta-1/2}\cos(2\pi kx) + (2k+1)^{-\beta-1/2}\sin(2\pi kx)\big],$$

where $\beta > 3/2$. Show that both priors lead to the same posterior convergence rate $n^{-\alpha\wedge\beta/(2\alpha+1)}$, the Bernstein–von Mises theorem holds for the prior Π_1, but the Bernstein–von Mises theorem fails for the prior Π_2.

12.8 Suppose that a and b are elements of a Hilbert space such that $h - \langle h, a\rangle b \perp b$, for every h. Show that $a = b/\|b\|^2$ and $b = a/\|a\|^2$.