7

Consistency: Examples

In this chapter we apply the result of the preceding chapter in a number of statistical settings: density estimation with a variety of priors (Pólya trees, Dirichlet process mixtures, Gaussian processes), nonparametric and semiparametric regression problems, and time series models. The Kullback-Leibler property of priors plays a central role, and is discussed extensively.

7.1 Priors with the Kullback-Leibler Property

In this section we prove that common priors such as the Pólya tree process and Dirichlet process mixtures and for density estimation possess the Kullback-Leibler property under appropriate conditions. Another important class of priors, namely those based on Gaussian processes, also posses the same property under appropriate conditions. However, we shall prove much stronger results in Chapter 11 under essentially the same conditions, so we refrain from presenting separate results on Kullback-Leibler property of Gaussian processes in this section.

7.1.1 Pólya Trees

The canonical Pólya tree $\mathrm{PT}^*(\lambda, a_m)$ is defined as a prior on probability measures in Definition 3.20 (and the subsequent discussion). Its parameters are a probability measure λ and a sequence of constants a_m. When $\sum_{m=0}^{\infty} a_m^{-1} < \infty$ the realizations from this prior are a.s. absolutely continuous relative to λ, by Theorem 3.16, and the prior can be viewed as a prior on λ-densities. The following theorem shows that it automatically possesses the Kullback-Leibler property.

Theorem 7.1 *Let P follow a canonical Pólya tree prior $\mathrm{PT}^*(\lambda, a_m)$ and such that $\sum_{m=1}^{\infty} a_m^{-1} < \infty$. Then P possesses a density p with respect to λ a.s. and the prior $\mathrm{PT}^*(\lambda, a_m)$ possesses the Kullback-Leibler property at any density p_0 such that $K(p_0; \lambda) < \infty$. Furthermore, if $V_2(p_0; \lambda) < \infty$, then $V_2(p_0; p) < \infty$ a.s. $[\mathrm{PT}^*(\lambda, a_m)]$.*

Proof As shown in Theorem 3.16, the density p of $P \sim \mathrm{PT}^*(\lambda, a_m)$ exists a.s. and can be represented as the infinite product (3.18). In the present situation, the variables V_ε possess $\mathrm{Be}(a_{|\varepsilon|}, a_{|\varepsilon|})$-distributions, all variables $V_{\varepsilon 0}$ are independent, and $V_{\varepsilon 1} = 1 - V_{\varepsilon 0}$, for $\varepsilon \in \mathcal{E}^*$. The discretization of p_0 to the partition at stage m has density $p_{0,m} = 2^m \sum_{\varepsilon \in \mathcal{E}^m} P(A_\varepsilon) \mathbb{1}\{A_\varepsilon\}$ relative to λ (which has $\lambda(A_\varepsilon) = 2^{-|\varepsilon|}$). For $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \mathcal{E}^m$ we can factorize $P(A_\varepsilon) = \prod_{j=1}^{m} v_{\varepsilon_1 \cdots \varepsilon_j}$, for $v_{\varepsilon\delta} = P(A_{\varepsilon\delta} | A_\varepsilon)$. As $p_{0,m} = \mathrm{E}_\lambda(p_0 | \mathcal{T}_m)$, where

$\mathcal{T}_m = \sigma\langle \mathcal{T}_m \rangle$, the martingale convergence theorem gives that $p_{0,m} \to p_0$ λ-a.e. as $m \to \infty$. This leads to the representations

$$p_{0,m}(x) = \prod_{j=1}^{m} (2v_{x_1 \cdots x_j}), \qquad p_0(x) = \prod_{j=1}^{\infty} (2v_{x_1 \cdots x_j}). \tag{7.1}$$

The quotient p_0/p can also be written as an infinite product, and the Kullback-Leibler divergence $K(p_0; p) = \int p_0 \log(p_0/p)\, d\lambda$ can be decomposed as, for any $m \geq 1$,

$$\int \Big[\sum_{j=1}^{m} \log \frac{v_{x_1 \cdots x_j}}{V_{x_1 \cdots x_j}} + \sum_{j=m+1}^{\infty} \log(2v_{x_1 \cdots x_j}) - \sum_{j=m+1}^{\infty} \log(2V_{x_1 \cdots x_j}) \Big] p_0(x)\, d\lambda(x).$$

We split this in the three terms indicated in square brackets. For given m, the event that the first term is smaller than some given $\delta > 0$ can be written as the event that a finite set of independent beta-variables $V_{x_1 \cdots x_j}$ falls into a nonempty open set of the appropriate unit cube (it is nonempty as $(v_{x_1 \cdots x_j} \colon j = 1, \ldots, m,\ (x_1, \ldots, x_m) \in \mathcal{E}^m)$ belongs to it). This has positive probability for any $\delta > 0$. The second term is small for large m by the last assertion of Lemma B.10. The third term is independent of the first. The expectation and variance of the integrand are both bounded by $\sum_{j=m+1}^{\infty} a_j^{-1}$, uniformly in x, by Lemma G.12. It follows that the third term tends to zero in probability, and hence is certainly smaller than a given $\delta > 0$ with positive probability.

This concludes the proof of the Kullback-Leibler property.

Since $V_2(p_0; p) \leq 2P_0(\log p_0)^2 + 2P_0(\log p)^2$, for the final assertion it is enough to show that $P_0(\log p)^2 < \infty$, a.s. In view of (3.18) the expectation of the latter random variable is bounded above by

$$\mathrm{E} \int \Big[\sum_{j=1}^{\infty} \log(2V_{x_1 \cdots x_j}) \Big]^2 p_0(x)\, d\lambda(x).$$

The second moment of the series is the sum of its variance and the square of its expectation, which were already noted to be bounded by a multiple of $\sum_{j=1}^{\infty} a_j^{-1} < \infty$, uniformly in x, by Lemma G.12. □

7.1.2 Kernel Mixtures

By equipping the parameters F and φ in mixtures of the type $p_{F,\varphi}(x) = \int \psi(x; \theta, \varphi)\, dF(\theta)$ with priors, we obtain a prior on densities. In this section we derive conditions on the kernel $\psi(\cdot; \theta, \varphi)$, a given family of probability density functions, so that the resulting prior possesses the Kullback-Leibler property as soon as the priors on F and φ have full (weak) support.

We assume that the sample space $\mathfrak{X}$ and the parameter spaces Θ and Φ of θ and φ are Polish spaces, equipped with their Borel σ-fields. The parameter φ may be vacuous, in which case the conditions simplify.

The true density p_0 need not itself possess mixture form, but must be approximable by mixtures. For given F_ϵ, φ_ϵ we may decompose

$$K(p_0; p_{F,\varphi}) = K(p_0; p_{F_\epsilon,\varphi_\epsilon}) + P_0 \log \frac{p_{F_\epsilon,\varphi_\epsilon}}{p_{F,\varphi}}. \tag{7.2}$$

We conclude that p_0 possesses the Kullback-Leibler property if, for every $\epsilon > 0$, the pair $(F_\epsilon, \varphi_\epsilon)$ can be chosen such that $K(p_0; p_{F_\epsilon,\varphi_\epsilon}) < \epsilon$ and such that the set of $(F_\epsilon, \varphi_\epsilon)$ so that the second term is bounded by ϵ and has positive prior probability. The second is certainly true if $(F, \varphi) \mapsto P_0 \log(p_{F_\epsilon,\varphi_\epsilon}/p_{F,\varphi})$ is continuous at $(F_\epsilon, \varphi_\epsilon)$, and $(F_\epsilon, \varphi_\epsilon)$ is in the support of its prior. In most examples these conditions are easiest to verify if F_ϵ is chosen compactly supported.

The following lemma gives explicit assumptions. For $D \subset \Theta$ and $N \subset \Phi$ define

$$\underline{\psi}(x; D, N) = \inf_{\theta \in D, \varphi \in N} \psi(x; \theta, \varphi), \qquad \overline{\psi}(x; \theta, N) = \sup_{\varphi \in N} \psi(x; \theta, \varphi).$$

Theorem 7.2 (Mixtures) *Suppose that for every $\epsilon > 0$ there exist $F_\epsilon \in \operatorname{supp}(\Pi)$ and $\varphi_\epsilon \in \operatorname{supp}(\mu)$, and open sets $D \supset \operatorname{supp}(F_\epsilon)$ and $N \ni \varphi_\epsilon$ with $K(p_0; p_{F_\epsilon,\varphi_\epsilon}) < \epsilon$ and that the following conditions hold.*

(A1) *$\{\theta \mapsto \psi(x; \theta, \varphi), \varphi \in N\}$ is uniformly bounded and equicontinuous on D, for any x.*
(A2) *$\int \log (p_{F_\epsilon,\varphi_\epsilon}(x)/\underline{\psi}(x; D, N))\, dP_0(x) < \infty$.*
(A3) *$\varphi \mapsto \psi(x; \theta, \varphi)$ is continuous at φ_ϵ, for any x and $\theta \in D$.*
(A4) *$\int \overline{\psi}(x; \theta, N)\, dF_\epsilon(\theta) < \infty$, for every x.*

Then $p_0 \in \mathrm{KL}(\Pi^*)$ for the prior Π^* on $p_{F,\varphi}$ induced by the prior $\Pi \times \mu$ on (F, φ).

Proof Finiteness of $K(p_0; p_{F_\epsilon,\varphi_\epsilon})$ implies that $p_{F_\epsilon,\varphi_\epsilon}(x) > 0$ for P_0-a.e. x. Therefore the decomposition (7.2) is valid. It suffices to show that the second term on the right is bounded by a multiple of ϵ with positive prior probability. Because F_ϵ and φ_ϵ are contained in the supports of their priors by assumption, this is true if the limit superior of this term as $F \rightsquigarrow F_\epsilon$ and $\varphi \to \varphi_\epsilon$ is nonpositive. If F_D is the restriction of F to D, then $p_{F,\varphi} \geq p_{F_D,\varphi}$, and hence the second term on the right increases if F is replaced by F_D. Because D contains the support of F_ϵ in its interior, so that $F_\epsilon(\partial D) = 0$, we have $F_D \rightsquigarrow F_\epsilon$ if $F \rightsquigarrow F_\epsilon$. Therefore, it suffices to show the limit of this term is nonpositive for *sub*probability measures $F \rightsquigarrow F_\epsilon$ supported within D.

By condition (A1), we have that $p_{F,\varphi}(x) \to p_{F_\epsilon,\varphi}(x)$ uniformly in $\varphi \in N$, for every x, as $F \rightsquigarrow F_\epsilon$ (and all are supported on D). By (A3) and (A4) and the dominated convergence theorem we also have that $p_{F_\epsilon,\varphi}(x) \to p_{F_\epsilon,\varphi_\epsilon}(x)$ as $\varphi \to \varphi_\epsilon$, for every x. Combined with the last assertion this shows that $p_{F,\varphi}(x) \to p_{F_\epsilon,\varphi_\epsilon}(x)$ as $F \rightsquigarrow F_\epsilon$ and $\varphi \to \varphi_\epsilon$. Also for F supported on D,

$$\log \frac{p_{F_\epsilon,\varphi_\epsilon}}{p_{F,\varphi}}(x) \leq \log \frac{p_{F_\epsilon,\varphi_\epsilon}(x)}{\underline{\psi}(x; D, N)}.$$

By (A2) the functions on the right are dominated by a P_0-integrable function. It follows that $\limsup P_0 \log(p_{F_\epsilon,\varphi_\epsilon}/p_{F,\varphi}) \leq P_0 \log 1 = 0$, by the dominated convergence theorem. □

Next, specialize the kernel function to be of location-scale type

$$\psi(x; \mu, h) = \frac{1}{h^d} \chi\Big(\frac{x - \mu}{h}\Big),$$

for a given probability density function χ defined on $\mathfrak{X} = \mathbb{R}^d$. We first consider location-scale mixtures obtained by putting a distribution on $\theta := (\mu, h)$, and next location mixtures created by mixing only over $\theta := \mu$ and considering $\varphi := h$ as an additional parameter.

Theorem 7.3 (Location-scale mixtures) *Assume that*

(B1) χ *is bounded, continuous and positive everywhere,*
(B2) $\int p_0(x) \log p_0(x)\, dx < \infty$,
(B3) $-\int p_0(x) \log \inf_{\|y\|<\delta} p_0(x-y)\, dx < \infty$, *for some* $\delta > 0$,
(B4) $\inf_{\|y\|<\|x\|^\eta} \chi(x-y) \geq \underline{\chi}(x)$ *for large* $\|x\|$ *and a function* $\underline{\chi}$ *that is decreasing as its argument moves away from zero and satisfies* $-\int p_0(x) \log \underline{\chi}(2x|x|^\eta)\, dx < \infty$, *for some* $\eta \in (0,1)$.

Then $p_0 \in \mathrm{KL}(\Pi)$ *for the prior* Π *on* $p_F = \int h^{-d} \chi((\cdot - \mu)/h)\, dF(\mu, h)$ *induced by a prior on* F *with full support* $\mathfrak{M}(\mathbb{R}^d \times (0,\infty))$; *and also for the prior on* $p_{F,h} = \int h^{-d} \chi((\cdot - \mu)/h)\, dF(\mu)$ *induced by a product prior on* (F, h) *with full support on* $\mathfrak{M}(\mathbb{R}^d) \times (0,\infty)$.

Proof Since there is no parameter φ, it suffices to verify conditions (A1)–(A2) of Theorem 7.2. We choose the measure F_ϵ on $\mathbb{R}^d \times (0,\infty)$ of the form $F_m := P_m \times \delta_{h_m}$, where P_m is P_0 restricted and renormalized to $\{\mu : \|\mu\| \leq m\}$, and δ_h is the Dirac measure at $h \in (0,\infty)$. We shall show that $K(p_0; p_{F_m}) \to 0$ as $m \to \infty$, for $h_m = m^{-\eta}$ and fixed $\eta > 0$, so that $K(p_0; p_{F_m}) < \epsilon$ for large enough m. The measure F_m is supported within the open set $D = \{\mu : \|\mu\| < m'\} \times (a, b)$, for $m' > m$ and $0 < a < h_m < b$. Boundedness and continuity of the map $(\mu, h) \mapsto \chi_h(x-\mu) := \chi((x-\mu)/h)/h^d$ on D as in condition (A1) is clear from the continuity of χ. Only Condition (A2) remains to be verified.

As is known from the theory of kernel approximation, $p_0 * \chi_h \to p_0$ in $\mathbb{L}_1$, as $h \to 0$. Since $\|p_m * \chi_h - p_0 * \chi_h\|_1 \leq \|p_m - p_0\|_1 \to 0$, as $m \to \infty$, we see by the triangle inequality that $p_{F_m} = p_m * \chi_{h_m} \to p_0$ in $\mathbb{L}_1$. To show that $K(p_0; p_{F_m}) \to 0$ it suffices to show that $\log(p_0/p_{F_m})$ is uniformly integrable in $\mathbb{L}_1(p_0)$. As $-\log x \geq 1 - x$ for $x > 0$, this function is bounded from below by $1 - p_{F_m}/p_0$, which is uniformly integrable relative to p_0, as $p_{F_m} \to p_0$ in $\mathbb{L}_1$. It suffices to show that $\log(p_0/p_{F_m})$ is bounded above by a p_0-integrable function.

Letting $c_m^{-1} = \int_{\|x\|<m} p_0(x)\, dx$, we can write

$$p_{F_m}(x) = c_m \int_{\|y\|<m} p_0(y) \chi_{h_m}(x-y)\, dy.$$

This immediately gives the estimate, for $\|x\| \leq m/2$ so that $\|y\| < m$ if $\|x - y\| < h_m$,

$$p_{F_m}(x) \geq c_m \inf_{\|y\|<h_m} p_0(x-y) \chi_{h_m}(y : \|y\| < h_m).$$

For $\|x\| > m/2$ and $\|y\| < C$, we have (easily, as $h_m = m^\eta$) for sufficiently large m that $\|y/h_m\| \leq \|x/h_m\|^\eta$ and hence $\chi_{h_m}(x-y) \geq \underline{\chi}(x/h_m)$, by the definition of $\underline{\chi}$. Since $\|x/h_m\| = \|x\| m^\eta$ is closer to the origin than $2x\|x\|^\eta$, if $\|x\| > m/2$, the assumed monotonicity of $\underline{\chi}$ yields $\chi_{h_m}(x-y) \geq \underline{\chi}(2x\|x\|^\eta)$, so that, for $\|x\| > m/2$,

$$p_{F_m}(x) \geq c_m P_0(y : \|y\| < C) \underline{\chi}(2x\|x\|^\eta).$$

Combining the preceding displays and noting that $c_m \to 1$, we find, for sufficiently large m,

$$p_{F_m}(x) \gtrsim \inf_{\|y\|<\epsilon} p_0(x-y) \wedge \underline{\chi}(2x\|x\|^\eta).$$

Hence $\log(p_0/p_{F_m})$ is bounded above by $\log p_0$ minus the negative logarithm of the function on the right side, which is integrable from above by assumptions (B2), (B3) and (B4).

For $D = \{\mu: \|\mu\| < m'\} \times (a, b)$ we have for $\|\mu/h\| \le \|x/h\|^\eta$

$$\inf_{(\mu,h)\in D} \frac{1}{h}\chi\Big(\frac{x-\mu}{h}\Big) \ge \frac{1}{a}\underline{\chi}\big(\frac{x}{h}\big).$$

In particular, this is valid for $\|x\|^\eta \ge K/a^{1-\eta}$, and then the right side is further bounded below by a multiple of $\underline{\chi}(x\|x\|^\eta)$. For $\|x\| < C$ and any given $C > 0$ the function on the left is bounded away from zero by the continuity and positivity of χ. It follows that minus the logarithm of the left side is integrable above. Because p_{F_m} is uniformly bounded, for fixed m, by the boundedness of χ, it follows that (A2) is satisfied.

The proof of the second part of the assertion, for location mixtures with a prior on scale, proceeds by similar arguments. □

Example 7.4 (Skew-normal kernel) The *skew-normal kernel* with skewness parameter λ is given by

$$\chi(x) = \frac{1}{\pi} e^{-x^2/2} \int_{-\infty}^{\lambda x} e^{-t^2/2} dt.$$

The value $\lambda = 0$ corresponds to the normal kernel. For this kernel the Kullback-Leibler property holds at every continuous density p_0 that has a finite moment of order $2 + \eta > 2$ and satisfies (B2)–(B3).

Because the skew-normal density is monotone at its two extremes, for (B4) it suffices to verify finiteness of $\int p_0(x) \log \chi(2x|x|^\eta)\, dx$, which is clear from the moment condition.

Example 7.5 (Multivariate normal kernel) For the standard multivariate normal kernel $\chi(x) = (2\pi)^{-d/2} e^{-\|x\|^2/2}$, the Kullback-Leibler property holds for every p_0 that satisfies conditions (B2)–(B3) and has $\int \|x\|^{2+\eta} p_0(x)\, dx < \infty$, for some $\eta > 0$.

As the kernel decreases monotonically with increasing $\|x\|$, condition (B4) can be checked by a monotonicity argument, as in the one-dimensional case.

Example 7.6 (Histograms) For given $\theta \in (0, 1]$ and $m \in \mathbb{N}$ let $\psi(\cdot; \theta, m)$, the density of the uniform distribution on the interval $((i-1)/m, i/m]$ if $(i-1)/m < \theta \le i/m$ (i.e. $\psi(x; , \theta, m) = m$ if both x and θ belong to $((i-1)/m, i/m]$ and is 0 otherwise). Then the mixture $\int \psi(\cdot; \theta, m)\, dF(\theta)$, for a measure F on $(0, 1]$ and fixed m, is a histogram with bins $((i-1)/m, i/m]$. Alternatively, the mixture $\int \psi(\cdot; \theta, m)\, dF(\theta, m)$, for a measure F on $(0, 1] \times \mathbb{N}$ is a histogram with variable bins, which may be used to estimate a density p_0 on the unit interval. We may construct a prior on histograms by putting a prior on F and/or m. In both cases any continuous density p_0 is contained in Kullback-Leibler support of the prior, as soon as the prior on F has full weak support, and in the case of fixed bins, the number m of bins is given a prior with infinite support in $\mathbb{N}$.

If p_0 is bounded away from zero, then this is easy to derive from Theorem 7.2, and the fact that any continuous function can be uniformly approximated by a histogram. If

p_0 is bounded away from zero this automatically also gives approximation relative to the Kullback-Leibler divergence. Extension to any continuous density next follows with the help of Corollary 6.35.

Example 7.7 (Bernstein polynomial kernel) The Bernstein polynomial prior on densities on [0, 1] described in Example 5.10, with $(w_0, \ldots, w_k)|\, k \sim \Pi_k$ and $k \sim \mu$, for a prior Π_k with full support $\mathbb{S}_k$ and μ with infinite support in $\mathbb{N}$ contains every continuous density p_0 in its Kullback-Leibler support.

As every continuous p_0 is uniformly approximated by its associated Bernstein polynomial, this can be proved similarly as in Example 7.6.

Example 7.8 (Lognormal kernel) The two parameter *lognormal kernel*

$$\psi(x; \mu, \varphi) = (2\pi)^{-1/2} x^{-1} \varphi^{-1} e^{-(\log x - \mu)^2/(2\varphi^2)}$$

is supported on $(0, \infty)$ rather than on the full line. We can investigate whether a density is in the Kullback-Leibler support of lognormal mixtures by applying Theorem 7.3 after mapping the positive half line to $\mathbb{R}$ by the logarithmic function. Since this map is a bimeasurable bijection, the Kullback-Leibler divergence between the induced distributions is the same as between the original distributions, whence the Kullback-Leibler property is preserved.

Under the map $y = \log x$ the density p_0 transforms into $y \mapsto q_0(y) = e^y p_0(e^y)$, and the lognormal mixtures into normal mixtures. The conditions (B2)–(B3) on q_0 translate back into the conditions on the original density p_0 that the two integrals $\int_0^\infty p_0(x) \log(x p_0(x))\, dx$ and $-\int_0^\infty p_0(x) \log \inf_{\|u-1\|<\delta} (xu\, p_0(xu))\, dx$ are finite, while the condition that q_0 has a finite $2+\eta$ moment translates into convergence of the integral $\int_0^\infty p_0(x) |\log x|^{2+\eta}\, dx$. Any p_0 with these properties is in the Kullback-Leibler support of the prior.

Example 7.9 (Exponential kernel) A mixture p_F of the exponential scale family on $\mathfrak{X} = (0, \infty)$ satisfies

$$p_F(x) = \int \lambda e^{-\lambda x}\, dF(\lambda) = -\frac{d}{dx} \int e^{-\lambda x}\, dF(\lambda).$$

Consequently, the survival function $\int_x^\infty p_F(y)\, dy$ is the Laplace transform of a measure F. The set of Laplace transforms of measures on $(0, \infty)$ is exactly the set of *completely monotone functions*: functions G such that $(-1)^n d^n G(x)/\, dx^n \geq 0$, for all $n, x > 0$ (Feller 1971, Chapter XIII).

If F is given a prior with full support $\mathfrak{M}(0, \infty)$, then any density p_0 with a completely monotone survival function, finite first absolute moment, and $\int p_0(x) \log p_0(x)\, dx < \infty$ belongs to the Kullback-Leibler support of the induced prior on p_F.

This may be proved using Theorem 7.2, by taking F_ϵ the distribution P_0 truncated to a sufficiently large compact set in $(0, \infty)$.

7.1.3 Exponential Densities

In Section 2.3.1 it is seen that the Kullback-Leibler divergence between densities of the form

$$p_f(x) = e^{f(x) - c(f)}, \qquad c(f) = \log \int e^f\, d\nu, \tag{7.3}$$

is bounded above by a function of the uniform distance between the corresponding functions f. This readily gives the Kullback-Leibler property for the prior induced on these densities by a prior on f.

Theorem 7.10 *If f_0 is bounded and belongs to the uniform support of the prior Π, then $p_{f_0} \in \mathrm{KL}(\Pi^*)$ for Π^* the prior induced on p_f by the prior Π on f.*

Proof If $\|f - f_0\|_\infty < \epsilon$, then $\|f - f_0\|_\infty \le \|f_0\|_\infty + \epsilon$, and hence $K(p_{f_0}; p_f) < \epsilon^2 e^\epsilon (1+\epsilon)$, by Lemma 2.5. Then we have $\Pi^*(p_f : K(p_{f_0}; \bar{p}_f) < 6\epsilon^2) \ge \Pi(f : \|f - f_0\|_\infty < \epsilon)$, for $\epsilon < 1$, which is positive by assumption. □

Example 7.11 If $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$ and p_0 is continuous and strictly positive, then $f_0 = \log p_0$ is well defined and continuous, and hence belongs to the support of any fully supported Borel prior on $\mathfrak{C}(\mathcal{X})$. Since $p_0 = p_{f_0}$ it follows that p_0 is in the Kullback-Leibler support of the induced exponential prior. By Corollary 6.35 this extends to *any* continuous density.

Concrete examples are fully supported Gaussian process priors, and series priors as in Lemma 2.2, which lead to infinite-dimensional exponential families.

7.2 Density Estimation

In Section 7.1 the Kullback-Leibler property was seen to hold under mild conditions for common priors, including Pólya tree processes, Dirichlet process mixtures, and random series. The Kullback-Leibler property implies consistency with respect to the weak topology (see Example 6.20), but this topology is too weak to be really useful for density estimation. Theorem 6.23 shows that the Kullback-Leibler property together with a complexity bound gives consistency for the total variation distance (or equivalently the Hellinger distance). The following counterexample confirms that the Kullback-Leibler property alone is not sufficient for consistency relative to this stronger topology.

In the remainder of this section we next investigate consistency in the total variation distance for normal mixtures, general Dirichlet process mixtures, Pólya trees and series priors. Other examples are treated in later chapters in the context of rates of contraction; in particular, priors based on Gaussian processes in Chapter 11.

Example 7.12 (Inconsistency) We shall construct a prior on the set of probability densities on $[0, 1]$ that has the Kullback-Leibler property relative to the uniform density p_0, but yields a posterior distribution such that $\limsup_{n\to\infty} \Pi_n(\|p - p_0\|_1 = 1 | X_1, \ldots, X_n) = 1$, almost surely, if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_0$.

The prior puts mass $1/2$ on a model $\mathcal{P}_0$ and masses $3/(\pi N)^2$ on models $\mathcal{P}_N$, for $N = 1, 2, \ldots$ Here $\mathcal{P}_0 = \{p_\theta : 0 \le \theta \le 1\}$ is the parametric model consisting of the probability densities $p_\theta(x) = e^{-\theta + \sqrt{2\theta}\Phi^{-1}(x)}$, for Φ the standard normal cumulative distribution function, and the prior distributes the mass within the model according to the prior $\pi(\theta) \propto e^{-1/\theta}$. For every $N \ge 1$ the model $\mathcal{P}_N$ is the collection of all histograms on the partition $((j-1)/(2N^2), j/(2N^2)]$, for $j = 1, \ldots, 2N^2$, with N^2 ordinate values 2 and

N^2 ordinate values 0; the prior spreads the mass uniformly over the $\binom{2N^2}{N^2}$ elements of the model. The values 0 and 2 are chosen so that $\|p - p_0\|_1 = 1$, for all $p \in \mathcal{P}_N$ and all $N \geq 1$. Thus the claim on the posterior distribution follows if $\limsup \Pi_n(\mathcal{P}_0 | X_1, \ldots, X_n) = 0$, almost surely. The prior possesses the Kullback-Leibler property in virtue of the fact that p_θ is equal to the uniform density if $\theta = 0$ and $\theta \mapsto p_\theta$ is smooth; in fact $K(p_0; p_\theta) = \theta$, for all $\theta \in [0, 1]$.

Intuitively the inconsistency is caused by the overly large number of densities at distance one from p_0. Because $\int \psi(x) p_N(x)\, dx \to \int \psi(x) p_0(x)\, dx$ as $N \to \infty$, for any bounded continuous function ψ and any $p_N \in \mathcal{P}_N$, the distance of $\mathcal{P}_N$ to p_0 for the weak topology actually tends to zero. This explains that the posterior can be consistent relative to the weak topology, as it should be since p_0 possesses the Kullback-Leibler property.

It remains to be shown that $\limsup \Pi_n(\mathcal{P}_0 | X_1, \ldots, X_n) = 0$, almost surely. For $N^2 \geq n$ and any given observed values $X_1, \ldots, X_n$, there are at least $\binom{2N^2 - n}{N^2}$ densities $p \in \mathcal{P}_N$ such that $\prod_{i=1}^n p(X_i) = 2^n$. Indeed, this value of the likelihood pertains if all observations fall in a bin with ordinate value 2; since the n observations fall in maximally n different bins, this leaves at least $2N^2 - n$ bins from which to choose the N^2 bins with value 0; each choice gives a density in $\mathcal{P}_N$. It follows that

$$\begin{aligned}\int_{\cup_{N>0}\mathcal{P}_N} \prod_{i=1}^n p(X_i)\, d\Pi(p) &\geq 2^n \sum_{N \geq \sqrt{n}} \frac{1}{2} \frac{6}{\pi^2 N^2} \binom{2N^2 - n}{N^2} \Big/ \binom{2N^2}{N^2} \\ &\geq 2^n \frac{3}{\pi^2} \sum_{N \geq \sqrt{n}} \frac{1}{N^2} \frac{1}{2^n} \gtrsim \frac{1}{\sqrt{n}}.\end{aligned}$$

On the other hand, for c the norming constant of the prior,

$$\int_{\mathcal{P}_0} \prod_{i=1}^n p(X_i)\, d\Pi(p) = c \int_0^1 e^{-n\theta + \sqrt{2\theta} \sum_{i=1}^n \Phi^{-1}(X_i)} e^{-1/\theta}\, d\theta.$$

As $-n\theta - 1/\theta \leq -2\sqrt{n}$ for all n and $0 \leq \theta \leq 1$, and $\sum_{i=1}^n \Phi^{-1}(X_i) < 0$ infinitely often a.s. $[P_0^\infty]$ by the law of the iterated logarithm, the right-hand side of the display is at most $e^{-2\sqrt{n}}$ infinitely often a.s. $[P_0^\infty]$. The quotient of the two integrals on the left gives the relative mass that the posterior assigns to $\cup_{N>0}\mathcal{P}_N$ and $\mathcal{P}_0$, respectively. The claim follows, as the lim sup of the quotient is ∞ almost surely.

7.2.1 Normal Mixtures

Let $p_{F,\sigma} = \phi_\sigma * F$, for ϕ_σ the density of the normal distribution with mean 0 and variance σ^2, and consider the prior on densities induced by equipping σ and F with independent priors $F \sim \Pi$ and $\sigma \sim \mu$. Theorem 7.3 gives conditions for the Kullback-Leibler property.

Theorem 7.13 *If for every $\delta, \epsilon > 0$ there exists sequences a_n and $\sigma_n < S_n$ and a constant $C > 0$ such that $n^{-1} \log\log(S_n/\sigma_n) \to 0$ and*

$$\frac{a_n}{\sigma_n} < n\delta, \qquad \Pi(F: F[-a_n, a_n]^c > \epsilon) < e^{-Cn}, \qquad \mu([\sigma_n, S_n]^c) \leq e^{-Cn},$$

then the posterior distribution $\Pi_n(\cdot|X_1,\ldots,X_n)$ *for* $p_{F,\sigma}$ *in the model* $X_1,\ldots,X_n|(F,\sigma) \overset{iid}{\sim} p_{F,\sigma}$, *for* $(F,\sigma) \sim \Pi \times \mu$, *is strongly consistent relative to the total variation norm at every* p_0 *in the Kullback-Leibler support of the prior of* $p_{F,\sigma}$. *In particular, for* $\Pi = \mathrm{DP}(\alpha)$ *and* μ *a distribution with finite moment consistency holds if one of the following conditions holds:*

(i) α *has compact support and* $\mu(\sigma < h) \le e^{-c/h}$, *for* $h \downarrow 0$.
(ii) α *has sub-Gaussian tails and* $\mu(\sigma < h) \le e^{-c/h^2}$, *for* $h \downarrow 0$.

Proof Given $\epsilon' > 0$, let a_n, σ_n, S_n and C be as in the assumption for $\epsilon = \epsilon'/6$ and $\delta = (\epsilon')^3/(100\log(60/\epsilon'))$. We apply Theorem 6.23 with $\mathcal{P}_{n,1}$ the set of mixtures $\phi_\sigma * F$ when $\sigma \in [\sigma_n, S_n]$ and $F[-a_n, a_n] \ge 1-\epsilon$, and define $\mathcal{P}_{n,2}$ the complementary set of mixtures. Clearly

$$\Pi(\mathcal{P}_{n,2}) \le \mu([\sigma_n, S_n]^c) + \Pi(F\colon F[-a_n,a_n]^c > \epsilon) \le 2e^{-Cn}.$$

We shall show that the $\mathbb{L}_1$-entropy of $\mathcal{P}_{n,1}$ at ϵ' is bounded above by $n(\epsilon')^2/4$.

Fix an integer $N \sim a_n/(\sigma_n\epsilon)$, and a regular partition of $[-a_n, a_n]$ into N adjacent intervals $E_1,\ldots,E_N$ of equal length, with midpoints $\theta_1,\ldots,\theta_N$. Construct a set S^* of points in $(\sigma_n, S_n]$ by choosing an $\epsilon 2^{k-1}$-net in $(\sigma_n 2^{k-1}, \sigma_n 2^k]$, for $k = 1, 2, \ldots, \lceil \log_2(S_n/\sigma_n)\rceil$. Let $\mathcal{P}^*$ be the set of all densities of the form

$$\phi_{\sigma^*} * \sum_{i=1}^N p_i^* \delta_{\theta_i},$$

when σ^* ranges over S^*, and $p^* = (p_1^*,\ldots,p_N^*)$ ranges over an ϵ-net P^* over the N-dimensional unit simplex equipped with the ℓ_1-norm. The number of points in $\mathcal{P}^*$ is equal to $\#S^*\#P^*$, and hence $\log \#\mathcal{P}^* \le \log\log_2(S_n/\sigma_n) + \log(3/\epsilon) + N\log(5/\epsilon)$, in view of Proposition C.1. The parameter δ has been defined so that $N\log(5/\epsilon)$ is bounded above by $n(\epsilon')^2/8$, under the condition $a_n/\sigma_n < n\delta$. Furthermore, $\log(2/\epsilon\sigma_n) + \log(2S_n/\epsilon) \ll n$ by assumption, as $n \to \infty$.

For any $\sigma \in [\sigma_n, S_n]$ and arbitrary F, there exists $\sigma^* \in S^*$ and $p^* \in P^*$ such that $|\sigma - \sigma^*| < \epsilon\sigma_n 2^{k-1}$ if $\sigma \in [\sigma_n 2^{k-1}, \sigma_n 2^k]$, and $\sum_{i=1}^N |p_i^* - F(E_i)/F[-a_n,a_n]| < \epsilon$. If $F[-a_n,a_n] \ge 1-\epsilon$, then also $\sum_{i=1}^N |p_i^* - F(E_i)| < 2\epsilon$. Because $\|\phi_{\sigma_1} - \phi_{\sigma_2}\|_1 \le 2|\sigma_1-\sigma_2|/(\sigma_1\wedge\sigma_2)$ and $\|\phi_\sigma * \delta_{\theta_1} - \phi_\sigma * \delta_{\theta_2}\|_1 \le |\theta_1-\theta_2|/\sigma$, for any $\sigma_1,\sigma_2,\sigma,\theta_1,\theta_2$, we have

$$\|\phi_\sigma * F - \phi_{\sigma^*} * F\|_1 \le \frac{2|\sigma-\sigma^*|}{\sigma\wedge\sigma^*},$$

$$\Big\|\phi_{\sigma^*} * F - \phi_{\sigma^*} * \sum_{i=1}^N F(E_i)\delta_{\theta_i}\Big\|_1 \le F[-a_n,a_n]^c + \sum_{i=1}^N \int_{E_i} \|\phi_{\sigma^*} * \delta_\theta - \phi_{\sigma^*} * \delta_{\theta_i}\|_1\, dF(\theta),$$

$$\Big\|\phi_{\sigma^*} * \sum_{i=1}^N F(E_i)\delta_{\theta_i} - \phi_{\sigma^*} * \sum_{i=1}^N p_i^*\delta_{\theta_i}\Big\|_1 \le \sum_{i=1}^N |F(E_i) - p_i^*|.$$

By the triangle inequality it follows that every element of $\mathcal{P}_{n,1}$ is approximated within $\mathbb{L}_1$-distance $6\epsilon < \epsilon'$ by some element of $\mathcal{P}^*$.

By Markov's inequality and by Proposition 4.2,

$$\mathrm{DP}_\alpha(F\colon F[-a_n, a_n]^c > \epsilon) \le \epsilon^{-1}\bar{\alpha}([-a_n, a_n]^c).$$

If μ has a finite first moment, then $\mu([S_n, \infty)) \le e^{-Cn}$ for $S_n \gtrsim e^{Cn}$, and $\log\log S_n = o(n)$. If α has compact support, then a_n can be chosen equal to a fixed, large constant, and $a_n/\sigma_n < n\delta$ if $\sigma_n \gtrsim 1/n$. The bound $\mu([0, \sigma_n)) \le e^{-Cn}$ is then implied by $\mu([0, h)) \le e^{-c/h}$. If α has sub-Gaussian tails, then a_n can be chosen of the order $\sqrt{n}$, and σ_n of the order $1/\sqrt{n}$, and a similar argument yields the stated bound. □

Example 7.14 (Inverse-gamma) The conjugate *inverse-gamma* prior $\sigma^{-2} \sim \mathrm{Ga}(a, b)$ satisfies $\mathrm{E}\sigma < \infty$ for $a > 1$, and $\mathrm{P}(\sigma < h) \sim e^{-b/h^2} h^{-2a+2}$. Thus both conditions (i) and (ii) are satisfied if $a > 1$.

For a multivariate generalization of the result, with matrix-valued bandwidth, see Problem 7.10. For Dirichlet mixtures, see also Section 7.2.2 and Section 9.4.

7.2.2 Dirichlet Process Mixtures of a General Kernel

Let $p_{F,\varphi}(x) = \int \psi(x; \theta, \varphi)\, dF(\theta)$ for a given family of probability densities $x \mapsto \psi(x; \theta, \varphi)$, indexed by Euclidean parameters $\theta \in \Theta \subset \mathbb{R}^k$ and $\varphi \in \Phi \subset \mathbb{R}^l$. Equip F with the Dirichlet process prior and φ by some other prior. The following theorem shows that consistency relative to the $\mathbb{L}_1$-norm holds under mild conditions as soon as the true density is in the Kullback-Leibler support of the prior. Section 7.1.2 gives examples of the latter.

Theorem 7.15 *If for any given $\epsilon > 0$ and n, there exist subsets $\Theta_n \subset \mathbb{R}^k$ and $\Phi_n \subset \mathbb{R}^l$ and constants $a_n, A_n, b_n, B_n > 0$ such that*

(i) $\|\psi(\cdot; \theta, \varphi) - \psi(\cdot; \theta', \varphi')\|_1 \le a_n\|\theta - \theta'\| + b_n\|\varphi - \varphi'\|$, *for all* $\theta, \theta' \in \Theta_n$ *and* $\varphi, \varphi' \in \Phi_n$,
(ii) $\mathrm{diam}(\Theta_n) \le A_n$ *and* $\mathrm{diam}(\Phi_n) \le B_n$,
(iii) $\log(a_n A_n) \le C \log n$ *for some* $C > 0$, *and* $\log(b_n B_n) \le n\epsilon^2/(8l)$,
(iv) $\max(\bar{\alpha}(\Theta_n^c), \pi(\Phi_n^c)) \le e^{-Cn}$, *for some* $C > 0$,

then the posterior distribution $\Pi_n(\cdot \mid X_1, \ldots, X_n)$ for $p_{F,\varphi}$ in the model $X_1, \ldots, X_n \mid (F, \varphi) \overset{iid}{\sim} p_{F,\varphi}$, for $(F, \varphi) \sim \mathrm{DP}(\alpha) \times \pi$, is strongly consistent relative to the total variation norm at every p_0 in the Kullback-Leibler support of the prior of $p_{F,\varphi}$.

Proof We apply Theorem 6.23 with d equal to the $\mathbb{L}_1$-distance divided by 2. For given $\epsilon > 0$ we choose $\mathcal{P}_{n,1}$ equal to, with $N \sim n\delta/\log n$ and δ sufficiently small, to be determined,

$$\mathcal{P}_{n,1} = \Big\{\sum_{j=1}^{\infty} w_j \psi(\cdot; \theta_j, \varphi)\colon (w_j) \in \mathbb{S}_\infty, \sum_{j>N} w_j < \frac{\epsilon}{8},\ \theta_1, \ldots, \theta_N \in \Theta_n, \varphi \in \Phi_n\Big\}.$$

By the series representation $F = \sum_j W_j \delta_{\theta_j}$ of the Dirichlet process, given in Theorem 4.12, the prior density $p_{F,\varphi}$ is contained in $\mathcal{P}_{n,1}$, unless the weights of the representation satisfy

$\sum_{j>N} W_j \geq \epsilon/8$, or at least one of $\theta_1, \ldots, \theta_N \overset{\text{iid}}{\sim} \bar{\alpha}$ falls outside Θ_n, or $\varphi \notin \Phi_n$. It follows that

$$\Pi(\mathcal{P}_{n,1}^c) \leq \mathrm{P}\Big(\sum_{j>N} W_j \geq \frac{\epsilon}{8}\Big) + N\bar{\alpha}(\Theta_n^c) + \pi(\Phi_n^c).$$

The last two terms are exponentially small by assumption and the choice of N. The stick-breaking weights in the first term satisfy $W_j = V_j \prod_{l=1}^{j-1}(1 - V_l)$, for $V_l \overset{\text{iid}}{\sim} \mathrm{Be}(1, |\alpha|)$ and $\sum_{j>N} W_j = \prod_{j=1}^{N}(1 - V_j)$. Since $-\log(1 - V_l)$ possesses an exponential distribution, $R_N := -\log \sum_{j>N} W_j$ is gamma distributed with parameters N and $|\alpha|$, and hence $\mathrm{P}(R_N < r) \leq (|\alpha|r)^N/N! \leq (e|\alpha|r/N)^N$. Therefore, the first term is bounded above by $(e|\alpha| \log(8/\epsilon)/N)^N$, which is also exponentially small, by choice of N, for any $\delta > 0$

The functions of the form $\sum_{j=1}^{N} w_j \psi(\cdot; \theta_j, \varphi)$ with $(w_1, \ldots, w_N) \in \mathbb{S}_N$ form an $\epsilon/4$-net over $\mathcal{P}_{n,1}$ for the $\mathbb{L}_1$-norm. To construct an $3\epsilon/4$-net over these finite sums we restrict $(w_1, \ldots, w_N)$ to an $\epsilon/4$-net over $\mathbb{S}_N$, restrict $(\theta_1, \ldots, \theta_N)$ to an $\epsilon/(4a_n)$-net over Θ_n and φ to an $\epsilon/(4b_n)$-net over Φ_n. The cardinality of such a net is bounded above by

$$\Big(\frac{20}{\epsilon}\Big)^N \times \Big(\frac{12A_n a_n}{\epsilon}\Big)^{kN} \times \Big(\frac{12B_n b_n}{\epsilon}\Big)^{l}.$$

By the triangle inequality, it follows that $\log N(\epsilon, \mathcal{P}_{n,1}, \|\cdot\|_1) \leq n\epsilon^2$, provided δ is small enough. □

Conditions of the theorem typically hold for a location kernel with an additional scale parameter if the prior density of the scale parameter has a thin tail at zero and the base measure of the Dirichlet process has a sufficiently thin tail. In particular, consistency for Dirichlet process mixtures of a normal kernel can be derived from this theorem. The result can however go much beyond location kernels.

7.2.3 Pólya Tree Process

For parameters satisfying $\sum_{m=1}^{\infty} a_m^{-1} < \infty$, the canonical Pólya tree process $\mathrm{PT}^*(\lambda, a_m)$ possesses the Kullback-Leibler property at densities with finite Kullback-Leibler divergence relative to λ (see Section 7.1.1). Under a much more restrictive condition the posterior is consistent with respect to the total variation distance.

Theorem 7.16 *If $\sum_{m>k} a_m^{-1} \leq C2^{-k}$ for some constant C and every large k, then the posterior distribution $\Pi_n(\cdot \mid X_1, \ldots, X_n)$ in the model $X_1, \ldots, X_n \mid p \overset{iid}{\sim} p$ and $p \sim \mathrm{PT}^*(\lambda, a_m)$ is strongly consistent with respect to the total variation distance at every p_0 with $K(\lambda; p_0) < \infty$.*

Proof Since $p_0 \in \mathrm{KL}(\mathrm{PT}^*(\lambda, a_m))$ by Theorem 7.1, it suffices to verify conditions (i) and (ii) of Theorem 6.23. The random density p generated by the Pólya tree process can be represented as in (3.18), where $x_1 x_2 \cdots$ is the expansion of x relative to the partition, the $V_{\varepsilon 0}$ are independent $\mathrm{Be}(a_{|\varepsilon|}, a_{|\varepsilon|})$ variables, and $V_{\varepsilon 1} = 1 - V_{\varepsilon 0}$, for $\varepsilon \in \mathcal{E}^*$. Write p_k for the corresponding finite products $p_k(x) = \prod_{j=1}^{k}(2V_{x_1 \cdots x_j})$. For given $k = k_n \in \mathbb{N}$, consider the partition of the model in the sets $\mathcal{P}_{n,1} = \{p: \|\log(p/p_k)\|_\infty \leq \epsilon^2/8\}$ and their complements $\mathcal{P}_{n,2}$.

Because p_k is constant within every kth level partitioning set, we have

$$\big|\log p(x) - \log p(y)\big| \le 2\big\|\log(p/p_k)\big\|_\infty,$$

whenever x and y belong to the same partitioning set. Thus the modulus $\Delta \log p$ of every $p \in \mathcal{P}_{n,1}$ relative to the kth level partition is bounded above by $\epsilon^2/4$. Because $\epsilon^2/4 + \sqrt{2\epsilon^2/4} < \epsilon$, we conclude by Corollary C.3 that $\log N(\epsilon, \mathcal{P}_{n,1}, \|\cdot\|_1) \le 2^k C_\epsilon$ for some positive number C_ϵ not depending on k. Since the squared Hellinger distance is bounded by the $\mathbb{L}_1$-distance, the entropy relative to the Hellinger distance satisfies the same inequality, with a different constant C_ϵ. Condition (i) of Theorem 6.23 is satisfied provided k is chosen to satisfy $2^k \le D_\epsilon n$, for a specific constant D_ϵ that depends on ϵ only.

By Lemma G.12 the variables $\log(2V_{x_1,\ldots,x_j})$ have mean bounded in absolute value by a multiple of a_j^{-1} and Orlicz ψ_2-norm bounded by $a_j^{-1/2}$. The second implies that $\sum_{j>k}\big[\log(2V_{x_1\cdots x_j}) - \mathrm{E}(\log(2V_{x_1\cdots x_j}))\big]$ possesses Orlicz ψ_2-norm bounded by a multiple of the square root of $\sum_{j>k} a_j^{-1}$, which is bounded by $C2^{-k}$ by assumption. (See Lemma 2.2.1 and Proposition A.1.6 in van der Vaart and Wellner 1996) for the relevant results on Orlicz norms.) It follows that for $2^{-k} \lesssim \epsilon^2/16$ there exists a constant D such that

$$\Pi(\mathcal{P}_{n,2}) = \mathrm{P}\Big(\Big|\sum_{j>k}\log(2V_{x_1\cdots x_j})\Big| > \frac{\epsilon^2}{8}\Big) \le e^{-D\epsilon^4 2^k}.$$

Therefore condition (ii) of Theorem 6.23 is satisfied for $2^k \gtrsim n$. In particular this is true for the maximal value $2^k \sim D_\epsilon n$ found in the preceding paragraph. □

The exponential growth condition for the parameters a_m of the preceding theorem seems a bit too strong for practical use. In Section 6.8.5 the posterior mean was seen to be consistent under only the condition $\sum_m a_m^{-1} < \infty$ which is needed even for absolute continuity.

7.2.4 Exponential Densities

In Theorem 7.10 a prior on a density $p = p_f$ of the form (7.3) induced by a prior Π on f is seen to possess the Kullback-Leibler property as soon as $\log p_0$ belongs to the uniform support of Π. Because the Hellinger distance $d_H(p_f, p_g)$ is bounded by a continuous function of $\|f - g\|_\infty$ (see Lemma 2.5), the entropy condition of Theorem 6.23 similarly is implied by an entropy condition on the functions f under the uniform norm.

Theorem 7.17 *If Π is the distribution of a stochastic process $(f(x): x \in [0,1]^d)$ with sample paths belonging to $\mathfrak{C}^\alpha([0,1]^d)$ and with $\mathrm{P}(\|f\|_{\mathfrak{C}^\alpha} > M) \le e^{-CM^r}$, for some constants C, r and $\alpha \ge d/r$, then the posterior distribution $\Pi_n(\cdot\,|\,X_1,\ldots,X_n)$ for p in the model $X_1,\ldots,X_n|\,p \overset{iid}{\sim} p$, $p = p_f$, $f \sim \Pi$ is strongly consistent at every p_0 such that $\log p_0$ belongs to the uniform support of Π.*

Proof The uniform entropy $\log N(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty)$ of the set of functions $\mathcal{F}_n = \{f: \|f\|_{\mathfrak{C}^\alpha} \le cn^{1/r}\}$ is bounded above by a multiple of $(cn^{1/r}/\epsilon)^{d/\alpha}$. Hence for given ϵ and c' it is bounded above by $c'n\epsilon^2$, if c is chosen sufficiently small. In view of Lemma 2.5 the ϵ-entropy of the induced set of densities $\{\bar{p}_f: f \in \mathcal{F}_n\}$ relative to the Hellinger distance is bounded above the 2ϵ-entropy of $\mathcal{F}_n$ relative to the uniform norm. By assumption $\Pi(\mathcal{F}_n^c) \le e^{-Cc^r n}$. The theorem follows from Theorem 6.23. □

Example 7.18 (Gaussian process) According to *Borell's inequality* a Gaussian random element f in a separable Banach space satisfies $\mathrm{P}(\|f\| > M) \leq 2e^{-M^2/2}$. Thus if the prior process f is a Gaussian, Borel measurable map in a separable subspace of $\mathfrak{C}^\alpha([0,1]^d)$, then the preceding theorem applies with $r = 2$. For a d-dimensional domain its condition is then that the sample paths of f are $d/2$-smooth ($\alpha \geq d/2$).[1]

Example 7.19 (Exponential family) Let Π be the distribution of $f := \sum_j \theta_j \psi_j$ for given basis functions $\psi_j : [0,1]^d \to \mathbb{R}$ and independent random variables θ_j. If the basis functions are contained in $\mathfrak{C}^\alpha([0,1]^d)$ with norms b_j, then $\|\sum_j \theta_j \psi_j\|_{\mathfrak{C}^\alpha} \leq \sum_j |\theta_j| b_j$. Thus Π concentrates on $\mathfrak{C}^\alpha([0,1]^d)$ as soon as $\sum_j \mathrm{E}|\theta_j| b_j < \infty$. The tail condition is certainly verified if $\mathrm{P}(\sum_j |\theta_j| b_j > M) \leq e^{-CM^r}$, but due to cancellation of positive and negative terms in the series $\sum_j \theta_j \psi_j$, this condition may be pessimistic.

For instance, if $\theta_j \overset{\text{ind}}{\sim} \mathrm{Nor}(0, \tau_j^2)$, then the process is Gaussian, and satisfies a tail bound with $r = 2$ as soon as $\sum_j \tau_j b_j < \infty$.

For $d = 1$ and the trigonometric basis $\psi_0(x) = 1$, and $\psi_j(x) = \sqrt{2}\cos(j\pi x)$ for $j \in \mathbb{N}$, we can choose $b_0 = 1$ and $b_j \sim j^\alpha$, for $\alpha \in (0, 1]$. For Gaussian variables we choose $\alpha = 1/2$ and obtain the condition that $\sum_j \tau_j \sqrt{j} < \infty$. E.g. $\tau_j \sim j^{-3/2-\delta}$ works for any $\delta > 0$. For further discussion of Gaussian priors, see Chapter 11, or Section 6.8.4.

7.3 Other Nonparametric Models

In this section we consider a few other nonparametric function estimation problems and investigate conditions for consistency in appropriate distance measures.

7.3.1 Nonparametric Binary Regression

In nonparametric binary regression the relationship between a binary response $Y \in \{0, 1\}$ and a covariate X is modeled as

$$\mathrm{P}_f(Y = 1 \mid X = x) = H(f(x)),$$

for a known "link function" H and an unknown function f that ranges over a large model $\mathcal{F}$. We assume that H is strictly monotone and continuously differentiable, with bounded derivative h. The logistic function is the most important example of a link function and is also the most convenient to deal with.

We consider two versions of the problem, both with as complete set of observations a sequence of pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$. In *random design* these are a random sample from a fixed distribution P_f, determined by the marginal distribution G of the covariates and the conditional Bernoulli distribution $Y_i \mid X_i \sim \mathrm{Bin}(1, H(f(X_i)))$ of the response. In *fixed design* the covariates are given and only the responses are random; they are assumed independent, with Bernoulli distributions as before.

In both cases we construct a prior on the distribution of the observations by endowing f with a prior Π.

The random design model is a special case of the density model of Section 6.4. Theorem 6.23 translates into the following result.

[1] Sufficient is that its sample paths are in $\mathfrak{C}^\beta([0,1]^d)$, for some $\beta > \alpha$; see Lemma I.7.

Theorem 7.20 (Random design) *If for every $\epsilon > 0$ there exists a partition $\mathcal{F} = \mathcal{F}_{n,1} \cup \mathcal{F}_{n,2}$ and a constant $C > 0$, such that $\log N(\epsilon, \mathcal{F}_{n,1}, \mathbb{L}_2(G)) \leq \|h\|_\infty^2 n\epsilon^2$, and $\Pi(\mathcal{F}_{n,2}) \leq e^{-Cn}$, then the posterior distribution is strongly consistent relative to the distance $d(p_f, p_g) = \|H(f) - H(g)\|_{G,2}$ at every p_{f_0} in the Kullback-Leibler support of the prior. For the logistic link function the latter is true for every f_0 in the $\mathbb{L}_2(G)$-support of Π. For a general link function it suffices that f_0 is uniformly bounded and is in the uniform support of Π.*

Proof We apply Theorem 6.23 with d equal to $(1/2)$ times the $\mathbb{L}_2$-distance on the densities p_f, which is bounded by the Hellinger distance, as the densities p_f are bounded by 1 (see Lemma B.1). By Lemma 2.8(i) this distance is equivalent to the distance $\|H(f) - H(g)\|_{2,G}$ of the theorem, and bounded above by $\|h\|_\infty$ times the $\mathbb{L}_2(G)$-distance on the regression functions. Thus (i)–(ii) of Theorem 6.23 are verified by the present conditions.

By Lemma 2.8 the Kullback-Leibler divergence between the densities p_f is bounded above the $\mathbb{L}_2(G)$-distance between the functions f, always in case of the logistic link function, and if the functions f are bounded for general link functions. This yields the Kullback-Leibler property. □

The fixed design model can be handled by Theorem 6.41. Let $G_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ be the empirical distribution of the design points.

Theorem 7.21 (Fixed design) *If for every $\epsilon > 0$ there exist partitions $\mathcal{F} = \mathcal{F}_{n,1} \cup \mathcal{F}_{n,2}$ and a constant $C > 0$, such that $\log N(\epsilon, \mathcal{F}_{n,1}, \mathbb{L}_2(G_n)) \leq 3\|h\|_\infty^2 n\epsilon^2$ and $\Pi(\mathcal{F}_{n,2}) \leq e^{-Cn}$, then $\Pi_n(\|H(f) - H(g)\|_{2,G_n} > \epsilon | Y_1, \ldots, Y_n) \to 0$ in probability for every $\epsilon > 0$ in the random design model with regression function f_0, for any f_0 such that $\liminf \Pi_n(\|f - f_0\|_{2,G_n} < \epsilon) > 0$ in case of the logistic link function, and such that $\liminf \Pi_n(\|f - f_0\|_\infty < \epsilon) > 0$ in general.*

Proof By Lemma 2.8(ii), the root average square Hellinger distance $d_{n,H}$ on the (multivariate Bernoulli) density of $(Y_1, \ldots, Y_n)$ is bounded by the $\mathbb{L}_2(G_n)$-distance on the functions f. Thus (i)–(ii) of Theorem 6.41 are verified by the present conditions. Similarly by Lemma 2.8(iii)–(iv) the Kullback-Leibler divergence and Kullback-Leibler variation are bounded above by a multiple of the square $\mathbb{L}_2(G_n)$-distance on the functions f, which allows to verify the first condition of Theorem 6.41. The theorem gives consistency for $d_{n,H}$, but then also for the root average square $\mathbb{L}_2$-distance, as this is bounded by a multiple of the root average square Hellinger distance. This distance is equivalent to the distance $\|H(f) - H(g)\|_{2,G_n}$. □

The marginal distribution G or G_n in the consistency assertion can be removed in favor of a neutral, not problem-dependent measure, such as the Lebesgue measure, if the former is sufficiently regular (see e.g. Problem 7.11). This may not be true of link function in the assertion: estimating f in its tails may be harder than estimating it in the middle as the distributions of the observations depends on f through $H(f)$ and thus is insensitive to perturbations of f in its tails.

Gaussian processes provide concrete examples of priors. They are discussed in detail in Chapter 11.

7.3.2 Nonparametric Regression with Normal Errors

In the nonparametric regression problem the observations are pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ that follow the model, for an unknown function f and mean zero errors $\varepsilon_1, \ldots, \varepsilon_n$,

$$Y_i = f(X_i) + \varepsilon_i.$$

We consider that the errors are a random sample from a normal distribution $\mathrm{Nor}(0, \sigma^2)$, and only equip the regression function f with a prior. In particular we treat the standard deviation σ of the errors as known; modification to unknown σ is possible for light-tailed priors.

We restrict to the *random design* model, where the observations (X_i, Y_i) are i.i.d.; the fixed design case is similar. The marginal distribution G of $X_1, \ldots, X_n$ cancels from the posterior distribution for f, even if it were equipped with an (independent) prior, and hence can be assumed known without loss of generality.

Let $[f]_M := (f \vee -M) \wedge M$ be f truncated to the interval $[-M, M]$.

Theorem 7.22 *If for every $\epsilon > 0$ there exist partitions $\mathcal{F} = \mathcal{F}_{n,1} \cup \mathcal{F}_{n,2}$ and a constant $C > 0$ such that $\log N(2\sigma\epsilon, \mathcal{F}_{n,1}, \|\cdot\|_{2,G}) \leq n\epsilon^2$ and $\Pi(\mathcal{F}_{n,2}) \leq e^{-cn}$, then $\Pi_n(f\colon \|[f]_M - [f_0]_M\|_{2,G} > \epsilon | (X_1, Y_1), \ldots, (X_n, Y_n)) \to 0$ a.s. $[P_{f_0}^\infty]$, for every M, for any f_0 that belongs to the $\mathbb{L}_2(G)$-support of Π.*

Proof The inequality $K(\phi_{\theta_1}; \phi_{\theta_2}) \lesssim |\theta_1 - \theta_2|^2$ for the Kullback-Leibler divergence between two $\mathrm{Nor}(\theta_i, \sigma^2)$-densities ϕ_{θ_1} and ϕ_{θ_2}, implies $K(p_{f_0}; p_f) \leq \|f_0 - f\|_{2,G}^2$ for the density p_f of (X, Y). Therefore, the Kullback-Leibler property of p_{f_0} follows from the assumption that f_0 is in the $\mathbb{L}_2(G)$-support of Π.

Similarly the inequality $d_H(\phi_{\theta_1}; \phi_{\theta_2}) \leq \frac{1}{2}|\theta_1 - \theta_2|/\sigma$ implies that $d_H(p_f, p_g) \leq \frac{1}{2}\|f - g\|_{2,G}/\sigma$, whence $N(\epsilon, \{p_f\colon f \in \mathcal{F}_{n,1}\}, d_H) \leq N(2\sigma\epsilon, \mathcal{F}_{n1}, \|\cdot\|_{2,G})$. Consequently, the entropy condition of Theorem 6.23 is implied by the present condition.

We conclude that the posterior distribution is consistent relative to the Hellinger distance. Finally $\|[f]_M - [f_0]_M\|_{G,2} \lesssim d_H(f, f_0)$. □

7.3.3 Spectral Density Estimation

The second order moments of a stationary time series $(X_t\colon t \in \mathbb{Z})$ are completely described by its *spectral density* defined by, for $\gamma(h) = \mathrm{cov}(X_{t+h}, X_t)$ the autocovariance function,

$$f(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) e^{-i\,h\omega}, \qquad 0 \leq \omega \leq \pi.$$

This may be estimated from data $X_1, \ldots, X_n$ by (weighted) averages of the *periodogram* I_n, defined as $I_n(\omega) = (2\pi n)^{-1} \left| \sum_{t=1}^n X_t e^{-it\omega} \right|^2$. Under general conditions the periodogram $I_n(\omega)$ at a frequency $\omega \in (0, \pi)$ tends in distribution to an exponential distribution with mean $f(\omega)$, and periodogram values at sufficiently separated frequencies are asymptotically independent. Furthermore, the periodogram values $U_l = I_n(\omega_l)$ at the *natural frequencies* (or *Fourier frequencies*) $\omega_l = 2\pi l/n$, for $l = 1, \ldots, \nu = \lfloor (n-1)/2 \rfloor$, are asymptotically distributed as independent exponential variables with means $f(\omega_l)$. This motivates the

Whittle likelihood, which is an approximate likelihood for f based on the assumption that the variables $U_1, \ldots, U_\nu$ are exactly exponentially distributed:

$$L_n(f \mid X_1, \ldots, X_n) = \prod_{l=1}^{\nu} \frac{1}{f(\omega_l)} e^{-U_l / f(\omega_l)}, \tag{7.4}$$

If the time series is Gaussian with $\sum_{r=0}^{\infty} r^\alpha |\gamma(r)| < \infty$ for some $\alpha > 0$, and the spectral density is bounded away from 0, then by Theorem L.8 the sequence of true distributions of $(U_1, \ldots, U_\nu)$ and the sequence of distributions under the assumption that the U_i are independent exponential variables with means $f(\omega_l)$ are mutually *contiguous*. This means that any convergence in probability, such as consistency, automatically holds under both true and "Whittle" distributions, if it holds under one of them.

Here we consider Bayesian estimation using a prior on f and forming the posterior distribution $\Pi_n(\cdot \mid U_1, \ldots, U_\nu)$ with the Whittle likelihood in place of the true likelihood. As a metric on the spectral densities consider

$$d_n(f, g) = \sqrt{\frac{1}{\nu} \sum_{l=1}^{\nu} \frac{\left(\sqrt{f}(\omega_l) - \sqrt{g}(\omega_l)\right)^2}{f(\omega_l) + g(\omega_l)}}.$$

Theorem 7.23 *If for every $\epsilon > 0$ there exist partitions $\mathcal{F} = \mathcal{F}_{n,1} \cup \mathcal{F}_{n,2}$ and a constant $C > 0$ such that $\log N(\epsilon, \mathcal{F}_{n,1}, d_n) \leq n\epsilon^2$ and $\Pi(\mathcal{F}_{n,2}) \leq e^{-cn}$, then $\Pi_n(f : d_n(f, f_0) > \epsilon \mid U_1, \ldots, U_\nu) \to 0$ a.s. under every f_0 that is bounded away from 0 and ∞ and belongs to the $\|\cdot\|_\infty$-support of the prior Π. This is true both if $U_1, \ldots, U_\nu$ are independent exponential variables with means $f(\omega_l)$, and if these variables are the periodogram at the natural frequencies of a stationary Gaussian times series with $\sum_{h=0}^{\infty} h^\alpha |\gamma(h)| < \infty$, for some $\alpha > 0$.*

Proof The final assertion follows from Theorem L.8; it suffices to prove the theorem for $(U_1, \ldots, U_\nu)$ exponential variables. For p_θ the density of the exponential distribution with mean θ,

$$\begin{aligned} d_H(p_{\theta_1}, p_{\theta_2}) &= \frac{\sqrt{2}|\sqrt{\theta_1} - \sqrt{\theta_2}|}{\sqrt{\theta_1 + \theta_2}}, \\ K(p_{\theta_0}; p_\theta) &= -\log \frac{\theta_0}{\theta} - 1 + \frac{\theta_0}{\theta} \leq \frac{|\theta - \theta_0|^2}{(\theta \wedge \theta_0)^2}, \\ V_{2,0}(p_{\theta_0}; p_\theta) &= \frac{|\theta - \theta_0|^2}{\theta^2}. \end{aligned}$$

From the second and third, it follows that if $\|f - f_0\|_\infty < \epsilon$ and f_0 is bounded away from zero, then the corresponding distributions $P_{f,n}$ of $(U_1, \ldots, U_\nu)$ possess Kullback-Leibler divergence and variation $K(P_{f_0,n}; P_{f,n})$ and $V_{2,0}(P_{f_0,n}; P_{f,n})$ bounded by a multiple of ϵ^2. The first inequality shows that the root average square Hellinger distance on the distributions $P_{f,n}$ is equal to the distance d_n on the spectral densities f. Thus the theorem follows from Theorem 6.41. □

The integral of the spectral density is the variance $\operatorname{var} X_0 = \int f(\omega)\, d\omega$ of the time series. Norming the spectral density to a probability density entails dividing it by this variance, and gives the Fourier transform of the auto*correlation* function. Consider estimating this function in the case that the variance is known. (As the variance may be estimated by $n^{-1}\sum_{t=1}^{n} X_t^2$, this also gives a data-dependent prior for estimating the full spectral density.)

Theorem 7.24 *If Π is a prior on $q = f/\|f\|_1$ such that $\Pi(\|q\|_{\mathrm{Lip}} > L_n) \lesssim e^{-Cn}$, for some $C > 0$ and $L_n = o(n)$, then the posterior distribution as in Theorem 7.23 is consistent relative to the $\mathbb{L}_1$-norm on q, at any Lipschitz continuous q_0 that is bounded away from zero and is in the support of Π for the uniform norm.*

Proof That the conditions of Theorem 6.41 concerning $K(P_{f_0,n}; P_{f,n})$ and $V_{2,0}(P_{f_0,n}; P_{f,n})$ hold follow as in the preceding proof. Rather than invoking the general construction of tests through the metrics d_n we use a direct argument.

We may assume that $\|q\|_{\mathrm{Lip}} \le L_n$. If Q_n is the probability measure putting equal mass at each of the Fourier frequencies, then $\int q\, dQ_n = 1 + O(L_n/n) = 1 + o(1)$, and $\|q - q_0\|_1 > 2\epsilon$ implies that $\|q - q_0\|_{1,Q_n} > \epsilon$ for sufficiently large n. Thus it follows from part (ii) of Lemma K.9 that $Q_n(q < q_0 - \epsilon) > \epsilon'$ for some $\epsilon' > 0$: the fraction of Fourier frequencies where $f(\omega_j) < f_0(\omega_j) - \epsilon$ is of the order of n. Since this gives a separation of the means under q and q_0 of at least cn of the observations U_j, a test with exponentially small error probabilities can be constructed by Lemma D.10. For all f inside a ball $\{f: \|f - f_1\|_\infty < \epsilon\}$ around some f_1, the same set of Fourier frequencies pertains and hence the same test can be used. We can cover $\mathcal{F}_{n,1} = \{f: \|q\|_{\mathrm{Lip}} \le L_n\}$ with a multiple of $L_n/\epsilon \ll n$ balls of radius ϵ. Thus the overall error probabilities are of the order $e^{o(n)}e^{-Cn}$. Furthermore, by assumption $\mathcal{F}_{n,2} = \{f: \|q\|_{\mathrm{Lip}} > L_n\}$ has exponentially small prior mass. □

Example 7.25 The conditions of Theorem 7.24 hold for a Bernstein polynomial prior on q (with range rescaled to $[0, \pi]$) provided the prior ρ on the index k satisfies $0 < \rho(k) \lesssim e^{-Ck^2 \log k}$. In this case we may choose $\mathcal{F}_{n,1} = \{q: k \le \delta\sqrt{n/\log n}\}$, which leads to $L_n = \delta' n/\log n$. The condition is also satisfied for the prior $q \propto e^W$, where W is a Gaussian process with sufficiently regular paths.

7.4 Semiparametric Models

In semiparametric problems, such as location or linear regression problems with unknown error distribution, the Euclidean part of the parameter is usually of most interest. In a proof of consistency of the posterior of this parameter the topology on the functional part of the parameter may then be chosen for convenience. For instance, in the location or regression problems one may work with the weak topology on the error density.

7.4.1 Location Problem

Suppose we observe a random sample from the model $X = \theta + \varepsilon$, where $\varepsilon \sim p$ and p is a probability density with a fixed known quantile. Specifically, we shall consider the more

involved case that p is symmetric about zero, but similar results are true if the median or mean of the error is fixed.

The key to consistency is a prior on the error density that has the Kullback-Leibler property, so as to avoid misbehavior as exhibited in Example 6.14. Let $\bar{\Pi}$ be the prior induced by $p \sim \Pi$ on symmetric densities $\bar{p}$ under the map $p \mapsto \bar{p}$, given by $\bar{p}(x) = \frac{1}{2}(p(x) + p(-x))$.

Theorem 7.26 *If $(\theta, p) \sim \mu \times \bar{\Pi}$ for $\theta_0 \in \operatorname{supp}(\mu)$, then the posterior distribution for (θ, p) is consistent at (θ_0, p_0) with respect to the product of the Euclidean and the weak topology if for every $C > 1$ there exist a density r with $p_0 \leq Cr$, $r_\theta \in \mathrm{KL}(\Pi)$ for every sufficiently small $|\theta|$ and $K(r; r_\theta) \to 0$ as $\theta \to 0$. In particular, if the Kullback-Leibler support of Π contains all densities p with $K(p; \lambda) < \infty$ for some density λ, then this is true if one of the following conditions holds:*

(a) *$K(p_{0,\theta}; \lambda) < \infty$ for sufficiently small $|\theta|$, and $K(p_0; p_{0,\theta}) \to 0$ as $\theta \to 0$.*
(b) *p_0 is continuous with compact support and $-\int \phi_{\mu,\sigma^2} \log \lambda < \infty$, for all (μ, σ^2).*

Proof Let Π^* be the prior on the shifted symmetric densities p_θ induced by the prior $(p, \theta) \sim \bar{\Pi} \times \mu$ under the map $(p, \theta) \mapsto p_\theta$. By Proposition A.9 the latter map is continuously invertible relative to the weak topology on p and p_θ and the Euclidean topology on θ. Hence it suffices that the posterior distribution for p_θ is weakly consistent at p_{0,θ_0}. By Example 6.20 this is the case as soon as $p_{0,\theta_0} \in \mathrm{KL}(\Pi^*)$.

If $p_0 \leq Cr$, then $p_{0,\theta_0} \leq Cr_{\theta_0}$, whence it suffices to show that $r_{\theta_0} \in \mathrm{KL}(\Pi^*)$, in view of Proposition 6.34. Now $r_\theta \in \mathrm{KL}(\Pi)$ implies $\overline{r_\theta} \in \mathrm{KL}(\bar{\Pi})$, by Proposition 6.32. This being true for every $\theta \approx 0$ and also $K(r; r_\theta) \to 0$ as $\theta \to 0$ by assumption, the desired result follows from Proposition 6.36.

(a) It suffices to apply the above argument with $r = p_0$.

(b) First assume that the support of p_0 is an interval $[-a, a]$. For $\eta \in (0, 1)$ define $r = \phi_{-a,\sigma^2} \mathbb{1}\{(-\infty, -a)\} + (1 - 2\eta)p_0 + \eta\phi_{-a,\sigma^2} \mathbb{1}\{(a, \infty)\}$, where $\sigma = \sigma(\eta)$ is adjusted so that r is continuous, at the points $\pm a$ and hence everywhere. For $C(1 - 2\eta) > 1$ we have $p_0 \leq Cr$, and $K(r_\theta; \lambda) = R \log r - R_\theta \log \lambda < \infty$. Furthermore $\log(r/r_\theta)(x) \to 0$ as $\theta \to 0$ for every x, while the function is dominated by a multiple of $1 + |x|$.

Next assume that the support of p_0 is contained in $[-a, a]$, but not necessarily the whole interval. Define r to be proportional to $(p_0 \vee \eta)\mathbb{1}\{[-a, a]\}$. Then $p_0 \leq Cr$ for sufficiently small η, and r has support $[-a, a]$. Next, proceed as for the special case. □

7.4.2 Linear Regression with Unknown Error Density

Consider the univariate regression problem with observations $Y_1, \ldots, Y_n$ following the model $Y_i = \alpha + \beta x_i + \varepsilon_i$, for fixed covariates $x_1, \ldots, x_n$ not all equal and errors $\varepsilon_i \overset{\text{iid}}{\sim} f$ with symmetric, but unknown density. The main interest is in estimating the regression parameters $(\alpha, \beta) \in \mathbb{R}^2$. Let $P_{\alpha,\beta,f,i}$ be the distribution of Y_i.

Given a prior Π on (α, β, f) assume that, for every $\epsilon > 0$,

$$\liminf_{n\to\infty} \Pi_n\Big((\alpha, \beta, f)\colon \frac{1}{n}\sum_{i=1}^n K(P_{\alpha_0,\beta_0,f_0,i};\, P_{\alpha,\beta,f,i}) < \epsilon, \qquad (7.5)$$
$$\frac{1}{n^2}\sum_{i=1}^n V_{2,0}(P_{\alpha_0,\beta_0,f_0,i};\, P_{\alpha,\beta,f,i}) < \epsilon\Big) > 0.$$

Theorem 7.27 *If* $\liminf n^{-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2 > 0$*, then the posterior distribution for* (α, β) *is consistent at every* (α_0, β_0, f_0) *at which* (7.5) *holds. If* (α, β) *and* f *are a priori independent and the prior of* (α, β) *possesses full support, then sufficient conditions for the latter to be true are that the covariates are uniformly bounded and either (i) or (ii) holds:*

(i) *For every* $C > 1$ *there exists* $\eta > 0$ *and a symmetric density* g *with* $\sup_{|\theta|<\eta} f_0(\cdot - \theta) \le Cg(\cdot)$ *and* $\Pi(f\colon K(g; f) < \epsilon, V_2(g; f) < \infty) > 0$ *for every* $\epsilon > 0$.
(ii) *The maps* $\theta \mapsto K(f_0; f(\cdot - \theta))$ *and* $\theta \mapsto V^+(f_0; f(\cdot - \theta))$ *are continuous at* $\theta = 0$ *and* $\Pi(f\colon K(f_0; f) < \epsilon, V_2(f_0; f) < \infty) > 0$ *for every* $\epsilon > 0$.

If the prior on f *has the form* $f = \int \phi_h(\cdot - z)\, dQ(z)$ *with* $Q \sim \tilde{\Pi}$*, then (ii) is verified if* $\int\int t^2 dQ(t)\, d\tilde{\Pi}(Q) < \infty$ *and* $\int (y^4 + \log_+^2 f_0(y)) f_0(y)\, dy < \infty$.

Proof As (7.5) repeats the condition on existence of sets B_n of Theorem 6.41, for the first assertion it suffices to show the existence of exponentially powerful tests of (α_0, β_0, f_0) versus $\tilde{\Theta}_n := \{(\alpha, \beta, f)\colon |\alpha - \alpha_0| > \epsilon, \text{ or } |\beta - \beta_0| > \epsilon\}$, so that the posterior distribution of these sets tends to zero by (a) of that theorem.

For $\psi = -g'/g = (2G - 1)$ the score function of the logistic distribution, consider the test statistics $T = n^{-1}\sum_{i=1}^n \psi(Y_i - \alpha_0 - \beta_0 x_i)(1, x_i)^\top$. By the symmetry of the error density $\mathrm{E}_0 T = 0$, while

$$\mathrm{E}_{\alpha,\beta,f} T = \frac{1}{n}\sum_{i=1}^n \int \psi(y + \alpha - \alpha_0 + (\beta - \beta_0)x_i) f(y)\, dy \begin{pmatrix} 1 \\ x_i \end{pmatrix} =: H_f(\alpha - \alpha_0, \beta - \beta_0).$$

Because $\psi' = 2g$, the Hessian matrix of the function H_f can be computed as

$$H_f'(\alpha, \beta) = \frac{1}{n}\sum_{i=1}^n \int 2g(y + \alpha + \beta x_i) f(y)\, dy \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}.$$

We show that the expression above is positive definite. To see this, fix y and abbreviate $2g(y + \alpha + \beta x_i) f(y)$ by c_i. Then for any $(a, b) \neq (0, 0)$, we observe that $a^2\sum_{i=1}^n c_i + 2ab\sum_{i=1}^n c_i x_i + b^2\sum_{i=1}^n c_i x_i^2 > 0$ because by the Cauchy-Schwarz inequality,

$$2ab\sum_{i=1}^n c_i x_i \le 2ab\sqrt{\sum_{i=1}^n c_i}\sqrt{\sum_{i=1}^n c_i x_i^2} \le a^2\sum_{i=1}^n c_i^2 + b^2\sum_{i=1}^n c_i x_i^2$$

and equality is ruled out as all x_i values are not equal. Clearly integration over y preserves the positivity, and hence H_f' is positive definite. Hence for any w in the unit sphere, the

function $t \mapsto w^\top H_f(tw)$, which has derivative $w^\top H_f'(tw)w$, is nondecreasing on $[0, \infty)$. This shows that the infimum of $H_f(\alpha, \beta)$ over (α, β) outside a sphere around the origin is assumed on the sphere. In view of the form of $H_f'(0)$ and continuity, it follows that on a sufficiently small sphere the infimum is bounded away from 0, uniformly in f such that $\int gf\, d\lambda$ is bounded away from 0. By Hoeffding's inequality the test that rejects when $\|T\|$ is bigger than a small constant has exponential power.

We are left with testing (α_0, β_0, f_0) versus alternatives (α, β, f) with $\int gf\, d\lambda < \eta$, for arbitrary constant $\eta > 0$. The statistic $T = n^{-1} \sum \mathbb{1}\{|Y_i - \alpha_0 - \beta_0 x_i| > M\}$ possesses means $\mathrm{E}_{\alpha_0, \beta_0, f_0} T = \mathrm{P}_{f_0}(|\varepsilon| > M)$ and

$$\mathrm{E}_{\alpha, \beta, f} T = \frac{1}{n} \sum_{i=1}^{n} \mathrm{P}_f(|\varepsilon_i + \alpha - \alpha_0 + (\beta - \beta_0)x_i| > M) \geq \mathrm{P}_f(\varepsilon > M),$$

as follows by splitting the sum over the terms with $\alpha - \alpha_0 + (\beta - \beta_0)x_i \geq 0$ (when the event contains $\varepsilon_i < -M$) and $\alpha - \alpha_0 + (\beta - \beta_0)x_i < 0$ (when the event contains $\varepsilon_i > M$), combined with the symmetry of f. Because $g(M)\mathbb{1}\{[-M, M]\} \leq g$, we have that $\mathrm{P}_f(\varepsilon > M) \geq \frac{1}{2}(1 - \int gf\, d\lambda / g(M))$, for any $M > 0$. Thus the expectations under null and alternative hypotheses are separated for M such that $\frac{1}{2}(1 - \eta/g(M)) > \mathrm{P}_{f_0}(|\varepsilon| > M)$. For sufficiently small η such M exists (e.g. $\mathrm{P}_{f_0}(|\varepsilon| > M) = 1/8$ and $\eta = \frac{1}{2} g(M)$).

To prove (7.5) under (i) we note that for bounded x_i all shifts $\alpha - \alpha_0 + (\beta - \beta_0)x_i$ are arbitrarily small if (α, β) ranges over a sufficiently small neighborhood of (α_0, β_0). The densities $f(\cdot - \alpha - \beta x_i)$ are then bounded above by $Cg(\cdot - \alpha_0 - \beta_0 x_i)$, whence $K(P_{\alpha_0, \beta_0, f_0, i}; P_{\alpha, \beta, f, i}) = K(f(\cdot + \alpha - \alpha_0 + (\beta - \beta_0)x_i); f) \leq \log C + C(\epsilon + \sqrt{\epsilon/2})$ if $K(g; f) < \epsilon$, by Lemma B.14. The Kullback-Leibler variation $V_{2,0} \leq V = V^+ + V^-$ is similarly bounded by Lemma B.14 (for V^+) and Lemma B.13 (for $V^- \leq 4K$).

For a proof of (7.5) under (ii) let $f_\theta = f(\cdot - \theta)$ and $A_\eta := \cap_{|\theta| < \eta} \{f : K(f_0; f_\theta) < \epsilon, V_2(f_0; f_\theta) < \infty\}$, for given $\epsilon > 0$. Then $\cup_{\eta > 0} A_\eta = \{f : K(f_0; f) < \epsilon, V_2(f_0; f) < \infty\}$ by the assumed continuity, and hence $\Pi(f \in A_\eta) > 0$ for some $\eta > 0$, by the assumption on Π. For all shifts $\alpha - \alpha_0 + (\beta - \beta_0)x_i$ bounded by η, we have $K(P_{\alpha_0, \beta_0, f_0, i}; P_{\alpha, \beta, f, i}) < \epsilon$ and $V_2(P_{\alpha_0, \beta_0, f_0, i}; P_{\alpha, \beta, f, i}) < \infty$, and then (ii) holds.

By Jensen's inequality a density of the form $f(y) = \int \phi_t(y - \theta)\, dQ(t)$ satisfies the inequality $\log(f_0/f) \leq \log f_0 - \int \log \phi_t\, dQ(t)$. For ϕ_t a shifted scaled normal kernel the absolute value of the second term is bounded above by a multiple of $y^2 + \int t^2\, dQ(t)$, uniformly in small shifts. The required continuity therefore follows by continuity of these functions relative to shifts and the dominated convergence theorem. □

Part (i) of the theorem applies, for instance, to symmetrized canonical Pólya tree priors with $\sum_{m=1}^{\infty} a_m^{-1/2} < \infty$, and part (ii) to Dirichlet mixtures with mixing measure $Q \sim \mathrm{DP}(\alpha)$ satisfying $\int t^2\, d\alpha(t) < \infty$.

Generalizations of Theorem 7.27 to multidimensional covariates are considered in Problems 7.18 and 7.19.

7.4.3 Binary Nonparametric Monotone Regression

Suppose we observe $Y_1, \ldots, Y_n$ for independent Bernoulli variables $Y_i \overset{\text{ind}}{\sim} \mathrm{Bin}(1, H(x_i))$, with success probabilities $H(x_i)$ depending on real deterministic covariates $x_1, \ldots, x_n$,

where $H\colon \mathbb{R} \to [0,1]$ is a monotone increasing unknown link function. We are interested in finding the covariate value x for which the probability of success $H(x)$ is equal to a given value ξ, i.e. the ξth quantile $x = H^{-1}(\xi)$ of H.

As a prior consider modeling the function H as $H(x) = F(\alpha + \beta x)$, and equipping the parameters F, α and β with priors. The latter parameters are not identifiable, but this is of no concern when estimating $H^{-1}(\xi)$, which is identifiable.

Let $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{x_i}$ be the empirical measure of the covariates.

Theorem 7.28 *Assume that H_0 is strictly increasing on a neighborhood of $H_0^{-1}(\xi)$. Let the covariates be contained within the support of H_0 and assume that $\liminf \mathbb{P}_n[a,b] > 0$ for every $a < b$ in a neighborhood of $H_0^{-1}(\xi)$. If $F \sim \Pi$ and $(\alpha,\beta) \sim \mu$ are a priori independent with priors with (full) supports $\mathfrak{M}$ and $\mathbb{R}\times[0,\infty)$, respectively, then $\Pi_n(|H^{-1}(\xi) - H_0^{-1}(\xi)| > \epsilon \,|\, Y_1,\ldots,Y_n) \to 0$ a.s. under H_0, for every $\epsilon > 0$.*

Proof If $H_m \rightsquigarrow H_0$, then by Pólya's theorem $H_m \to H_0$ uniformly. Therefore for every $\epsilon > 0$ there exists a weak neighborhood $\mathcal{U}$ of H_0 such that, for all $H \in \mathcal{U}$,

$$\frac{1}{n}\sum_{i=1}^n K(\mathrm{Bin}(1, H_0(x_i)), \mathrm{Bin}(1, H(x_i))) \le \frac{1}{n}\sum_{i=1}^n |H_0(x_i) - H(x_i)|^2 < \epsilon^2,$$

$$\sum_{i=1}^\infty \frac{1}{i^2} V_2(\mathrm{Bin}(1, H_0(x_i)), \mathrm{Bin}(1, H(x_i))) \le \sum_{i=1}^\infty \frac{1}{i^2}|H_0(x_i) - H(x_i)|^2 < \infty.$$

(See Problem B.2.) By the assumptions on the supports of the priors and Kolmogorov's strong law of large numbers the condition of Theorem 6.40 is verified.

It now suffices to show that there exist exponentially powerful tests of the null hypothesis H_0 versus the alternative $\{H\colon \left|H_0^{-1}(\xi) - H^{-1}(\xi)\right| > \epsilon\}$. We can split the alternative into two one-sided alternatives; consider the right alternative $\{H\colon H^{-1}(\xi) > H_0^{-1}(\xi) + \epsilon\}$. Because H_0 is strictly increasing, it is strictly greater than ξ on the interval $\mathfrak{X}_0 := \left[H_0^{-1}(\xi) + \epsilon/2,\, H_0^{-1}(\xi) + \epsilon\right]$; let $\eta = H_0(H_0^{-1}(\xi) + \epsilon/2) - \xi$ be the minimum gain. Because a function H attains the level ξ not before the right end point of $\mathfrak{X}_0$, we have $\mathrm{E}_H(Y_i) \le \xi$ for every i with $x_i \in \mathfrak{X}$. Since $\mathrm{E}_{H_0}(Y_i) = H_0(x_i) \ge \xi + \eta$, the test that rejects the null hypothesis if $n^{-1}\sum Y_i \mathbb{1}\{x_i \in \mathfrak{X}_0\} \mathbb{P}_n(\mathfrak{X}_0) < \xi + \eta/2$ is exponentially powerful, by Hoeffding's inequality and the assumption that $\liminf \mathbb{P}_n(\mathfrak{X}_0) > 0$. □

An extension to generalized linear models is considered in Problem 7.21.

7.5 Historical Notes

The Kullback-Leibler property of Dirichlet process mixtures was proved by Ghosal et al. (1999b) and subsequently improved by Tokdar (2006) for normal mixtures. Consistency for the Bernstein polynomial prior was shown by Petrone and Wasserman (2002); and of Bayesian histograms by Gasparini (1996), by direct methods. The Kullback-Leibler property of general kernel mixtures was studied by Wu and Ghosal (2008a). Simplified versions of their results are presented in Section 7.1.2 as well as in Problems 7.2–7.7. Bhattacharya

and Dunson (2010) studied the Kullback-Leibler property of kernel mixture priors on manifolds, considering the complex Bingham and Watson kernels. The Kullback-Leibler property and consistency for the multivariate normal kernel were studied by Wu and Ghosal (2010). Consistency for the infinite-dimensional exponential family prior was studied by Barron et al. (1999) in the case of a polynomial basis and a priori normally distributed coefficients, although their proof appears to be based on an incorrect estimate. The inconsistency example in Example 7.12 is due to Barron et al. (1999). Walker et al. (2005) called this phenomenon data tracking, and provided more insight on it. Theorem 7.13 is due to Ghosal et al. (1999b). We learned the construction of a Dirichlet mixture sieve based on the stick-breaking property of a Dirichlet process used in the proof of Theorem 7.15 from Tokdar (personal communication). The results in Sections 7.2.3 and 7.2.4 are based on Barron et al. (1999). Consistency for nonparametric binary regression, or estimation of response probability, was addressed by Ghosal and Roy (2006), who specifically used Gaussian process priors. Coram and Lalley (2006) established consistency for priors supported on step functions using a delicate large deviation technique. Similar results for categorical, but infinitely many covarites were obtained earlier by Diaconis and Freedman (1993). Normal regression was considered by Choi and Schervish (2007), who also used Gaussian process priors, but used somewhat different methods. Spectral density estimation using Whittle's likelihood is common in the frequentist literature. In the Bayesian context, Carter and Kohn (1997), Gangopadhyay et al. (1999) and Liseo et al. (2001) used the Whittle likelihood with the integrated Brownian motion or free-knot spline based priors. Consistency under this setting using the contiguity technique is due to Choudhuri et al. (2004a), who specifically used the Bernstein polynomial prior. Theorem 7.26 is due to Ghosal et al. (1999a), who pointed out that the Diaconis-Freedman-type inconsistency can be avoided in the location problem by using a prior with the Kullback-Leibler property. Theorems 7.27 and 7.28 are due to Amewou-Atisso et al. (2003). Consistency proofs for many other problems, including the Cox model and Problems 7.18–7.22 can be found in Wu and Ghosal (2008b). The estimate in Problem 7.23 is adapted from Walker (2004) who considered density estimation in the i.i.d. case and used an infinite-dimensional exponential family.

Problems

7.1 (Barron et al. 1999) For infinite-dimensional exponential families in Section 2.3, positivity of the prior probability of Kullback-Leibler neighborhoods of p_0 can be verified under weaker conditions on p_0. Show that the uniform approximation requirement and continuity can be relaxed to $K(p_0;\lambda) < \infty$ and $\mathbb{L}_1$-approximation of functions bounded by B by finite basis expansion satisfying a bound rB, where $r > 1$ can be chosen to work for all possible B.

7.2 (Wu and Ghosal 2008a) Show that for the double exponential and the logistic kernel, the conditions of Theorem 7.3 hold under finiteness of $(1+\eta)$-moment of p_0 for some $\eta > 0$.

7.3 (Wu and Ghosal 2008a) Show that for the t-kernel, conditions of Theorem 7.3 hold under $\int \log_+ |x| p_0(x)\,dx < \infty$.

7.4 (Wu and Ghosal 2008a) Consider the triangular density kernel given by

$$\psi(x; m, n) = \begin{cases} \begin{cases} 2n - 2n^2 x, & x \in (0, \frac{1}{n}), \\ 0, & \text{otherwise}, \end{cases} & m = 0, \\ \begin{cases} n^2(x - \frac{m}{n}) + n, & x \in (\frac{m-1}{n}, \frac{m}{n}), \\ -n^2(x - \frac{m}{n}) + n, & x \in (\frac{m}{n}, \frac{m+1}{n}), \\ 0, & \text{otherwise}, \end{cases} & m = 1, 2, \ldots, n-1, \\ \begin{cases} 2n + 2n^2(x - 1), & x \in (0, \frac{1}{n}), \\ 0, & \text{otherwise}, \end{cases} & m = n. \end{cases}$$

Construct a kernel mixture prior by mixing both m and n according to a prior with weak support $\mathfrak{M}(\{(m, n): m \le n\})$. If p_0 is a continuous density on $[0, 1]$, then $p_0 \in \mathrm{KL}(\Pi^*)$.

7.5 (Wu and Ghosal 2008a) For the Weibull kernel, using log transformation to reduce the problem to a location-scale mixture, show that a nonzero, bounded, continuous density on $(0, \infty)$, f_0, is in the Kullback-Leibler support of the mixture prior if $\log f_0(x)$, $e^{2|\log x|^{1+\eta}}$ and $\log(f_0(x)/\phi_\delta(x))$ are f_0-integrable for some $\eta, \delta > 0$, where $\phi_\delta(x) = \inf_{|t-x|<\delta} f_0(t)$.

7.6 (Wu and Ghosal 2008a) Consider (a reparameterized) inverse-gamma density kernel defined by $\psi(x; \beta, z) = (\beta z)^\beta x^{-\beta-1} e^{-\beta z/x} / \Gamma(\beta)$. Consider a kernel mixture prior by mixing (β, z) according to a prior Π with weak support $\mathfrak{M}((2, \infty) \times \mathbb{R}^+)$. Find conditions for a continuous density p_0 on $[0, \infty)$ to satisfy $p_0 \in \mathrm{KL}(\Pi^*)$, where Π^* is the induced prior distribution of $p_F = \int \psi(\cdot; \beta, z)\, dF(\beta, z)$ for $F \sim \Pi$.

7.7 (Wu and Ghosal 2008a) Consider a kernel mixture prior based on the scaled uniform kernel $\psi(x; \theta) = \theta^{-1} \mathbb{1}\{0 \le x \le \theta\}$, where $\theta > 0$ is mixed according to F having a prior Π with weak support $\mathfrak{M}(\mathbb{R}^+)$. If p_0 is a continuous and decreasing density function on $\mathbb{R}^+$ such that $\int p_0(x) |\log p_0(x)|\, dx < \infty$, then show that $p_0 \in \mathrm{KL}(\Pi^*)$, where Π^* is the induced prior distribution of $p_F = \int \psi(\cdot; \theta)\, dF(\theta)$ for $F \sim \Pi$.

7.8 (Tokdar 2006) If the kernel is univariate normal location-scale mixture and the mixing distribution follows the Dirichlet process, show that the moment condition $\int |x|^{2(1+\eta)} p_0(x)\, dx < \infty$ for some $\eta > 0$, in Example 7.4 can be weakened to $\int |x|^\eta p_0(x)\, dx < \infty$ for some $\eta > 0$.

7.9 (Wu and Ghosal 2010) Consider a multivariate normal mixture prior $p \sim \Pi^*$ defined by $p(x) = \int \phi_d(x; \theta, \Sigma)\, dF(\theta)$, $F \sim \Pi$ and $\Sigma \sim \mu$, where Π has support $\mathfrak{M}(\mathbb{R}^d)$ and the prior μ for the scale matrix Σ has σI_d in its support for all sufficiently small $\sigma > 0$. Assume that the true density p_0 is everywhere positive, bounded above, that the integrals $\int p_0(x) |\log p_0(x)|\, dx$ and $\int p_0(x) \log(p_0(x)/\phi_\delta(x))\, dx$ are finite, for some $\delta > 0$, where $\phi_\delta(x) = \inf\{p_0(t): \|t - x\| < \delta\}$ and $\int \|x\|^{2(1+\eta)} p_0(x)\, dx < \infty$ for some $\eta > 0$. Show that $p_0 \in \mathrm{KL}(\Pi^*)$.

7.10 (Wu and Ghosal 2010) Consider the setting of Problem 7.9 with Π being α and Σ^m, $m \ge 2d$, has a Wishart distribution with nonsingular scale matrix A and degrees of freedom $q > d$, truncated to satisfy $\mathrm{tr}(\Sigma) \le M$ for some fixed $M > 0$. Show that the posterior is consistent at p_0. Further, if p_0 has sub-Gaussian tails, show that

$m > d$ suffices. [Hint: show that $\Pi^*(F(\|\theta\|_\infty > 2a_n) \geq \epsilon_n | X_1, \ldots, X_n) \to 0$ in P_0^n-probability whenever $a_n \to \infty$, $n^{-1}a_n^{2d}\epsilon_n \to \infty$ and $n^{-1}\epsilon_n \inf\{g(t): \|t\| \leq a_n\} e^{a_n^2/(2M)} \to \infty$, where g is the density of $\bar{\alpha}$. Now use Problem C.1.]

7.11 (Ghosal and Roy 2006) In Theorem 7.20, assume all conditions hold and the covariate is deterministic and one-dimensional. Further assume that given $\delta > 0$, there exist a constant K_1 and an integer N such that for $n > N$, we have that $\sum_{i:S_{i,n} > K_1 n^{-1}} S_{i,n} \leq \delta$, where $S_{i,n} = x_{i+1,n} - x_{i,n}$ are the spacings between consecutive covariate values. Show that the posterior for f is consistent at f_0 in the usual $\mathbb{L}_1$-distance $\int |f_1(x) - f_2(x)|\, dx$.

7.12 Derive a consistency result for nonparametric binary regression from that on the densities using Theorems 6.23 and 6.41 for stochastic and deterministic regressors, respectively.

7.13 Formulate and prove a consistency result for nonparametric Poisson regression given by $Y | X = x \sim \text{Poi}(\lambda(x))$.

7.14 (Coram and Lalley 2006) This example shows sometimes it is beneficial to directly bound the posterior probability of the complement of the sieve rather than using Theorem 6.17. In this case, delicate large deviation estimates are used.

Consider the binary regression problem $Y | X = x \sim \text{Bin}(1, f(x))$, where X has a nonsingular distribution on $[0, 1]$. Let (X_i, Y_i), $i = 1, 2, \ldots$ be i.i.d. observations. Consider a series prior for f defined as follows: A positive integer-valued random variable $N \sim \lambda$, where $\lambda_m > 0$ for infinitely many m, $U_1, \ldots, U_m | N = m \overset{\text{iid}}{\sim} \text{Unif}(0, 1)$, $W_1, \ldots, W_m | N = m \overset{\text{iid}}{\sim} \text{Unif}(0, 1)$ and $f(x) = \sum_{j=1}^{m+1} W_j \mathbb{1}\{U_{j-1:m} < x \leq U_{j:m}\}$, where $U_{j:m}$ is the jth order statistic among m, $j = 1, \ldots, m$, and $U_{0:m} = 0$, $U_{m+1:m} = 1$. Let the true response probability f_0 be a measurable function from $[0, 1]$ to $[0, 1]$ not identically $1/2$.

The posterior distribution for f given $\mathscr{F}_n := \sigma\langle (X_1, Y_1), \ldots, (X_n, Y_n)\rangle$ can be written as

$$\Pi(\cdot | \mathscr{F}_n) = \frac{\sum \lambda_m Z_{m,n} \Pi_m(\cdot | \mathscr{F}_n)}{\sum \lambda_m Z_{m,n}},$$

where $\Pi_m(\cdot | \mathscr{F}_n)$ is the "posterior based on model $N = m$" and "$Z_{m,n}$ is the marginal likelihood of the model $N = m$."

Show that if $m/n \to \alpha$, then $n^{-1} \log Z_{m,n} \to \psi(\alpha) < 0$ for all $\alpha > 0$, where the function $\psi(\alpha)$ has a unique maximum 0 at $\alpha = 0$. In other words, models with more than ϵn discontinuities for any $\epsilon > 0$ will have marginal likelihood decaying exponentially.

Show that if $K \leq m < \epsilon n$, then the likelihood of $N = m$ given values of U's decays exponentially like $\exp[-n\{f_u \log f_u + (1 - f_u) \log(1 - f_u)\}]$, where f_u is the step function obtained by averaging f_0 over the intervals defined by $U_1, \ldots, U_m$.

Show that the entropy of all models with $N \leq K$ grows logarithmically, and hence Schwartz's theory applies to give consistency at f_0 using only these models.

Combining the above three conclusions, obtain consistency of the posterior at any measurable f_0 not identically $1/2$.

7.15 (Diaconis and Freedman 1993) Consider a binary response variable Y and infinitely many binary predictors $X = (X_1, X_2, \ldots)$ with response function at $X = x$ given by $f(x)$, where $f: \{0, 1\}^\infty \to [0, 1]$ is measurable. Consider a series prior for f defined as follows: A positive integer-valued random variable $N \sim \lambda$, where $\lambda_m > 0$ for

infinitely many m. Given $N = m$, $f(x)$ depends only on $(x_1, \ldots, x_m)$ and let Π_m be the joint prior distribution of $(\theta_\varepsilon : \varepsilon \in \{0,1\}^m)$, where $\theta_\varepsilon = f(\varepsilon)$. Let the true response probability be f_0. Suppose that at stage n we observe Y for every of the 2^n combinations of covariate values. Let the metric on f be given by the $\mathbb{L}_1$-distance $\int |f(x) - g(x)|\, d\lambda(x)$, where λ is the infinite product of i.i.d. Bin(1, 1/2).

Show that if f_0 is not identically $\frac{1}{2}$, then the posterior is consistent.

Show that if λ_ms are given by geometric with success probability $1 - r$, then the posterior is consistent at $f_0 \equiv 1/2$ whenever $r < 2^{-1/2}$ and inconsistent whenever $r > 2^{-1/2}$.

7.16 In the semiparametric linear regression problem, if random fs are not symmetrized around zero, then α is not identifiable. Nevertheless, show that uniformly consistent tests for β can be obtained by considering the difference $Y_i - Y_j$, which has a density that is symmetric around $\beta(x_i - x_j)$, and hence consistency for β may be obtained.

7.17 (Ghosal et al. 1999a) In the location problem, consider a symmetrized Dirichlet mixture of normal prior Π on the error density f. Assume that the true error density f_0 is in the Kullback-Leibler support of Π, $\int z^2 f_0(z)\, dz < \infty$, $\int f_0 \log_+^2 f_0 < \infty$ and the base measure of the Dirichlet process has finite second moment. Show that the posterior for θ is consistent at any θ_0 belonging to the support of the prior for θ.

7.18 (Wu and Ghosal 2008b) Consider a multiple regression model $Y_i = \alpha + X_i^\top \beta + \varepsilon_i$, $X_i \overset{\text{iid}}{\sim} Q \in \mathfrak{M}(\mathbb{R}^d)$, independently $\varepsilon_i \overset{\text{iid}}{\sim} f$, $i = 1, 2, \ldots$, $\beta \in \mathbb{R}^d$, and $f \in \mathcal{F} = \{f : \int_{-\infty}^0 f(x)\, dx = a\}$ for some $0 < a < 1$ (i.e. has the same value of some fixed quantile). Let $f \sim \tilde{\Pi}$, independently $(\alpha, \beta) \sim \mu$. Let Π stand for $\tilde{\Pi} \times \mu$. Let $f_{\alpha,\beta} = f(y - \alpha - x^\top \beta)$, $p_0 = f_{0,\alpha_0,\beta_0}$, and $g_\theta(y) = g(y - \theta)$. Assume that the covariate X is compactly supported and f_0 is continuous with $f_0(0) > 0$.

Assume that

(a) For some $\zeta > 0$, $Q(\gamma_j X_j > \zeta,\ j = 1, \ldots, d) > 0$ for all $\gamma_j = \pm 1$;

Show that there exist exponentially consistent tests for testing $H_0 : (f, \alpha, \beta) = (f_0, \alpha_0, \beta_0)$ against $H_1 : \{(f, \alpha, \beta) : f \in \mathcal{U}, |\alpha - \alpha_0| < \epsilon, \|\beta - \beta_0\| < \epsilon\}$, where $\mathcal{U}$ is a weak neighborhood of f_0.

Further, assume that

(b) for $\eta > 0$ sufficiently small, there exists $g_\eta \in \mathcal{F}$ and constant $C_\eta > 0$ such that for $|\eta'| < \eta$, $f_0(y - \eta') < C_\eta g_\eta(y)$ for all y and $C_\eta \to 1$ as $\eta \to 0$;

(c) for all sufficiently small η and all $\xi > 0$, $\tilde{\Pi}\{K(g_\eta; f) < \xi\} > 0$ and $(\alpha_0, \beta_0) \in \text{supp}(\mu)$.

Show that for any weak neighborhood $\mathcal{U}$ of f_0, we have $\Pi\{(f, \alpha, \beta) : f \in \mathcal{U}, |\alpha - \alpha_0| < \epsilon, \|\beta - \beta_0\| < \epsilon | (X_1, Y_1), \ldots, (X_n, Y_n)\} \to 1$ a.s. P_0^∞. Let the covariate X be deterministic and assume values $x_1, \ldots, x_n$, replace Condition (a) by the following deterministic version: for some $\zeta > 0$

$$\liminf n^{-1} \#\{i : \gamma_j x_{ij} > \zeta,\ j = 1, \ldots, d\} > 0 \text{ for all } \gamma_j = \pm 1. \tag{7.6}$$

Show that exponentially consistent tests for testing the pair H_0 and H_1 defined above, exist. If, moreover, $V_2(g_\eta; f) < \infty$ a.s. $[\Pi]$, then the posterior for (f, α, β) is consistent at (f_0, α_0, β_0).

7.19 (Wu and Ghosal 2008b) In the multiple regression model of Problem 7.18 with stochastic covariates, suppose that we use a symmetrized Dirichlet mixture of normal prior given by $f(y) = f_{h,G}(y) = (\int \phi_h(y-t)\,dG(t) + \int \phi_h(y+t)\,dG(t))/2$, $G \sim \mathrm{DP}(\pi)$ and $h \sim \nu$. Show that Condition (ii) can be replaced by simpler moment conditions, namely, $\int y^2 p_0(y)\,dy < \infty$, $\int p_0 |\log p_0| < \infty$ and $\int t^2 d\pi(t) < \infty$. For deterministic covariates, show that Conditions (b) and (c) hold if $\int y^4 p_0(y)\,dy < \infty$ and $\int p_0(y)\log^2(y) p_0(y)\,dy < \infty$.

7.20 (Wu and Ghosal 2008b) Consider the exponential frailty model described by

$$(X_i, Y_i)|\, W_i \overset{\text{ind}}{\sim} (\mathrm{Exp}(W_i), \mathrm{Exp}(\lambda W_i)) \text{ independently, } W_i \overset{\text{iid}}{\sim} F,$$

and priors are given by $F \sim \tilde{\Pi}$ and $\lambda \sim \mu$ independently. Let the p.d.f. of (X, Y) be $p_{\lambda,F}(x, y) = \int \lambda_0 w^2 e^{-w(x+\lambda_0 y)} dF(w)$ and $p_0 = p_{\lambda_0, F_0}$, where F_0 is the true value of the frailty distribution F and λ_0 the true value of the frailty parameter λ.

Show that there exist exponentially consistent tests for testing $H_0: \lambda = \lambda_0$ against $H_1: |\lambda - \lambda_0| > \epsilon$ for any $\epsilon > 0$, and also for $H_0: \lambda = \lambda_0$, $F = F_0$ against $H_1: |\lambda - \lambda_0| \le \epsilon$, $F \in \mathcal{U}^c$, $F(r_n) - F(r_n^{-1}) \ge 1 - \delta$ for some $\delta > 0$, weak neighborhood $\mathcal{U}$ of F_0 and sequence $r_n < \sqrt{n\beta}$ for some sufficiently small $\beta > 0$.

Further assume that $\log(f_0(x, y))$, x and y are f_0-integrable, $F_0 \in \mathrm{supp}(\tilde{\Pi})$, $\lambda_0 \in \mathrm{supp}(\mu)$ and $w_E := \int w^2 dF_0(w) < \infty$. If for any $\delta > 0$ and $\beta > 0$ there exists a sequence $r_n < \sqrt{n\beta}$ and $\beta_0 > 0$ such that $\tilde{\Pi}\{F: F(r_n) - F(r_n^{-1}) < 1 - \delta\} < e^{-n\beta_0}$, then show that the posterior for (λ, F) is consistent at (λ_0, F_0).

7.21 (Wu and Ghosal 2008b) Consider a generalized linear model $X_i \overset{\text{iid}}{\sim} Q$, $Y_i | X_i \sim \exp[\theta_i y - b(\theta_i)]a(y)$, where $\mathrm{E}(Y_i | X_i) = b'(\theta_i) = g(X_i^\top \beta)$, for $i = 1, \ldots, n$, and $\|\beta\| = 1$ and g is differentiable and strictly increasing. Assume that $\|X\| \le L$. Let $\mathcal{G}$ stand for the space of all possible link functions, and $M := \{\mu: \mu = g(x^\top \beta), \|x\| \le L, \|\beta\| = 1\}$ be a compact interval. Let $T: M \to [0, 1]$ be a fixed strictly increasing and differentiable function and put a prior on g by $g = F \circ T^{-1}$, where F is a c.d.f. on $[-L, L]$ following a prior $\tilde{\Pi}$ distributed independently of β. Let $\Pi = \tilde{\Pi} \times \pi$. Assume that the covariates satisfy Condition (a) in Problem 7.18. Furthermore, assume that the following conditions hold:

- $\int \tilde{f}_0(t) |\log \tilde{f}_0(t)| dt < \infty$ and $\int \tilde{f}_c(t + \xi) |\log \bar{\tilde{F}}_0(t)| dt < \infty$ for all $\xi \in \mathbb{R}$;
- for some $\delta > 0$, $\int \tilde{f}_0(t) \log(\tilde{f}_0/\phi_\delta)(t)\,dt < \infty$, where $\phi_\delta(t) := \inf\{\tilde{f}_0(s): |s - t| < \delta\}$;
- there exists $\eta > 0$ such that $\int (e^{2|\log t|^\eta} + e^{(\log t - a)/b}) f_0(t)\,dt < \infty$, for any $a \in \mathbb{R}$ and $b \in (0, \infty)$;
- for Π-almost all P and any given $h > 0$, $\int (e^{-t/h} + t/h)\,dP(t) < \infty$;
- the weak support of $\tilde{\Pi}$ is the space of all probability measures on $\mathbb{R}$;
- for some $\delta > 0$ and any $\xi \in \mathbb{R}$, $\int \tilde{f}_c(t + \xi) \log(\bar{\tilde{F}}_{0,\beta_0}/\phi_\delta)(t)\,dt < \infty$;
- there exists $\eta > 0$ such that $\int f_c(t)(e^{2(\log t)^{1+\eta}} + e^{(\log t - a)/b})\,dt < \infty$, for $a \in \mathbb{R}$ and $b > 0$.

(a) Show that g is identifiable.

(b) Show that for any $\epsilon > 0$, there exists an exponentially consistent sequence of tests for testing $H_0: (\beta, F) = (\beta_0, F_0)$ against $H_1: \|\beta - \beta_0\| > \epsilon$, $F \in \mathcal{U}$, for some weak neighborhood $\mathcal{U}$ of F_0.

(c) Show that for any weak neighborhood $\mathcal{U}$ of F_0, almost surely $[P^{\infty}_{\beta_0, f_0}]$,

$$\Pi\Big((\beta, F): \|\beta - \beta_0\| < \epsilon, \tilde{F} \in \mathcal{U}, \,|\, (Z_1, X_1, \Delta_1), \ldots, (Z_n, X_n, \Delta_n)\Big) \to 1.$$

(d) Show that the Kullback-Leibler property holds at the true density of Z.
(e) Conclude that the posterior is consistent at (β_0, f_0).

[Hint: Show that there is an exponentially consistent sequence of tests for testing

- $H_0: (\beta, F) = (\beta_0, F_0)$ against $H_1: F \notin \mathcal{U}$, for any weak neighborhood $\mathcal{U}$ of F_0;
- $H_0: (\beta, F) = (\beta_0, F_0)$ against $H_1: d_{KS}(F, F_0) < \Delta, \|\beta - \beta_0\| > \delta$.]

7.22 (Wu and Ghosal 2008b) Consider a partial linear model $Y_i = X_i^{\top}\beta + f(Z_i) + \varepsilon_i$, where the covariates $X_i \in \mathbb{R}^d$ and $Z_i \in \mathbb{R}$, errors $\varepsilon_i \overset{\text{iid}}{\sim} \text{Nor}(0, \sigma^2)$ for $i = 1, \ldots, n$, and f is an odd function. Assume that X has distribution $Q_X(x)$ and Z has symmetric p.d.f. $q_Z(z)$. Let $q_{f,Z}$ denote the p.d.f. of $f(Z)$. Assume that X and Z are compactly supported and Condition (a) in Exercise 7.18 holds. Consider fully supported priors μ for β, Π^* for $f(Z)$ and ρ for σ. Let $\tilde{\Pi}$ stand for the prior induced on $q_{f,Z} * \phi_\sigma$ and $\Pi = \tilde{\Pi} \times \mu$. Let f_0, β_0 and σ_0 denote the true values of f, β and σ respectively. Assume that

(a) there exists $\eta > 0$, C_η and a density g_η, such that $q_{f_0,Z} * \phi_{\sigma_0}(y - \eta') < C_\eta g_\eta(y)$, for $|\eta'| < \eta$;
(b) for all sufficiently small η and for all $\delta > 0$, $\tilde{\Pi}\{f: K(g_\eta; f) < \delta\} > 0$.

Show that the posterior is consistent at (β_0, f_0, σ_0). [Hint: Consider $\xi = f(Z) + \varepsilon$ as the error variable in a multiple regression problem $Y = X^{\top}\beta + \xi$ and apply Exercise 7.18. The oddness of f and symmetry of $q_Z(z)$ ensures that ξ is symmetrically distributed about 0.]

7.23 Consider a Markov process with transition density $f(y|x) = g(y - \rho x)$, where g is a density function on $[0, 1]$ and $-1 < \rho < 1$ is known. Assume that the true density g_0 is continuous and bounded away from 0, so that the corresponding Markov process is ergodic. Consider a finite random series prior based on trigonometric polynomials on g as described in Example 7.19: $g(u; \theta) = \exp[\sum_{j=0}^{N} \theta_j \psi_j(u) - c(\theta)]$, for $\psi_0(u) = 1$ and $\psi_j(u) = \sqrt{2}\cos(j\pi x)$ if $j \in \mathbb{N}$; $\text{P}(N = k) = \rho_k$ for $\sum_{k=1}^{\infty} \sqrt{\rho_k} < \infty$; and $\theta_j \overset{\text{ind}}{\sim} \text{Nor}(0, \tau_j^2)$, where $\tau_j \lesssim j^{-q}$ for some $q > 1$.

For given $N = k$ consider a countable partition $\{\mathcal{A}_{n_1,\ldots,n_k}, n_j \in \mathbb{Z}, k \in \mathbb{N}\}$ of the form $\{g(y - \rho x): g \in \mathcal{B}_{n_1,\ldots,n_k}, n_j \in \mathbb{Z}, k \in \mathbb{N}\}$. Let $\gamma_j = cj^{-\beta}$ be such that $\sum_{j=1}^{\infty} \gamma_j \le 1$. Let $B_{n_1,\ldots,n_k} = \prod_{j=1}^{k} (\delta n_j \gamma_j, \delta(n_j + 1)\gamma_j]$ and $\mathcal{B}_{n_1,\ldots,n_k} = \{g(\cdot; (\theta_1, \ldots, \theta_k)): (\theta_1, \ldots, \theta_k) \in B_{n_1,\ldots,n_k}\}$.

Show that $d_H(g, g^*) \le 2\sqrt{\delta}$ if $g, g^* \in \mathcal{B}_{n_1,\ldots,n_k}$.
[Hint: Integrate $|\exp[\sum \theta_j \psi_j(x)] - \exp[\sum \theta_j^* \psi_j(x)]|$ to obtain $|e^{c(\theta)} - e^{c(\theta^*)}| < 2\delta e^{c(\theta^*)}$ for $\delta < \log 2$. Hence obtain $|c(\theta) - c(\theta^*)| < 2\delta$. Now bound the Kullback-Leibler divergence and hence the Hellinger distance.]

Show consistency of the posterior by verifying that

$$\sum_{k=1}^{\infty} \sum_{n_1 \in \mathbb{Z}} \cdots \sum_{n_k \in \mathbb{Z}} \rho_k^{1/2} \prod_{j=1}^{k} \sqrt{\Pi(\delta n_j \gamma_j < \theta_j \le \delta(n_j + 1)\gamma_j]} < \infty.$$