CHAPTER 4. ASYMPTOTICS OF POSTERIORS AND MODEL SELECTION

4.1 Consistency of posteriors. Given a measurable family $\{P_\theta,\ \theta \in \Theta\}$, dominated by a σ-finite measure v, for a measurable space $(\Theta, \mathcal{T})$, a prior π on Θ, and observations $X_1,\ X_2, \ldots$ i.i.d. (P_{θ_0}) for some $\theta_0 \in \Theta$, we have posteriors $\pi_{x,n}$ on Θ where $x = (X_1, \ldots, X_n)$ for each n. Recall that we have defined the posteriors by multiplying the prior by the likelihood function $\Pi_{j=1}^n f(\theta, X_j)$ and normalizing the result, if possible, to be a probability measure (Proposition 1.3.5). (For Theorem 4.1.4 below, where $\{P_\theta,\ \theta \in \Theta\}$ is not necessarily dominated, a more general definition of posteriors will be used.) The posteriors will be called *consistent* if for every neighborhood U of θ_0, $\pi_{x,n}(U) \to 1$ almost surely as $n \to \infty$. This form of consistency is not for estimators T_n, but is just a property of the prior and the likelihood function.

In some situations, consistency of posteriors can lead to consistency of estimators. For example, if Θ is an interval in $\mathbb{R}$, and T_n is a median of the posterior law $\pi_{x,n}$, then consistency of posteriors will imply that T_n are consistent. If the interval is bounded, T_n could also be taken as the mean of $\pi_{x,n}$.

If the prior π has $\pi(U) = 0$ for some neighborhood U of the true parameter θ_0, then $\pi_{x,n}(U) = 0$ for all x and n, so the posteriors can't be consistent. On the other hand, if $\pi(U) > 0$ for every neighborhood U of θ_0, then under some conditions as in Section 3.3, it will be shown that the posteriors are consistent. It can happen in pathological cases that the posteriors are not consistent, for example if as the neighborhoods U shrink to $\{\theta_0\}$, $\pi(U) \to 0$ very fast, and if the likelihood function doesn't behave well. Such an example will be given in Proposition 4.1.2 and after it.

4.1.1 Theorem. Assume that:

(i) $\{P_\theta,\ \theta \in \Theta\}$ is a measurable family, dominated by a σ-finite measure v, and identifiable, so that $P_\theta \neq P_\phi$ for $\theta \neq \phi$;

(ii) Θ is a locally compact separable metric space, with Borel σ-algebra $\mathcal{T}$,

(iii) $(dP_\theta/dv)(x) \equiv f(\theta, x)$ where $f(\cdot, \cdot)$ is jointly measurable,

(iv) $P = P_{\theta_0}$ for some $\theta_0 \in \Theta$, and $X_1, X_2, \ldots$, are i.i.d. (P);

(v) $h(\cdot, x) := \log f(\theta_0, x) - \log f(\cdot, x)$ is continuous on Θ,

(vi) For some positive, continuous function $b(\cdot)$ on Θ and integrable function $u(\cdot)$ on X for P_{θ_0}, $|h(\theta, x)| \leq b(\theta)u(x)$ for all θ and almost all x,

(vii) (3.3.6) and (3.3.7) hold for the given h and $b(\cdot)$, that is, $\lim_{\theta\to\infty} b(\theta) > \gamma(\theta_0) = 0$ and $E[\liminf_{\theta\to\infty} h(\theta, x)/b(\theta)] \geq 1$.

Then for any prior π such that $\pi(U) > 0$ for every neighborhood U of θ_0, the posteriors are consistent.

Notes. In (v), $h(\cdot, \cdot)$ has been chosen to incorporate an adjustment function so that, in the notation of Section 3.3, $a(x) \equiv 0$. Here continuity of h in θ is assumed in (v), rather than the lower semicontinuity assumed in Section 3.3, (A-2). This is needed in order to allow general priors. Suppose that $f(\theta_0, x)$ were the lim sup, not the limit, of $f(\theta, x)$ as $\theta \to \theta_0$ and that for some $\varepsilon > 0$ and sequence $\theta_k \to \theta_0$, $f(\theta_k, x) < f(\theta_0, x) - \varepsilon$ for x in a set A with $P(A) > 0$ and all k. Then if the prior π is concentrated on points θ where

$f(\theta, x) < f(\theta_0, x) - \varepsilon$ for $x \in A$, we can have $\pi(U) > 0$ for every neighborhood U of θ_0, but the posteriors might not be consistent.

Proof. Continuity of $h(\cdot, x)$ implies assumptions (A-1) and (A-2) of Section 3.3, where S is any countable dense set in Θ and $A = \emptyset$. Assumption (vi) implies that $E_{\theta_0}|h(\theta, x)| < \infty$ for all θ, which is stronger than (A-3). In Theorem 3.3.16, using identifiability, (A-4) is shown to hold in this case. Condition (vi) is stronger than (3.3.5), and the other parts of (A-5) are assumed, so all of (A-1) through (A-5) hold. Lemma 3.3.9 doesn't involve any estimators T_n, so it still holds. Also, now that continuity of $h(\cdot, x)$ and (vi) are assumed, the Lemma also holds with sup instead of inf, so $\gamma(\cdot)$ is continuous. Note that $\gamma(\theta_0) = 0$ in the present case.

For any neighborhood U of θ_0, there is an $\varepsilon > 0$ such that almost surely for n large enough,

$$\Pi_{i=1}^n f(\theta, X_i)/f(\theta_0, X_i) \ \leq \ \exp(-n\varepsilon)$$

for all $\theta \notin U$: this follows from the proof of (3.3.14) for θ in some compact set C and from (3.3.12) for $\theta \notin C$. These proofs do not involve T_n. On the other hand, for a small enough neighborhood $V \subset U$, by Lemma 3.3.9, almost surely for n large enough, by the strong law of large numbers, for each $\theta \in V$,

$$\Pi_{i=1}^n f(\theta, X_i)/f(\theta_0, X_i) \ \geq \ \exp(-n\varepsilon/2).$$

Then for each $\phi \notin U$ and $\theta \in V$, the likelihood ratio for n observations satisfies

$$R_{\theta,\phi}(X_1, \ldots, X_n) \ \geq \ e^{n\varepsilon/2}.$$

Since $\pi(V) > 0$, the ratio of posteriors $\pi_{x,n}(V)/\pi_{x,n}(\Theta \setminus U) \to \infty$, which implies that $\pi_{x,n}(U) \to 1$. □

The following will give examples where posteriors are not consistent:

4.1.2 Proposition. Suppose $\{P_\theta, \ \theta \in \Theta\}$ is a family with densities $f(\theta, x)$ such that for a metric d on Θ, some $\theta_0 \in \Theta$, $P = P_{\theta_0}$, and a sequence θ_m converging to θ_0 for d, we have

$$0 \ < \ a_m \ := \ I(P_{\theta_0}, P_{\theta_m}) = -E \log(f(\theta_m, \cdot)/f(\theta_0, \cdot)) \uparrow C$$

strictly as $m \to \infty$ where $0 < C \leq +\infty$. Then there is a prior π on the sequence $\{\theta_m\}$ with $\pi(\theta_m) > 0$ for all $m \geq 1$, and so with $\pi(U) > 0$ for every neighborhood U of θ_0, such that for $X_1, X_2, \ldots$ i.i.d. P, the posteriors are not consistent.

Proof. Let $Y_m(x) := \log(f(\theta_m, x)/f(\theta_0, x))$. Since $a_m = -EY_m \geq 0$ by Lemma 3.3.15 and $a_m < C \leq +\infty$, it follows that $Y_m > -\infty$ a.s., so $f(\theta_m, \cdot) > 0$ a.s. Let $\overline{Y}_{mn} := \sum_{j=1}^n Y_m(X_j)/n$. Then $\overline{Y}_{mn} \to -a_m$ a.s. as $n \to \infty$ by the strong law of large numbers, so for each $m = 1, 2, \ldots$, there is an $n_0 := n_0(m)$ such that

$$\Pr\{\text{for some } \ n \geq n_0 : \ \overline{Y}_{mn} \leq -a_m - (a_{m+1} - a_m)/3 \ \text{ or}$$
$$\overline{Y}_{m+1,n} \ \geq \ -a_{m+1} + (a_{m+1} - a_m)/3\} \ < \ 1/2^m. \tag{4.1.3}$$

Also choose $n_0(m)$ large enough so that $\exp(-n_0(m)(a_{m+1} - a_m)/3) < 1/2$. Let $L_{nm} := \Pi_{j=1}^{n} f(\theta_m, X_j)$. Since $f(\theta_m, X_j) > 0$ a.s. for each j, there is always an $a_{nm} > 0$ such that

$$\Pr(L_{nm} < a_{nm}) \ < \ 1/(2^m n_0(m)).$$

Also, there are always $b_{nm} < \infty$ large enough so that

$$\Pr(L_{nm} \geq b_{nm}) \ < \ 1/(2^m n_0(m)).$$

Let $r_m := 2 + 2 \cdot \max\{b_{n,m+1}/a_{nm} : \ n < n_0(m)\}$. Then the prior π will be defined so that $\pi(\theta_{m+1})/\pi(\theta_m) = 1/r_m$. In other words, for the suitable normalizing constant c, let $\pi(\theta_1) := c$ and $\pi(\theta_m) := c/\Pi_{j=1}^{m-1} r_j$ for $m \geq 2$. Then for each $n < n_0(m)$, the posterior probability $\pi_{x,n}(\theta_{m+1})$ can be larger than $\pi_{x,n}(\theta_m)/2$ only if $L_{n,m+1}/L_{nm} > r_m/2$, which requires either $L_{n,m+1} > b_{n,m+1}$ or $L_{nm} < a_{nm}$. For a given m, the probability that any of these events occur for $n < n_0(m)$ is at most $2/2^m$. Also, except on the event in (4.1.3), we have for $n \geq n_0(m)$ that $\overline{Y}_{m+1,n} < \overline{Y}_{mn}$ and so $L_{n,m+1}/L_{nm} < 1$. So, for each m, except on an event of probability at most $3/2^m$, we have $\pi_{x,n}(\theta_{m+1}) \leq \pi_{x,n}(\theta_m)/2$ for all n. Since $\sum_m 3/2^m < \infty$, by the Borel-Cantelli lemma (RAP, Theorem 8.3.4), almost surely there is an m_0 such that for all $m \geq m_0$ and all n, $\pi_{x,n}(\theta_{m+1}) \leq \pi_{x,n}(\theta_m)/2$. Then $\sum_{m>m_0} \pi_{x,n}(\theta_m) \leq \pi_{x,n}(\theta_{m_0})$, so the left side of the latter inequality can't converge to 1, and the posteriors can't be consistent. □

For a specific example where the last proposition applies, let the sample space be the open interval $0 < x < 1$ with v = Lebesgue measure. For $m \geq 2$ let f_m be continuous, $f_m(x) := e^{-m}$ for $0 < x \leq 1/m$, let f_m be linear on the interval $1/m \leq x \leq 1/m + e^{-m}$ and let f_m be constant for $1/m + e^{-m} \leq x \leq 1$. The constant is $1 + 1/m + o(1/m)$. (A simpler example could be defined, constant on $(0, 1/m)$ and on $[1/m, 1)$.) Let $\theta_m := 1/m$ and $f(\theta_m, x) := f_m(x)$. Let $\theta_0 := 0$ and $f(0, x) \equiv 1$, giving the uniform distribution. Then $f_m(x) \to 1$ as $m \to \infty$ for $0 < x < 1$, so $f(\cdot, x)$ is continuous in θ on the sequence where it is defined. We have $E(\log f_m) = -1 + 1/m + o(1/m)$ as $m \to \infty$ and, taking a subsequence, we can assume that the convergence of these integrals is strictly monotone. Then Proposition 4.1.2 applies with $C = 1$.

In another example, if P_θ is uniform on $[\theta, 1]$, where $0 \leq \theta < 1$, we will have consistency if the true $\theta_0 = 0$, for any prior π with $\pi(U) > 0$ for every neighborhood U of 0, even though $a_m = +\infty$ for all m.

The non-consistency at one point θ_0 in Proposition 4.1.2 and the example after it result from (a) peculiar behavior of the likelihood function as $\theta \to \theta_0$, so that although $f(\cdot, x)$ is continuous, P_{θ_m} moves further away from P_{θ_0} in terms of Kullback-Leibler divergence as $m \to \infty$, and (b) very fast decrease of the prior probabilities of neighborhoods of θ_0 as they shrink to θ_0. It may not be surprising, then, that such behavior is exceptional and can only happen on a set of prior probability 0, under quite general conditions, as follows.

Suppose given a parameter space Θ with a σ-algebra $\mathcal{B}$ of subsets and a prior probability distribution π on $\mathcal{B}$. Let $(X, \mathcal{A})$ be a sample space and let X^∞ be the set of all sequences $\{X_n\}_{n=1}^\infty$ with $X_n \in X$ for all n. On X^∞ we have the product σ-algebra, the smallest σ-algebra making each X_n measurable. Suppose that for each $\theta \in \Theta$, a probability measure $\Pr_\theta$ is given on X^∞ and that the family $\Pr_\theta$, $\theta \in \Theta$, is measurable. We get by

Proposition 1.3.5 with X^∞ in place of X a joint distribution Pr on $\Theta \times X^\infty$ where θ has marginal distribution π and, for each $\theta \in \Theta$, $\{X_n\}_{n\geq 1}$ have conditional distribution $\Pr_\theta$.

Posterior probabilities can be defined for families that are not necessarily dominated by a σ-finite measure, as follows. Let Pr be a probability on the product σ-algebra in $\Theta \times X^\infty$. Let $X^{(n)} := (X_1, ..., X_n)$. A function $(A, X^{(n)}) \mapsto \Pr(A|X^{(n)})$ from $\mathcal{B} \times X^n$ into $[0,1]$ is a *regular conditional probability* for Pr of A given $X^{(n)}$ if, for each $A \in \mathcal{B}$, $\Pr(A, \cdot)$ equals the conditional probability of A given $X^{(n)}$, and for each $X^{(n)} \in X^n$, $\Pr(\cdot, X^{(n)})$ is a countably additive probability measure on $\mathcal{B}$. Then a general definition of the posterior probability $\pi_{x,n}$ on Θ is that it is a regular conditional probability $\Pr(\cdot, X^{(n)})$ if one exists, for Pr defined as above via Proposition 1.3.5. For dominated families $\{P_\theta,\ \theta \in \Theta\}$, posteriors as defined via likelihood functions were shown in Theorem 1.3.7 to exist almost surely, and the two definitions of posteriors will agree.

Let's say that the family $\Pr_\theta$, $\theta \in \Theta$, is *empirically identifiable* if for some measurable function T from X^∞ into Θ, for each $\theta \in \Theta$, $T(x) = \theta$ for $\Pr_\theta$-almost all $x \in X^\infty$. This occurs, for example, if there are estimators $T_n(X_1, \ldots, X_n)$ converging to θ $\Pr_\theta$-a.s. as $n \to \infty$ for each θ.

4.1.4 Theorem (Doob). Let $\Pr_\theta$, $\theta \in \Theta$, be a measurable, empirically identifiable family. Suppose that Θ is a Borel subset of a complete separable metric space with Borel σ-algebra. Let π be a prior probability on Θ with $\pi(U) > 0$ for every non-empty open set U. Then the posteriors $\pi_{x,n}$ exist almost surely and are consistent for π-almost all θ.

Proof. There is a Borel isomorphism of Θ onto a complete separable metric space (RAP, Theorem 13.1.1). Then the posteriors exist in the sense of regular conditional probabilities of θ given $X_1, \ldots, X_n$ by RAP, Theorem 10.2.2.

Let U be a non-empty open subset of Θ (for the original topology, not another metric obtained via Borel isomorphism). Then $1_{\theta \in U}$ is an integrable function. Its conditional expectation

$$E(1_{\theta\in U}|X_1, \ldots, X_n) \ = \ \Pr(\theta \in U|X_1, \ldots, X_n) \ = \ \pi_{x,n}(U).$$

Let $\mathcal{F}_n$ be the smallest σ-algebra with respect to which $X_1, \ldots, X_n$ are measurable. The conditional expectations of a fixed integrable function with respect to an increasing sequence of σ-algebras $\mathcal{F}_n$ form a martingale (RAP, Sec. 10.3), which converges almost surely, in this case to $1_{\theta\in U}$ (RAP, Theorem 10.5.1), since by empirical identifiability, this function is measurable with respect to the σ-algebra generated by the union of the $\mathcal{F}_n$.

Since the topology of Θ has a countable base (RAP, Proposition 2.1.4), let $\{U_k\}_{k=1}^\infty$ be such a base. Let the convergence of the martingale for $U = U_k$ hold for $\Pr_\theta$-almost all x for all $\theta \notin A_k$ where $\pi(A_k) = 0$. Let $A := \bigcup_{k=1}^\infty A_k$. Then $\pi(A) = 0$. Let $\theta \notin A$. Then θ has a neighborhood-base consisting of a subsequence $\{U_{k(j)}\}_{j\geq 1}$. By convergence of the martingales we have a.s. $\Pr_\theta$, $\pi_{x,n}(U_{k(j)}) \to 1$ as $n \to \infty$ for all j, so the posteriors are consistent at θ. This completes the proof. □

Thus, if a prior π has an atom with $\pi(\phi) > 0$ for some singleton $\{\phi\}$, $\phi \in \Theta$, then the posteriors will be consistent for such a ϕ under very general conditions. The above proof can be applied without the assumption that Θ is a separable metric space, and even if ϕ is an isolated point, because the posterior probability of $\{\phi\}$ will converge to 1 a.s. $\Pr_\phi$.

NOTES

At this writing I do not have a reference for Theorem 4.1.1 but it is presumably known. Schwartz (1965) gave sufficient conditions for consistency of posteriors at particular θ_0's. Freedman (1963) and Schwartz (1965) gave examples of non-consistent posteriors. Proposition 4.1.2 and the example after it are related to their examples. Theorem 4.1.4 is attributed to Doob (1949). I learned it from Le Cam (1986), p. 616, Prop. 2 and Corollary.

REFERENCES

*Doob, J. L. (1949). Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, Colloques Internat. CNRS, Paris.

Freedman, David A. (1963). On the asymptotic behavior of Bayes estimates in the discrete case. *Ann. Math. Statist.* **34**, 1386-1403.

Le Cam, Lucien (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York.

Schwartz, Lorraine (1965). On Bayes procedures. *Z. Wahrscheinlichkeitsth. verw. Geb.* **4**, 10-26

* I learned of this reference from a secondary source and have not seen it in the original.