2.6 Bayes estimation. The definition of Bayes estimator is a special case of the general definition of Bayes decision rule given in Sec. 1.3. Given a family $\{P_\theta,\ \theta \in \Theta\}$ of laws, where $(\Theta, \mathcal{T})$ is a measurable space, a loss function $L(\theta, y)$, the risk for an estimator U at θ defined by $r(\theta, U) := E_\theta L(\theta, U)$, and a prior π defined on $(\Theta, \mathcal{T})$, an estimator T is *Bayes* for π iff the Bayes risk $r(\pi, U) := \int r(\theta, U) d\pi(\theta)$ has its minimum for all statistics U when $U = T$. Recall that by Theorem 1.3.8, if a decision problem for a measurable family and a given prior has a decision rule with finite risk and some decision rule $a(\cdot)$ minimizes the posterior risk for almost all x, then it is Bayes. Recall also that if a family $\{P_\theta,\ \theta \in \Theta\}$ is dominated by a σ-finite measure v, we can choose v equivalent to the family by Lemma 2.1.6. For squared-error loss, Bayes estimates are just expectations for the posterior:

2.6.1 Theorem. Let $\{P_\theta,\ \theta \in \Theta\}$ be a measurable family equivalent to a σ-finite measure v. Let π be a prior on Θ and g a measurable function from Θ into some $\mathbb{R}^d$. Then for squared-error loss, there exists a Bayes estimator for $g(\theta)$ if and only if there exists an estimator U for $g(\theta)$ with finite risk,

$$r(\pi, U) \ = \ \int \int |U(x) - g(\theta)|^2 dP_\theta(x) d\pi(\theta) \ < \ \infty.$$

Then a Bayes estimator is given by $T(x) := \int g(\theta) d\pi_x(\theta)$ where the integral with respect to the posterior π_x exists and is finite for v-almost all x. T is the unique Bayes estimator up to equality v-almost everywhere. Thus T is an admissible estimator of g.

Proof. Since $|\cdot|^2$ is the sum of squares of coordinates, we can assume $d = 1$. By Propositions 1.3.5 and 1.3.13, the posterior distributions π_x have the properties of regular conditional probabilities of θ given x as defined in RAP, Section 10.2.

"Only if" holds since by definition, a Bayes estimator has finite risk. To prove "if," let U have finite risk, $r(\pi, U) < \infty$. Let $dQ(\theta, x) := dP_\theta(x) d\pi(\theta)$ be the usual joint distribution of θ and x. Then the function $(\theta, x) \mapsto U(x) - g(\theta)$ is in $\mathcal{L}^2(Q)$, even though possibly neither $x \mapsto U(x)$ nor $\theta \mapsto g(\theta)$ is. Thus $U(x) - g(\theta) \in \mathcal{L}^1(Q)$, and we have the conditional expectation (by RAP, Theorem 10.2.5)

$$E(U(x) - g(\theta)|x) \ = \ \int U(x) - g(\theta) d\pi_x(\theta) \ = \ U(x) - \int g(\theta) d\pi_x(\theta)$$

for v-almost all x, since $U(x)$ doesn't depend on θ. Thus $T(x)$ is well-defined for v-almost all x. Now $x \mapsto U(x) - T(x)$ is the orthogonal projection in $\mathcal{L}^2(Q)$ of $U(x) - g(\theta)$ into the space H of square-integrable functions of x for Q (RAP, Theorem 10.2.9), which is unique up to a.s. equality (RAP, Theorem 5.3.8). Thus $\int (U(x) - g(\theta) - f(x))^2 \, dQ(\theta, x)$ is minimized over all square-integrable functions f of x when and only when $f(x) = U(x) - T(x)$ for v-almost all x. For any other estimator $V(x)$ of $g(\theta)$ with finite risk, $U - V \in H$. Thus $\int (V(x) - g(\theta))^2 dQ(\theta, x)$ is minimized among all estimators $V(x)$ of $g(\theta)$ when $V = T$, in other words, T is a Bayes estimator of $g(\theta)$, unique up to v-almost everywhere equality.

A Bayes estimator or decision rule, unique up to almost sure equality for all P_θ, is always admissible by Theorem 1.2.5. □

Example. Let $\Theta = \mathbb{R}$ and consider the normal location family with $P_\theta = N(\theta, 1),\ \theta \in \mathbb{R}$. Let π be the Cauchy prior, $d\pi(\theta) = d\theta/[\pi(1+\theta^2)],\ \theta \in \mathbb{R}$. Let $g(\theta) := \theta$ and $U(x) := x$. Then for squared-error loss, $r(\theta, U) = 1$ for all θ, so $r(\pi, U) = 1$ for any prior, specifically the Cauchy prior. Since this risk is finite, Theorem 2.6.1 tells us that a Bayes estimator exists. Note however that g is not in $\mathcal{L}^2$ or even in $\mathcal{L}^1$ of the prior. We know that conditional expectations given x, when they exist, are given by integrals with respect to posterior distributions π_x. In this case, however, where the expectation is undefined, so are conditional expectations, but the integrals with respect to π_x are defined and give the Bayes estimator. It is easily seen that multiplication by the normal likelihood function makes the integrals finite.

Given a parameter space Θ with a metric d defined on it, where (Θ, d) is separable and $\mathcal{T}$ is the Borel σ-algebra, and given a loss function, a decision rule T will be called *Bayes admissible* if there exists some prior π, with $\pi(U) > 0$ for every non-empty open set $U \subset \Theta$, such that T is Bayes for π.

Consider again the estimator of p^2 from two binomial trials given near the end of Sec. 2.5, which is unbiased and admissible for many loss functions, but gives the estimate 0 for p^2 when one success is observed in the two trials. This estimator is not Bayes admissible. Moreover, it is not Bayes for any prior π such that $\pi((0,1)) > 0$ for the open interval $(0,1)$.

Other, familiar estimators are not Bayes admissible. For example, to estimate p given that there were k successes in n independent trials with probability p of success on each, the classical estimator of p is k/n, which is unbiased, sufficient and LS. It estimates that $p = 0$ when $k = 0$ and $p = 1$ when $k = n$. These estimates make it not Bayes admissible. A Bayes estimator for any prior giving positive probability to the open interval $(0,1)$, for squared-error loss, must make a strictly positive estimate even when $k = 0$ and an estimate < 1 even when $k = n$. For the prior uniformly distributed over $[0,1]$, the Bayes estimator is $(k+1)/(n+2)$ which, of course, is not unbiased (e.g. for $p = 0, 1$), but on the whole the possible bias of Bayes or Bayes admissible estimators seems a lesser fault than those of unbiased estimators in examples such as we have just recalled.

Surprisingly, Bayes admissibility for squared-error loss is quite incompatible with unbiasedness:

2.6.2 Theorem. For any dominated, measurable family $\{P_\theta,\ \theta \in \Theta\}$, prior π on Θ and measurable real-valued function g on Θ, an unbiased estimator T of g is Bayes for π and squared-error loss if and only if it has risk $r(\pi, T) = 0$, so that $T(x) = g(\theta)$ P_θ-almost surely for π-almost all θ.

Remarks. The condition of 0 Bayes risk is extremely restrictive: note that whenever $g(\theta) \neq g(\phi)$, the laws P_θ and P_ϕ must be singular (the opposite extreme from an equivalent family). So, for any equivalent family, a 1-1 function g has no unbiased, Bayes estimator for any prior which is not trivial (concentrated at one point).

Proof. "If" is clear. To prove "only if," by definition of Bayes decision rule (Section 1.2), T must have finite risk. Thus by Theorem 2.6.1, $T(x) = \int g(\theta) d\pi_x(\theta)$. Let p be the "predictive" distribution for x, in other words its marginal distribution under Q, as

defined just before (1.3.6). Then $dQ(\theta, x) = d\pi_x(\theta)dp(x)$. We have

$$r(\pi, T) = \int\int T(x)^2 - 2T(x)g(\theta) + g(\theta)^2 \, d\pi_x(\theta)dp(x)$$

$$= \int \left[T(x)^2 - 2T(x)^2 + \int g(\theta)^2 d\pi_x(\theta) \right] dp(x).$$

Since $r(\pi, T) < \infty$, we have $\int g(\theta)^2 d\pi_x(\theta) < \infty$ for p-almost all x, and

$$r(\pi, T) = \int g(\theta)^2 - T(x)^2 dQ(\theta, x).$$

Similarly, by unbiasedness,

$$r(\pi, T) = \int\int T(x)^2 - 2T(x)g(\theta) + g(\theta)^2 dP_\theta(x)d\pi(\theta)$$

$$= \int \left[\int T(x)^2 dP_\theta(x) - 2g(\theta)^2 + g(\theta)^2 \right] d\pi(\theta),$$

so $r(\pi, T) < \infty$ implies $\int T(x)^2 dP_\theta(x) < \infty$ for π-almost all θ, and

$$r(\pi, T) = \int T(x)^2 - g(\theta)^2 dQ(\theta, x) = -r(\pi, T),$$

so $r(\pi, T) = 0$, finishing the proof. □

PROBLEMS

1. Let $\Gamma(a) := \int_0^\infty x^{a-1}e^{-x}dx$ for $a > 0$, and $f_{\lambda,a}(x) := \lambda^a x^{a-1}e^{-\lambda x}/\Gamma(a)$ for $x > 0$, 0 for $x \le 0$, where $a > 0$ and $\lambda > 0$. Then $f_{\lambda,a}$ is a gamma probability density with *scale parameter* λ and *shape parameter* a. Let the parameter λ of a Poisson distribution with $P(X = k) = \lambda^k e^{-\lambda}/k!$ for $k = 0, 1, ...$, have a prior distribution $e^{-\lambda}d\lambda$ for $\lambda \ge 0$ (standard exponential distribution). Let k be the observed value of the Poisson random variable. What is the posterior distribution of λ?
2. For the binomial distribution with n trials and probability p of success, the variance is $np(1-p)$. More specifically, suppose we observe n i.i.d. Bernoulli variables $X_1, ..., X_n$ where $P(X_j = 1) = p = 1 - P(X_j = 0)$ for each j. Let $k := \sum_{j=1}^n X_j$ be the number of successes.
   (a) Show that the usual unbiased estimator s^2 of variance (for general distributions) equals $k - k^2/n$ in this case.
   (b) Show that s^2 is admissible in this case for any convex loss function. *Hint*: see the proof of Proposition 2.5.18.
   (c) Show that s^2 is not Bayes admissible for squared-error loss in this case. *Hint:* It is unbiased.

3. "Unbiased estimation on a circle." Let the sample space be the unit circle $S^1 := \{(x,y) : x^2+y^2=1\}$. Given n points on the circle all within half the circle (some arc of π radians), assign angular coordinates in an interval of length π radians. Any real number θ, viewed as an angle, defines a point $s_\theta := (\cos\theta, \sin\theta) \in S^1$. For example, if the observations were $(1,-1)/2^{1/2}$ and $(1,1)/2^{1/2}$ a suitable interval would be $-\pi/2 < \phi < \pi/2$ but not $0 < \phi < \pi$. Average these coordinates to get a $\overline{\theta}$ and then an estimator $T := s_{\overline{\theta}} \in S^1$.

   (a) Show that T is a well-defined statistic.

   (b) For any random variable (X,Y) taking values within some half-circle, with distribution depending on some parameter θ, define its circular expectation $E_{o,\theta}(X,Y)$ as a point in S^1 by a choice of angular coordinate as above. Show that $E_{o,\theta}(X,Y)$ is well-defined. Let $\{P_\theta : 0 \le \theta < 2\pi\}$ be the family of laws which are uniform on arcs of length π, given by $[\theta - \pi/2, \theta + \pi/2]$.

   (c) Show that T is unbiased for circular expectation for the given family $\{P_\theta : 0 \le \theta < 2\pi\}$ in the sense that $E_{o,\theta}T \in S^1$ is equal for any θ to s_θ.

   (d) Find the risk of T for each θ where the loss function is $\min_{k\in\mathbb{Z}}(\overline{\theta} - \theta - 2k\pi)^2$ and the minimum is over all integers $k \in \mathbb{Z}$. *Hint*: $|\overline{\theta} - \theta - 2k\pi| \le \pi/2$ for a unique $k \in \mathbb{Z}$. Reduce to a question about real-valued uniform random variables.

4. In the same situation as the previous problem, for n i.i.d. observations and a given choice of coordinates, take the order statistics $\theta(1) < \theta(2) < \cdots < \theta(n)$. Let $U := (\theta(1) + \theta(n))/2$.

   (a) Show that $(s_{\theta(1)}, s_{\theta(2)})$ form a sufficient statistic in $S^1 \times S^1$.

   (b) Show that s_U is well-defined and is an unbiased estimator of s_θ in the same sense as T in Problem 3.

   (c) Find the risk of U for each θ as in Problem 3 and show it is smaller than that of T.

5. Let $\{P_\theta : -\infty < \theta < \infty\}$ be a location family in $\mathbb{R}$ where P_θ has a density $f(\theta, x) := g(x-\theta)$ with respect to Lebesgue measure dx for a fixed probability density g. An estimator T for θ is called *equivariant* if for any $h \in \mathbb{R}$, the distribution of $T + h$ for P_θ is the same as the distribution of T for $P_{\theta+h}$. Show that an equivariant estimator of θ can't be Bayes for squared-error loss for any prior. *Hint*: an equivariant estimator minus a constant would be an unbiased estimator, and being Bayes implies that the estimator itself is unbiased. Apply Theorem 2.6.2. Then $\{x : g(x) > 0\}$ has Lebesgue measure > 0 and so cannot be disjoint from all its translates (RAP, Prop. 3.4.3).

6. An estimator $V = V(u_1, \ldots, u_n)$ taking $(S^1)^n$ into S^1 for the unit circle S^1 is called *equivariant* if for any rotation R of S^1, $V(Ru_1, \ldots, Ru_n) \equiv R(V(u_1, ..., u_n))$. Show that both estimators T and s_U in Problems 3 and 4 are equivariant.

7. Prove or disprove: the estimator s_U in Problem 3 is Bayes for the uniform prior on the circle and loss function as in Problem 3(b).

8. Suppose the probability p of success in n independent trials has a prior which is a beta density $p^{a-1}(1-p)^{b-1}/B(a,b)$ for $0 < p < 1$ with respect to Lebesgue measure. Here $a > 0$, $b > 0$, and $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$. Recall that $B(a,b) \equiv \Gamma(a)\Gamma(b)/\Gamma(a+b)$

where $\Gamma(a) := \int_0^\infty x^{a-1}e^{-x}dx$ for $a > 0$ and $\Gamma(m) = (m-1)!$ for $m = 1, 2, \ldots$. If k successes are observed in the n trials,

(a) What is the posterior distribution of p?

(b) What is the Bayes estimator of p, for squared-error loss?

NOTES

Theorems 2.6.1 and 2.6.2 are stated in Lehmann (1991), Corollary 4.1.1 p. 239 and Theorem 4.1.2 pp. 244-245, but for the former, Lehmann writes the Bayes estimator as $T(x) = E(g(\theta)|x)$, implicitly assuming that $\int |g(\theta)|d\pi(\theta) < \infty$. Lehmann's proof of the latter theorem uses the assumption that $g(\theta) \in \mathcal{L}^2(\pi)$. The Example given after Theorem 2.6.1 shows that these assumptions need not hold. Thus, other proofs have been given for Theorems 2.6.1 and 2.6.2 without any moment assumptions.

Lehmann apparently doesn't give any earlier references for these facts, although at least Theorem 2.6.1 for $g \in \mathcal{L}^2$ was presumably known well before 1983.

Bickel and Doksum (1977), Theorem 1.6.1 and (10.3.1), state that the Bayes estimator of $g(\theta)$, if it exists, must be $E(g(\theta)|x)$. Again, this is correct only when the expectation and thus the conditional expectation are defined.

Berger (1985, Sec. 4.4.2 p. 161) correctly states that the Bayes estimator for squared-error loss is the expectation for the posterior distribution, in the special case $g(\theta) \equiv \theta$, under the assumption that each of three integrals for the posterior distribution is finite (as they will be, almost surely, under the assumption of Theorem 2.6.1).

REFERENCES

Berger, James O. (1980). *Statistical Decision Theory: Foundations, Concepts and Methods.* Springer, New York.

Berger, James O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Second ed. of Berger (1980). Springer, New York.

Bickel, P. J., and Doksum, K. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* Holden-Day, San Francisco.

Lehmann, Erich (1991). *Theory of Point Estimation.* Wadsworth, Pacific Grove, CA; 1st ed. Wiley, New York, 1983.