Statistics 582, Problem Set 8 Solutions

Wellner; 3/5/98

1. Consider testing the simple hypothesis $H_0 : X \sim P_0$ versus the simple alternative $H_1 : X \sim P_1$. Let ϕ be a test of H_0 versus H_1, and suppose a prior distribution on $\{P_0, P_1\}$ is given by $(\lambda, 1-\lambda)$ with $\lambda \in (0,1)$.
A. For losses given by l_0 and l_1 as in Corollary 5.5.4, page 12, chapter 5, find the Bayes rule for this problem. Compute the ordinary risks and the Bayes risk for the Bayes rule.
B. When $l_0 = l_1 = 1$ and $\lambda = 1/2$, express the Bayes risk you found in A in terms of the total variation distance between P_0 and P_1. Explain why this relationship makes sense intuitively.
C. Does the relationship you found in B continue to hold for other values of the losses l_0 and l_1 and prior λ?

Solutions: A. By Corollary 5.5.4 of the notes, the Bayes rule for testing with the loss function $L(\theta, a_i) = l_i 1_{\Theta_i^c}(\theta)$ is given by

$$d_\Lambda(1|x) \equiv d_\Lambda(x) = \begin{cases} 1 & \text{if } P(\theta \in \Theta_1|X=x) > (l_1/l_0)P(\theta \in \Theta_1|X=x) \\ \gamma(x) & \text{if } P(\theta \in \Theta_1|X=x) = (l_1/l_0)P(\theta \in \Theta_1|X=x) \\ 0 & \text{if } P(\theta \in \Theta_1|X=x) < (l_1/l_0)P(\theta \in \Theta_1|X=x) \end{cases} .$$

When there are just two possible states of nature $P_{\theta_0} = P_0$ and $P_{\theta_1} = P_1$, with prior probabilities λ and $1-\lambda$ respectively, the posterior probabilites of $\Theta_i = \{\theta_i\}$ for $i = 0, 1$ become

$$P(\theta \in \Theta_i|X=x) = \frac{\lambda^{1-i}(1-\lambda)^i p_i(x)}{\lambda p_0(x) + (1-\lambda)p_1(x)}$$

where $p_i(x) = (dP_i/d\mu)(x)$ for some dominating measure μ ($\mu = P_0 + P_1$ will always work). Thus the Bayes rule in this case becomes

$$d_\Lambda(x) = \begin{cases} 1 & \text{if } (1-\lambda)p_1(x) > (l_1/l_0)\lambda p_0(x) \\ \gamma(x) & \text{if } (1-\lambda)p_1(x) = (l_1/l_0)\lambda p_0(x) \\ 0 & \text{if } (1-\lambda)p_1(x) < (l_1/l_0)\lambda p_0(x) \end{cases} .$$

Let $c \equiv (l_1\lambda/l_0(1-\lambda))$. Then the Bayes risk and ordinary risks of this rule are given by:

$$\mathcal{R}(\Lambda, d_\Lambda) \;=\; l_0 P(\theta \in \Theta_1) + \int_{p_1(x) > c p_0(x)} \{l_1\lambda p_0(x) - l_0(1-\lambda)p_1(x)\} d\mu(x)$$

$$= l_0(1-\lambda) + \int_{p_1(x)>cp_0(x)} \{l_1\lambda p_0(x) - l_0(1-\lambda)p_1(x)\}d\mu(x)\,,$$

$$R(0,d_\Lambda) = l_1 E_0(d_\Lambda(X)) = l_1 E_0\{1_{[p_1(X)>cp_0(X)]} + \gamma(X)1_{[p_1(X)=cp_0(X)]}\}\,,$$

$$R(1,d_\Lambda) = l_0 E_1(1-d_\Lambda(X)) = l_0(1 - E_0\{1_{[p_1(X)>cp_0(X)]} + \gamma(X)1_{[p_1(X)=cp_0(X)]}\})\,,$$

B. When $l_0 = l_1 = 1$, and $\lambda = 1/2$, then $c = 1$ and

$$\begin{aligned}
\mathcal{R}(\Lambda, d_\Lambda) &= \frac{1}{2} + \frac{1}{2}\int_{p_1>p_0} p_0 d\mu - \frac{1}{2}\int_{p_1>p_0} p_1 d\mu \\
&= \frac{1}{2}\{1 + \int_{p_1>p_0} p_0 d\mu + \int_{p_1\le p_0} p_1 d\mu - \int_{p_1>p_0} p_1 d\mu - \int_{p_1\le p_0} p_1 d\mu\} \\
&= \frac{1}{2}\{1 + \int_{p_1>p_0} p_0 d\mu + \int_{p_1\le p_0} p_1 d\mu - 1\} \\
&= \frac{1}{2}\int_{p_1>p_0} p_0 \wedge p_1 d\mu \\
&= \frac{1}{2}(1 - d_{TV}(P_0, P_1))
\end{aligned}$$

by problem 7.3.B. Thus when $d_{TV}(P_0, P_1)$ is small (P_0 and P_1 are close together) the Bayes risk is close to $1/2$, and when $d_{TV}(P_0, P_1)$ is close to 1 (so P_0 and P_1 are far apart), the Bayes risk is close to zero (and equals zero when $d_{TV}(P_0, P_1) = 1$). This makes good sense intuitively since it should be harder to distinguish P_0 from P_1 when they are close together than when they are far apart (and $d_{TV}(P_0, P_1)$ is close to 1).

C. Note that when $c = (l_1\lambda/l_0(1-\lambda)) = 1$, then $l_1\lambda = l_0(1-\lambda)$ and the Bayes risk is

$$\begin{aligned}
(1-\lambda)l_0\{1 + \int_{p_1>p_0} p_0 d\mu - \int_{p_1>p_0} p_1 d\mu\} &= (1-\lambda)l_0 \int p_0 \wedge p_1 d\mu \\
&= (1-\lambda)l_0(1 - d_{TV}(P_0, P_1))\,.
\end{aligned}$$

2. Suppose that $X \sim N(\theta, \sigma^2)$, and it is "known" that $\theta \in [-\tau, \tau]$.

A. Consider the class of *affine estimators* of θ given by $d(X) = cX + d$ for $c, d \in R$. Show that the minimax affine estimator of $\theta \in [-\tau, \tau]$ is given by $d_0 = c_0 X$ where $c_0 = \tau^2/(\sigma^2 + \tau^2)$ and that the minimax risk is

$$\inf_{c,d} \sup_{\theta\in[-\tau,\tau]} E_\theta(\theta - cX - d)^2 = \sigma^2\tau^2/(\sigma^2 + \tau^2)$$

B. Suppose that $\lambda(\theta)$ is a prior distribution for θ with all its mass concentrated on $[-\tau, \tau]$. Find the posterior distribution for θ for such a prior. Where is it concentrated? What can you say about the Bayes rule with respect to this prior

for squared error loss?
C. How is the minimax risk over all estimators related to the minimax risk of affine estimators which you computed in A? How is it related to the Bayes risk?
D. Suppose that a sample of i.i.d. X_i's is available so that $\overline{X}_n \sim N(\theta, \sigma^2/n)$ (i.e. we can replace σ^2 by σ^2/n) in C, and suppose that the true $\theta_0 > \tau$. Describe the behavior of the posterior distributions and Bayes rule in C as $n \to \infty$.

Solution: A. For each fixed $\theta \in R$ and $c, d \in R$ the risk of the rule $d(X) = cX+d$ is

$$\begin{aligned} R(\theta, d) &= E_\theta(\theta - cX - d)^2 \\ &= Var_\theta(cX + d) + (E_\theta(cX + d) - \theta)^2 \\ &= c^2\sigma^2 + (c\theta + d - \theta)^2 \\ &= c^2\sigma^2 + ((c-1)\theta + d)^2 \\ &= c^2\sigma^2 + (1-c)^2\,(\theta - \theta_0)^2 \end{aligned}$$

where $\theta_0 \equiv -d/(c-1) = d/(1-c)$. This function of θ has a minimum of $c^2\sigma^2$ at θ_0 and grows quadratically in both directions from θ_0. Thus it follows that

$$\sup_{\theta \in [-\tau, \tau]} R(\theta, d) = \begin{cases} c^2\sigma^2 + (1-c)^2\,(\tau - \theta_0)^2 & \text{if } \theta_0 < 0 \\ c^2\sigma^2 + (1-c)^2\tau^2 & \text{if } \theta_0 = 0 \\ c^2\sigma^2 + (1-c)^2\,(-\tau - \theta_0)^2 & \text{if } \theta_0 > 0 \end{cases} .$$

It is easily seen from the pictures of the risks that this function of c, d is minimized by $d = 0$, and then

$$\inf_d \sup_{\theta \in [-\tau, \tau]} R(\theta, d) = c^2\sigma^2 + (1-c)^2\tau^2$$

is minimized over c by easy calculations: the first derivative w.r.t. c is $2c\sigma^2 + 2(c-1)\tau^2$, and this is 0 if $c = c_0 \equiv \tau^2/(\sigma^2+\tau^2)$; the second derivative is $2(\sigma^2+\tau^2) > 0$, so c_0 yields a minimum. Hence we find that the minimum risk of affine estimators of θ is given by

$$\inf_{c,d} \sup_{\theta \in [-\tau, \tau]} E_\theta(\theta - cX - d)^2 = c_0^2\sigma^2 + (1-c_0)^2\tau^2 = \sigma^2\tau^2/(\sigma^2 + \tau^2)\,.$$

B. If Λ is a prior distribution with density $\lambda(\theta)$ on $[-\tau, \tau]$ then the posterior distribution of θ is given by

$$\lambda(\theta|x) = \frac{\sigma^{-1}\phi(\frac{x-\theta}{\sigma})\lambda(\theta)1_{[-\tau,\tau]}(\theta)}{\int_{-\tau}^{\tau} \sigma^{-1}\phi(\frac{x-s}{\sigma})\lambda(s)1_{[-\tau,\tau]}(s)\,ds}\,.$$

It is clear that the posterior also puts all its mass on the interval $[-\tau, \tau]$, and therefore the Bayes rule with respect to squared error, $E(\theta|X)$, is in the interval

$[-\tau, \tau]$.
C. Since the class of all estimators is larger than the class of affine estimators, we have (with "N" standing for "nonlinear" and "A" for "affine")

$$\rho_N(\tau,\sigma) \equiv \inf_{d\in\mathcal{D}} \sup_{\theta\in[-\tau,\tau]} R(\theta,d) \le \inf_{d\in\mathcal{D}_{affine}} \sup_{\theta\in[-\tau,\tau]} R(\theta,d) \equiv \rho_A(\tau,\sigma) = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\,.$$

In fact it is known that

$$\rho_A(\tau,\sigma) \le \frac{5}{4}\rho_N(\tau,\sigma)$$

so that

$$\frac{4}{5}\rho_A(\tau,\sigma) \le \rho_N(\tau,\sigma) \le \rho_A(\tau,\sigma)\,;$$

see e.g. Donoho (1994) (*Ann. Statist.* **22**, 238 - 270) and the references therein. The relationship with the Bayes risk for any prior Λ is as follows:

$$\begin{aligned}
\mathcal{R}(\Lambda,d_\Lambda) &= \inf_{d\in\mathcal{D}} \mathcal{R}(\Lambda,d) \\
&\le \inf_{d\in\mathcal{D}_{affine}} \mathcal{R}(\Lambda,d) \\
&= \inf_{d\in\mathcal{D}_{affine}} \int_{-\tau}^{\tau} R(\theta,d)d\Lambda(\theta) \\
&\le \inf_{d\in\mathcal{D}_{affine}} \sup_{\theta\in[-\tau,\tau]} R(\theta,d) = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\,.
\end{aligned}$$

Of course this also follows from the preceding since the minimax risk $\rho_N(\sigma,\tau)$ is, by Theorem 5.6.2, equal to the Bayes risk for the least favorable prior distribution Λ:

$$\mathcal{R}(\Lambda,d_\Lambda) \le \sup_{\Lambda} \mathcal{R}(\Lambda,d_\Lambda) \equiv r_\Lambda = \rho_N(\sigma,\tau) \le \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\,.$$

As $\tau\to\infty$ Bickel (1981) has shown that the "approximate least favorable" densities are of the form $\lambda_\tau(\theta) = \tau^{-1}\cos^2((\pi/2)\theta/\tau)1_{[-\tau,\tau]}(\theta)$.
D. When we have $\overline{X}_n \sim N(\theta,\sigma^2/n)$, then the posterior is

$$\begin{aligned}
\lambda(\theta|\underline{X}) &= \frac{(\sigma/\sqrt{n})^{-1}\phi(\frac{\sqrt{n}(\overline{X}_n-\theta)}{\sigma})\lambda(\theta)1_{[-\tau,\tau]}(\theta)}{\int_{-\tau}^{\tau}(\sigma/\sqrt{n})^{-1}\phi(\frac{\sqrt{n}(\overline{X}_n-s)}{\sigma})\lambda(s)ds} \\
&= \frac{(\sigma/\sqrt{n})^{-1}\phi(\frac{\sqrt{n}(\overline{X}_n-\theta_0)+\sqrt{n}(\theta_0-\theta)}{\sigma})\lambda(\theta)1_{[-\tau,\tau]}(\theta)}{\int_{-\tau}^{\tau}(\sigma/\sqrt{n})^{-1}\phi(\frac{\sqrt{n}(\overline{X}_n-s)}{\sigma})\lambda(s)ds}
\end{aligned}$$

where $\sqrt{n}(\overline{X}_n-\theta_0) = O_p(1)$ while $\sqrt{n}(\theta_0-\theta)\to\infty$ for each fixed $\theta\in[-\tau,\tau]$. Hence $\lambda(\theta|\underline{X}) \to_p 0$ for each fixed $\theta\in[-\tau,\tau]$. Since the normal density in the

numerator is centered at $\overline{X}_n \to_p \theta_0 > \tau$, we see that

$$\Lambda(\theta|\underline{X}_n) = \int_{-\infty}^{\theta} \lambda(s|\underline{X})\, ds \to_p 1_{[\tau,\infty)}(\theta)\,.$$

3. Suppose that $X_1, \ldots, X_n$ are i.i.d. Exponential(θ) (so the X's have density $p_\theta(x) = \theta e^{-\theta x} 1_{(0,\infty)}(x)$. with respect to Lebesgue measure on R, and that $\theta \sim \Gamma(\alpha, \beta)$:

$$\lambda(\theta) = \beta \frac{(\beta\theta)^{\alpha-1}}{\Gamma(\alpha)} \exp(-\beta\theta) 1_{[0,\infty)}(\theta).$$

A. Find the Bayes rule $d_B(\underline{X})$ for estimation of θ with squared error loss $L(\theta, a) = |\theta - a|^2$. Find the Bayes rule $d_{Bw}(\underline{X})$ for estimation of θ with weighted squared error loss $L(\theta, a) = (\theta - a)^2/\theta$. Is the maximum likelihood estimator among either of these families of Bayes estimators?
B. Are the Bayes estimators d_B and d_{Bw} consistent? What are the limit distributions of d_B and d_{Bw}? Compare them with the maximum likelihood estimator.
C. Prove a (conditional) limit theorem for the posterior distributions given $\underline{X}$.
D. Suppose that instead of the Gamma prior distribution, θ has the Pareto(θ_0, α) distribution with density λ given by

$$\lambda(\theta) = (\frac{\alpha}{\theta_0})(\frac{\theta_0}{\theta})^{\alpha+1} 1_{(\theta_0,\infty)}(\theta);$$

here $E(\theta) = \frac{\alpha}{\alpha-1}\theta_0$ where $\alpha > 1$ and $\theta_0 > 0$ are known. What can you say about the Bayes estimator for squared error loss with this prior? For what values of θ_0 is the Bayes rule consistent?

Solution: A. The posterior distribution is Gamma($\alpha + n, \beta + \sum X_i$). Thus the Bayes rule for $L(\theta, a) = (\theta - a)^2$ is

$$d_B(\underline{X}) = \frac{\alpha + n}{\beta + \sum X_i}\,.$$

For $L(\theta, a) = (\theta - a)^2/\theta$, the Bayes rule is

$$d_{Bw}(\underline{X}) = \frac{E(\theta K(\theta)|\underline{X})}{E(K(\theta)|\underline{X})} = \frac{1}{E(1/\theta|\underline{X})} = \frac{\alpha + n - 1}{\beta + \sum X_i}$$

since, for $\theta \sim$ Gamma(α, β) we have

$$E(1/\theta) = \frac{\beta}{\alpha - 1}$$

if $\alpha > 1$. Thus the MLE $1/\overline{X}_n$ is *not* among either of these families of estimators.
B. Both d_B and d_{Bw} are consistent and asymptotically equivalent to the MLE $1/\overline{X}_n$:

$$\begin{aligned}\sqrt{n}\left\{d_B(\underline{X}) - 1/\overline{X}_n\right\} &= \sqrt{n}\left\{\frac{1+n^{-1}\alpha}{\overline{X}_n + n^{-1}\beta} - \frac{1}{\overline{X}_n}\right\} \\ &= n^{-1/2}\frac{\alpha\overline{X}_n - \beta}{\overline{X}_n(\overline{X}_n + n^{-1}\beta)} = O(n^{-1/2})O_p(1) = o_p(1)\,,\end{aligned}$$

and similarly for d_{Bw}. Thus, for $d = d_B$ or $d = d_{Bw}$ we have, since $I(\theta) = \theta^{-2}$,

$$\sqrt{n}(d(\underline{X}) - \theta) = \sqrt{n}(\frac{1}{\overline{X}_n} - \theta) + o_p(1) \to_d N(0, 1/I(\theta)) = N(0, \theta^2)\,.$$

C. Now

$$\begin{aligned}\theta \sim \text{Gamma}(\alpha + n, \beta + \sum X_i) &=_d (\beta + \sum X_i)^{-1}\text{Gamma}(\alpha + n, 1) \\ &=_d (\beta + \sum X_i)^{-1}(Y_0 + \sum_{i=1}^n Y_i)\end{aligned}$$

where $Y_0 \sim \text{Gamma}(\alpha, 1)$, and $Y_i \sim \text{Gamma}(1,1) = \text{Exp}(1)$, $i = 1, \ldots, n$ are all independent. Thus conditionally on the X_i's we have, with $Z \sim N(0,1)$ and with θ_0 the true value of θ,

$$\begin{aligned}\sqrt{n}(\theta - E(\theta|\underline{X})) &=_d \sqrt{n}\frac{Y_0 + \sum_{i=1}^n Y_i - (\alpha + n)}{\beta + \sum_{i=1}^n X_i} \\ &= \sqrt{n}(\overline{Y}_n - 1)\frac{1}{\overline{X}_n + n^{-1}\beta} + \sqrt{n}(Y_0 - \alpha)\frac{1/n}{\overline{X}_n + n^{-1}\beta} \\ &\to_d Z\frac{1}{\theta_0^{-1}} \sim N(0, \theta_0^2)\end{aligned}$$

almost surely with respect to the distribution of $X_1, X_2, \ldots$. Note that the posterior mean $E(\theta|\underline{X})$ can be replaced here by either the MLE $1/\overline{X}_n$ or by $T_n = \theta_0 + (nI(\theta_0))^{-1}\sum_{i=1}^n \dot{l}_\theta(X_i) = 2\theta_0 - \theta_0^2\overline{X}_n$ since

$$\sqrt{n}(E(\theta|\underline{X}) - 1/\overline{X}_n) = o_p(1)$$

and similarly with T_n in place of $1/\overline{X}_n$.
D. When the prior is Pareto(θ_0, α), the posterior density is of the form

$$\begin{aligned}\lambda(\theta|\underline{X}) &= \frac{\theta^n \exp(-\theta\sum X_i)(\alpha\theta_0^{-1})(\theta_0/\theta)^{\alpha+1}1_{(\theta_0,\infty)}(\theta)}{\int_{\theta_0}^\infty s^n \exp(-s\sum X_i)(\alpha\theta_0^{-1})(\theta_0/s)^{\alpha+1}ds} \\ &= \frac{\theta^{n-\alpha-1}\exp(-\theta\sum X_i)1_{(\theta_0,\infty)}(\theta)}{\int_{\theta_0}^\infty s^{n-\alpha-1}\exp(-s\sum X_i)ds}\,,\end{aligned}$$

which is concentrated on (θ_0, ∞). Thus the Bayes rule $d_B(\underline{X}) = E(\theta|\underline{X})$ takes values in (θ_0, ∞) a.s.. Similar to the argument in class in the Bernoulli(θ) example, $Z_n = d_B(\underline{X}) = E(\theta|X_1, \ldots, X_n)$ is a martingale and hence $Z_n = d_B(\underline{X}) \to E(\theta|X_1, X_2, \ldots)$. But $\widehat{\theta} = \overline{X}_n^{-1} \to_{a.s.} \theta$ for each fixed $\theta \in (0, \infty)$, and hence

$$P_\Lambda(\widehat{\theta}_n \to \theta) = \int P_\theta(\widehat{\theta}_n \to \theta) d\Lambda(\theta) = 1\,.$$

Hence $\widehat{\theta}_n \to \theta$ a.s. P_Λ, and this implies that θ is $\mathcal{F}_\infty \equiv \sigma(X_1, X_2, \ldots)$ measurable. Therefore $E(\theta|X_1, X_2, \ldots) = \theta$ a.s. and $d_B(\underline{X}) \to \theta$ a.s. P_Λ. This in turn implies that $d_B(\underline{X}) \to_{a.s.} \theta$ for Λ-a.e. θ. this suggests that d_B might be inconsistent for $\theta \in (0, \theta_0)$, and this is in fact the case since $d_B(\underline{X}) < \theta_0$. When the true $\theta < \theta_0$, it is possible to show that $d_B(\underline{X}) \to_{a.s.} \theta_0 > \theta$ and that the posterior distributions convergen to point mass at θ_0.