Statistics 582, Problem Set 7, Solutions

Wellner; 2/25/98

1. Ferguson, problem 2.11.11, page 96.

Solution: A. Here for the rule $d_0(X) = X$, we compute

$$R(\theta, d_0) = E_\theta \left\{ \frac{(\theta - X)^2}{\theta} \right\} = \frac{Var_\theta(X)}{\theta} = \frac{\theta}{\theta} = 1 \,,$$

so $d_0(X) = X$ is an "equalizer rule": it has constant risk for the loss function $L(\theta, a) = (\theta - a)^2/\theta$.

B. When the (improper) prior is Lebesgue measure, we have $\lambda(\theta) = 1$ for all θ and the posterior is Gamma$(X + 1, 1)$. The Bayes rule for this posterior and the given loss function is

$$d_\Lambda(X) = \frac{E(\theta K(\theta)|X)}{E(K(\theta)|X)} = \frac{\Gamma(X+1)}{\Gamma(X)} = X \,.$$

Thus $d_0(X) = X$ is "generalized Bayes".

C. When the prior is Gamma(α, β), the posterior distribution is Gamma$(\alpha + X, \beta + 1)$, so that the Bayes estimator for the given loss function is

$$\begin{aligned} d_\Lambda(X) &= \frac{E(\theta K(\theta)|X)}{E(K(\theta)|X)} = \frac{1}{E(1/\theta|X)} \\ &= \frac{\alpha + X - 1}{\beta + 1} \,. \end{aligned}$$

D. Here I will choose $\alpha = 1$ for simplicity. Then the Bayes risk of the Bayes rules found in C is easily found to be

$$\begin{aligned} \mathcal{R}(\Lambda, d_\Lambda) &= E \left\{ \frac{(\theta - d_\Lambda(X))^2}{\theta} \right\} \\ &= E \left\{ \frac{(\theta(\beta+1) - X)^2}{(\beta+1)^2 \theta} \right\} \\ &= E \left\{ \frac{[(X - \theta) - \beta\theta]^2}{\theta(\beta+1)^2} \right\} \\ &= \frac{1}{(\beta+1)^2} + \frac{\beta}{(\beta+1)^2} = \frac{1}{\beta+1} \\ &\to 1 \qquad \text{as} \quad \beta \to 0 \\ &= \sup_\theta R(\theta, d_0) \,. \end{aligned}$$

Thus we conclude by Theorem 5.6.7 that d_0 is minimax.

2. Suppose that $X_n \equiv X \sim \text{Multinomial}_k(n, \underline{\theta})$.
A. Suppose that the prior distribution on θ is given by a Dirichlet distribution, Dirichlet($\underline{\alpha}$):

$$\lambda(\underline{\theta}) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_k)}{\prod_{j=1}^k \Gamma(\alpha_j)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} 1_{[\underline{\theta}: \sum \theta_i = 1]} .$$

Verify the computation of the Bayes estimator for squared error loss given in example 4.3.4
B. What is the posterior distribution for θ? Find the mode of the posterior distribution (along the lines of our computations of the MLE of the multinomial) and compare it with the MLE.
C. Find a minimax estimator d_M of $\underline{\theta}$.

Solution: 1. A. If $\underline{\theta} \sim$ Dirichlet($\underline{\alpha}$), then $\theta_j \sim \text{Beta}(\alpha_j, \sum_{k \neq j} \alpha_k)$, and hence from our computations of the mean of a Beta, $E(\theta_j) = \alpha_j / \sum_{i=1}^k \alpha_i$, and as a vector $E(\underline{\theta}) = \underline{\alpha} / \sum_{i=1}^k \alpha_i$. Since the posterior distribution of θ is Dirichlet($\underline{\alpha} + \underline{X}$), the Bayes estimator of θ for squared error loss is

$$E(\underline{\theta} | \underline{X}) = \frac{\underline{\alpha} + \underline{X}}{\sum \alpha_i + n} .$$

B. As noted in A, the posterior density is Dirichlet($\underline{\alpha} + \underline{X}$):

$$\lambda(\theta | \underline{X}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k + n)}{\prod_{j=1}^k \Gamma(\alpha_j + X_j)} \theta_1^{\alpha_1 + X_1 - 1} \cdots \theta_k^{\alpha_k + X_k - 1} 1_{[\underline{\theta}: \sum \theta_j = 1]}.$$

To find the mode of the posterior, we need to find the value of $\underline{\theta}$ which maximizes $\lambda(\underline{\theta} | \underline{X})$ over the set $\sum_j \theta_j = 1$, or equivalently which maximizes

$$\sum_{j=1}^k (\alpha_j + X_j - 1) \log \theta_j + c(\sum_{j=1}^k \theta_j - 1).$$

Thus we need to solve

$$\frac{\alpha_j + X_j - 1}{\theta_j} + c = 0 , \qquad j = 1, \cdots, k, \tag{0.1}$$

and

$$\sum_{j=1}^k \theta_j = 1. \tag{0.2}$$

The first equation yields

$$\theta_j^{mode} = \frac{\alpha_j + X_j - 1}{-c}, \qquad j = 1, \ldots, k\,.$$

Note that $\theta_j^{mode} \geq 0$ if $\alpha_j + X_j \geq 1$, which always holds if $\alpha_j \geq 1$, and holds for $X_j \geq 1$ if $0 < \alpha_j < 1$. Substitution of this into (0.2) yields

$$1 = \sum_{j=1}^{k} \theta_j^{mode} = -\frac{1}{c}(\sum_{j=1}^{k} \alpha_j + n - k)$$

and hence $-c = \sum_j \alpha_j + n - k$. Thus the mode of the posterior is given by

$$\underline{\theta}^{mode} = \frac{\underline{\alpha} + \underline{X} - \underline{1}}{\sum \alpha_j + n - k}.$$

When $\underline{\alpha} = \underline{1}$ (the vector of all 1's), then the mode of the posterior equals the MLE: $\widehat{\theta} = \underline{X}/n$. Note that $\underline{\alpha} = \underline{1}$ yields a uniform prior over θ.
C. If $\underline{X} \sim \text{Mult}_k(n; \underline{\theta})$ and $\underline{\theta} \sim \text{Dirichlet}(\underline{\alpha})$, then the Bayes estimator of $\underline{\theta}$ for squared error loss is $d_\Lambda(\underline{X}) = (\underline{\alpha} + \underline{X})/(\sum \alpha_i + n)$. For $\alpha_1 = \ldots = \alpha_k = \alpha$, this yields the Bayes estimator

$$d_\Lambda(\underline{X}) = \frac{\alpha\underline{1} + \underline{X}}{k\alpha + n} = \frac{k\alpha}{k\alpha + n}\frac{\underline{1}}{k} + \frac{n}{k\alpha + n}\frac{\underline{X}}{n}$$

with risk

$$\begin{aligned}
R(\underline{\theta}, \underline{d}_\Lambda) &= E_{\underline{\theta}}|\underline{\theta} - \underline{d}_\Lambda(\underline{X})|^2 \\
&= \sum_{i=1}^{k}\{Var_{\underline{\theta}}(d_{\Lambda i}(\underline{X})) + \text{bias}^2_{\underline{\theta}}(d_{\Lambda i})\} \\
&= \frac{1}{(k\alpha + n)^2}\sum_{i=1}^{k}\{n\theta_i(1-\theta_i) + (\alpha - k\alpha\theta_i)^2\} \\
&= \frac{1}{(k\alpha + n)^2}\{n - k\alpha^2 + (\alpha^2k^2 - n)\sum_{i=1}^{k}\theta_i^2\} \qquad \text{since} \qquad \sum\theta_i = 1 \\
&= \frac{(1 - 1/k)}{(1 + \sqrt{n})^2} \qquad \text{if} \qquad \alpha = \frac{\sqrt{n}}{k}
\end{aligned}$$

which is constant in $\underline{\theta}$. Hence by corollary 5.6.3

$$d_\Lambda(\underline{X}) = \frac{\sqrt{n}}{\sqrt{n} + n}\frac{1}{k} + \frac{n}{\sqrt{n} + n}\frac{\underline{X}}{n} \equiv (1 - \lambda_n)\frac{\underline{1}}{k} + \lambda_n\widehat{\underline{p}}_n$$

is minimax for estimation of $\underline{\theta}$.

3. Let P and Q be two probability measures on a measurable space $(\mathcal{X}, \mathcal{B})$. Then the Hellinger distance $d_H(P,Q)$ is defined by

$$d_H^2(P,Q) = \frac{1}{2}\int [\sqrt{p} - \sqrt{q}]^2 d\mu$$

where $p = dP/d\mu$, $q = dQ/d\mu$, and μ is any measure dominating P and Q (e.g. $P+Q$). [Note that this differs only by a factor of $1/2$ from the definition we used earlier.] Similarly, the total variation distance $d_{TV}(P,Q)$ can be defined by

$$d_{TV}(P,Q) = \frac{1}{2}\int |p-q| d\mu.$$

A. Show that

$$d_{TV}(P,Q) = \sup_{A \in \mathcal{B}} |P(A) - Q(A)|.$$

[Hint: Let $\delta = p - q$. Since $\int p d\mu = \int q d\mu = 1$, it follows that $\int \delta d\mu = 0$ and $\int_A \delta d\mu = -\int_{A^c} \delta d\mu$ for any set A.]
Comment: often $d_{TV}(P,Q)$ is defined by the supremum on the right side above (without the factor of 2); then it would be proved that this equals $1/2$ times the integral ($L_1(\mu)$ distance from p to q) that I used as the definition above.
B. Show that $d_{TV}(P,Q) = 1 - \int p \wedge q d\mu$.
C. Show that

$$d_H^2(P,Q) \le d_{TV}(P,Q) \le d_H(P,Q)\{2 - d_H^2(P,Q)\}^{1/2} \le \sqrt{2} d_H(P,Q).$$

[Hint: Use B to prove the first inequality; use $|p-q| = |p^{1/2} - q^{1/2}||p^{1/2} + q^{1/2}|$ and the Cauchy - Schwarz inequality to prove the second inequality.]

Solution: A. Let $\delta \equiv p - q$. For any fixed set $A \in \mathcal{A}$,

$$\int_A \delta d\mu = -\int_{A^c} \delta d\mu, \tag{0.3}$$

and hence

$$\begin{aligned}
2|P(A) - Q(A)| &= 2|\int_A \delta d\mu| && (0.4)\\
&= |\int_A \delta d\mu| + |\int_{A^c} \delta d\mu| \qquad \text{by (0.3)} && (0.5)\\
&\le \int_A |\delta| d\mu + \int_{A^c} |\delta| d\mu && (0.6)\\
&= \int |\delta| d\mu, && (0.7)
\end{aligned}$$

and equality holds if $A \equiv \{x : \delta(x) \geq 0\}$. Thus

$$2 \sup_{A \in \mathcal{A}} |P(A) - Q(A)| = \int |p - q| d\mu \equiv 2 d_{TV}(P, Q) .$$

B. From the proof of A,

$$\begin{aligned} d_{TV}(P,Q) &= \frac{1}{2} \int |p-q| d\mu = \int_{p>q} (p-q) d\mu \\ &= \{ \int_{p \leq q} p d\mu + \int_{p>q} p d\mu - \int_{p \leq q} p d\mu - \int_{p>q} q d\mu \} \\ &= \{1 - \int p \wedge q d\mu \}. \end{aligned} \tag{0.8}$$

C. By elementary reasoning

$$\begin{aligned} \int p \wedge q d\mu &= \int_{p \leq q} p + \int_{p>q} q \\ &= \int_{p \leq q} \sqrt{p}\sqrt{p} + \int_{p>q} \sqrt{q}\sqrt{q} \\ &\leq \int_{p \leq q} \sqrt{p}\sqrt{q} + \int_{p>q} \sqrt{q}\ \sqrt{p} = \int \sqrt{pq} d\mu \\ &= \rho(P,Q) . \end{aligned} \tag{0.9}$$

Consequently, from (0.8) (0.9), and $d_H^2(P,Q) = (1 - \rho(P,Q))$ we obtain

$$d_{TV}(P,Q) = (1 - \int p \wedge q d\mu) \geq (1 - \rho(P,Q)) = d_H^2(P,Q).$$

To prove the upper bound, use $|p-q| = |\sqrt{p} - \sqrt{q}||\sqrt{p} + \sqrt{q}|$ to write

$$\begin{aligned} d_{TV}(P,Q) &= \frac{1}{2} \int |p-q| d\mu \\ &= \frac{1}{2} \int |\sqrt{p} - \sqrt{q}||\sqrt{p} + \sqrt{q}| d\mu \\ &\leq \frac{1}{2} \left\{ \int |\sqrt{p} - \sqrt{q}|^2 d\mu \int |\sqrt{p} + \sqrt{q}|^2 d\mu \right\}^{1/2} \\ &= d_H(P,Q) \frac{1}{\sqrt{2}} \{2 + 2\rho(P,Q)\}^{1/2} \\ &= d_H(P,Q) \left\{ 1 + 1 - d_H^2(P,Q) \right\}^{1/2} \\ &\leq \sqrt{2} d_H(P,Q) . \end{aligned}$$

Thus the total variation and Hellinger metrics generate the same topology on probability distributions.