Statistics 582, Problem Set 6 Solutions

Wellner; 2/15/2018

1. 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) Suppose that instead of the Gamma prior distribution, θ has the Pareto(θ_0, α) distribution with density λ given by

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

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) 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 Section 5.8 of the course notes concerning 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 converge to point mass at θ_0.

2. Specialize the decision rule in Theorem 5.2 of the course notes to the case when P_i is the normal distribution $N_d(\mu_i, I)$, $i = 1, \ldots, k$ where $\mu_1, \ldots, \mu_k$ are distinct vectors in $\mathbb{R}^d$, $\mu_i \neq \mu_j$ for $i \neq j$. What happens if we replace I by Σ?

Solution: When $P_i = N_d(\mu_i, I)$, the inequality $\lambda_i p_i(x) > \lambda_j p_j(x)$ can be written as

$$\lambda_i \exp(-\frac{1}{2}(x-\mu_i)^T(x-\mu_i)) > \lambda_j \exp(-\frac{1}{2}(x-\mu_j)^T(x-\mu_j)),$$

or, equivalently, assuming that $\lambda_i \neq 0$ and $\lambda_j \neq 0$,

$$(\mu_i - \mu_j)^T x > \frac{1}{2}(\mu_i^T\mu_i - \mu_j^T\mu_j) + \log(\lambda_j/\lambda_i).$$

When $\mu_i \neq \mu_j$, this set of x's corresponds to a half-space bounded by a hyperplane orthogonal to $\mu_i - \mu_j$. Moreover, if $p_i = p_j$, this hyperplane is the bisector of the line segment from μ_i to μ_j. When we consider all the $k-1$ mean vectors μ_j with $j \neq i$, it becomes clear that the set of x's for which $d(i|x) = 1$ is the intersection of $k-1$ half spaces, and this yields a convex polyhedron with at most $k-1$ faces. When $\lambda_i = 1/k$ for $i = 1, \ldots, k$, then we classify X as belonging to the (normal) distribution with mean μ_i that is closest to X. When the identity covariance matrix I is replaced by an arbitrary nonsingular covariance matrix Σ, this remains true with ordinary Euclidean distance replaced by $d^2_\Sigma(x,\mu) \equiv (x-\mu_i)^T\Sigma^{-1}(x-\mu_i)$.

3. Lehmann and Casella, TPE, Problem 5.17, page 293, parts (a) and (c) (Also note Problems 5.18, 5.19, 5.20, page 293.) The original proof of Theorem 5.7 (Lehmann and Casella page 260), used Rényi's entropy functions (Rényi, 1961)

$$R_\alpha(f,g) = \frac{1}{\alpha-1}\log\int f^\alpha(x)g^{1-\alpha}(x)d\mu(x)$$

where f and g are densities, μ is a dominating measure, and $\alpha \neq 1$ is a constant.
(a) Show that $R_\alpha(f,g)$ satisfies $R_\alpha(f,g) \geq 0$ and $R_\alpha(f,f) = 0$.
(c) Show that $\lim_{\alpha\to 1} R_\alpha(f,g) = K(f,g)$.

Solution: (a) First,

$$\begin{aligned} R_\alpha(f,f) &= \log(\int f^\alpha f^{1-\alpha}d\mu)/(\alpha-1) = \log(\int f d\mu)/(\alpha-1) \\ &= \log(1)/(\alpha-1) = 0. \end{aligned}$$

Next, for $0 < \alpha < 1$, by Hölder's inequality with $p = 1/\alpha$, $q = 1/(1-\alpha)$, so that $1/p + 1/q = \alpha + (1-\alpha) = 1$,

$$\begin{aligned} 0 \leq \int f^\alpha g^{1-\alpha}d\mu &\leq \left(\int (f^\alpha)^{1/\alpha}d\mu\right)^\alpha \left(\int (g^{1-\alpha})^{1/(1-\alpha)}d\mu\right)^{1-\alpha} \\ &= \left(\int f d\mu\right)^\alpha \left(\int g d\mu\right)^{1-\alpha} = 1^\alpha 1^{1-\alpha} = 1, \end{aligned}$$

so $\log\left(\int f^\alpha g^{1-\alpha} d\mu\right) \le 0$, and it follows that $R_\alpha(f,g) \ge 0$. For $\alpha > 1$ or $\alpha < 0$, the function $r_\alpha(u) = u^\alpha$ is convex, and hence by Jensen's inequality

$$\begin{aligned}\int f^\alpha g^{1-\alpha} d\mu &= \int g(x) r_\alpha(f(x)/g(x)) d\mu(x) \\ &\ge r_\alpha\left(\int g(f/g) d\mu\right) = r_\alpha(1) = 1.\end{aligned}$$

Thus $R_\alpha(f,g) \ge \log 1/(\alpha-1) = 0$ when $\alpha > 1$. Rényi (1961) seems to have required $\alpha > 0$ (and this is missing from the statement in Lehmann and Casella). On the other hand with the factor $1/(\alpha-1)$ replaced by $1/(\alpha(\alpha-1))$, it continues to be true that $R_\alpha(f,g) \ge 0$ for $\alpha < 0$.

(b) By definition,

$$\begin{aligned}&R_\alpha(\pi(\lambda|X), \psi(\lambda)) \\ &= \log\left(\int \pi(\lambda|X)^\alpha \psi(\lambda)^{1-\alpha} d\lambda\right)/(\alpha-1) \\ &= \log\left(\int (\pi(\lambda|X)/\psi(\lambda))^\alpha \psi(\lambda) d\lambda\right)/(\alpha-1)\end{aligned}$$

where

$$\frac{\pi(\lambda|X)}{\psi(\lambda)} = \int_\Theta \frac{f(X|\theta)}{m(X)} \pi(\theta|\lambda) d\theta = E\left\{\frac{f(X|\theta)}{m(X)}\right\}$$

where the integration in the last expectation is with respect to $\pi(\theta|\lambda)$. Thus by Jensen's inequality for $\alpha > 1$

$$\begin{aligned}&R_\alpha(\pi(\lambda|X), \psi(\lambda)) \\ &= \log\left(\int \left[E\left\{\frac{f(X|\theta)}{m(X)}\right\}\right]^\alpha \psi(\lambda) d\lambda\right)/(\alpha-1) \\ &\le \log\left(\int E\left(\frac{f(X|\theta)}{m(X)}\right)^\alpha \psi(\lambda) d\lambda\right)/(\alpha-1) \\ &= \log\left(\int\left(\int\left(\frac{f(X|\theta)}{m(X)}\right)^\alpha \pi(\theta|\lambda) d\theta\right)\psi(\lambda) d\lambda\right)/(\alpha-1) \\ &= \log\left(\int\left(\frac{f(X|\theta)}{m(X)}\right)^\alpha \int \pi(\theta|\lambda)\psi(\lambda) d\lambda\, d\theta\right)/(\alpha-1) \\ &= \log\left(\int\left(\frac{f(X|\theta)\pi(\theta)}{m(X)}\right)^\alpha \pi(\theta)^{1-\alpha}\, d\theta\right)/(\alpha-1) \\ &= \log\left(\int \pi(\theta|X)^\alpha \pi(\theta)^{1-\alpha} d\theta\right)/(\alpha-1) \\ &= R_\alpha(\pi(\theta|X), \pi(\theta)).\end{aligned}$$

The same argument works when $0 < \alpha < 1$ using Jensen's inequality again, this time with concavity of $u \mapsto u^\alpha$ so that the inequality is reversed, and noticing that $\alpha - 1 < 0$. Note that the resulting family of inequalities given in Theorem 5.7 and this problem says that, in the sense of Rényi's divergence or Kullback-Leibler divergence, *the data has less effect on hyperpriors than priors*, or, said another way, *the posterior distribution of a hyperparameter is less affected by changes in the prior than the posterior distribution of a parameter.*
(c) When $\alpha \to 1$, both the numerator and denominator of the definition of $R_\alpha(f,g)$ converge to 0, so applying L'Hopital's rule (differentiating both numerator and denominator and taking limits again) yields

$$\begin{aligned}
\lim_{\alpha\to 1} R_\alpha(f,g) &= \lim_{\alpha\to 1} \frac{\int g \exp(\alpha \log(f/g)) \log(f/g) d\mu}{\int f^\alpha g^{1-\alpha} d\mu} \\
&= \frac{\int g \exp(\log(f/g)) \log(f/g) d\mu}{\int f d\mu} \\
&= \frac{\int g(f/g) \log(f/g) d\mu}{1} = \int f \log(f/g) d\mu \\
&= K(f,g).
\end{aligned}$$

Note that if we replace $\alpha - 1$ in the denominator by $\alpha(\alpha - 1)$, then the preceeding argument goes through with only a minor change, while now

$$\begin{aligned}
\lim_{\alpha\to 0} R_\alpha(f,g) &= \lim_{\alpha\to 0} \frac{\int g \exp(\alpha \log(f/g)) \log(f/g) d\mu}{(2\alpha - 1) \int f^\alpha g^{1-\alpha} d\mu} \\
&= \frac{\int g \exp(0 \cdot \log(f/g)) \log(f/g) d\mu}{-\int g d\mu} \\
&= \frac{\int g \log(f/g) d\mu}{-1} = \int g \log(g/f) d\mu \\
&= K(g,f).
\end{aligned}$$

4. Problem 3.9, Lehmann and Casella, TPE, page 286. For the natural exponential family $p_\eta(x)$ of (4.3.7) and the conjugate prior $\pi(\eta|k,\mu)$ of (4.3.19) establish that:
(a) $E(X) = A'(\eta)$ and $Var(X) = A''(\eta)$ where the expectation is with respect to the sampling density $p_\eta(x)$.
(b) $EA'(\eta) = \mu$ and $Var(A(\eta)) = (1/k)EA''(\eta)$, where the expectation is with respect to the prior distribution.

Solution: (a) Suppose that $p_\eta(x) = \exp(\eta x - A(\eta))$ is a density with respect to the dominating measure μ where $\eta \in H$, a non-empty open interval contained in $\mathbb{R}$ and $x \in \mathcal{X} \subset \mathbb{R}$. Since $1 = \int_{\mathcal{X}} p_\eta(x) d\mu(x)$ for all $\eta \in H$, it follows from Theorem

1.5.8 of Lehmann and Casella that we may change the order of differentiation to conclude that

$$\begin{aligned} 0 &= \int_{\mathcal{X}} \frac{\partial}{\partial \eta} p_\eta(x) d\mu(x) \\ &= \int_{\mathcal{X}} \exp(\eta x - A(\eta))(x - A'(\eta)) d\mu(x) \\ &= E_\eta(X) - A'(\eta). \end{aligned}$$

Therefore $E_\eta(X) = A'(\eta)$. Differentiation across this identity yields

$$0 = \int_{\mathcal{X}} e^{\eta x - A(\eta)} (x - A'(\eta))^2 d\mu(x) - A''(\eta).$$

Combining this with the identity in the previous display yields

$$Var_\eta(X) = A''(\eta).$$

(b) Now consider the prior density π with respect to Lebesgue measure on $\mathbb{R}$ given by

$$\pi(\eta | k, \mu) = c(k, \mu) \exp(k\mu\eta - kA(\eta)).$$

Here we differentiate with respect to the argument of the density rather than (either of the two) parameter(s). Since $1 = \int_H \pi(\eta|k,\mu) d\eta$, it follows that

$$\begin{aligned} 0 &= \int_H \frac{\partial}{\partial \eta} \pi(\eta|k,\mu) d\eta \\ &= \int_H \pi(\eta|k,\mu)(k\mu - kA'(\eta)) d\eta \\ &= E\{k\mu - kA'(\eta)\} = k(\mu - EA'(\eta)), \end{aligned}$$

which yields $EA'(\eta) = \mu$. Differentiation across this identity again yields

$$\begin{aligned} 0 &= \int_H \frac{\partial^2}{\partial \eta^2} \pi(\eta|k,\mu) d\eta \\ &= E(k\mu - kA'(\eta))^2 - kEA''(\eta), \end{aligned}$$

and hence $k^2 Var(A'(\eta)) = kEA''(\eta)$, or, equivalently

$$Var(A'(\eta)) = k^{-1} EA''(\eta).$$

For justification of the interchange of derivative and integral in the latter identities and further explanation of conjugate priors for $k-$dimensional exponential families, see Diaconis and Ylvisaker (1979), *Ann. Statist.* **7**, 269 - 281.