

Statistics 582, Problem Set 6 Solutions

Wellner; 2/18/2015

1. Continuation of problem 2, problem set 4: Suppose that X_1, \dots, X_n are i.i.d. Exponential(θ) (so the X 's have distribution P_θ and density $p_\theta(x) = \theta e^{-\theta x} 1_{(0, \infty)}(x)$) with respect to Lebesgue measure on \mathbb{R} , and that $\theta \sim \Gamma(\alpha, \beta)$:

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

In problem set 4 we found the Bayes rules with respect to squared error loss $L(\theta, a) = (\theta - a)^2$ and weighted squared error loss $L(\theta, a) = (\theta - a)^2/\theta$.

- (a) Prove a (conditional) limit theorem for the posterior distributions given \underline{X} .
(b) What does theorem 5.8.2 say about the limiting distribution of the Bayes rule for squared error loss (assuming that X_1, \dots, X_n are i.i.d. $P_{\theta_0} \equiv P$ with $\theta_0 \in (0, \infty)$)?
(c) Can you say anything about the asymptotic behavior of the Bayes rule with respect to weighted squared error loss?

Solution: (a) There are several possible ways of proceeding here: (i) verify the hypotheses of Theorem 8.1 of the notes; (ii) verify the hypotheses of Theorem 10.1 of van der Vaart's *Asymptotic Statistics*; (iii) give a direct proof in this special case of convergence in distribution; or (iv) give a direct proof in this special case of convergence in total variation distance by showing that the densities converge pointwise followed by Scheffé's lemma. Both (i) and (ii) are made difficult by conditions B2 and B3 (in the case of Theorem 8.1) and by the "separation by tests" condition in van der Vaart's Theorem 10.1. Thus proceed directly as in (iii). First, note that

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

where $Y_0 \sim \text{Gamma}(\alpha, 1)$, and $Y_i \sim \text{Gamma}(1, 1) = \text{Exp}(1)$, $i = 1, \dots, n$ 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 ,

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

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

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

and similarly with T_n in place of $1/\bar{X}_n$.

To show that the densities of $\sqrt{n}(\theta - E(\theta|\underline{X}))$ converge pointwise, first consider the distribution function and density of the unscaled version of the lead term, $\sqrt{n}(\bar{Y}_n - 1)$: since $n\bar{Y}_n = \sum_1^n Y_i \sim \text{Gamma}(n, 1)$,

$$\begin{aligned}
F_{\sqrt{n}(\bar{Y}_n - 1)}(z) &= P(\bar{Y}_n \leq 1 + n^{-1/2}z) = P\left(\sum_1^n Y_i \leq n + \sqrt{nz}\right) \\
&= \int_0^{n+\sqrt{nz}} \frac{t^{n-1}}{\Gamma(n)} \exp(-t) dt.
\end{aligned}$$

Thus, differentiating and then using

$$\Gamma(n) = (n-1)! \sim \sqrt{2\pi(n-1)} \left(\frac{n-1}{e}\right)^{n-1}$$

by Stirling's formula, we find that

$$\begin{aligned}
f_{\sqrt{n}(\bar{Y}_n - 1)}(z) &= \frac{(n + \sqrt{nz})^{n-1}}{\Gamma(n)} \exp(-(n + \sqrt{nz})) \sqrt{n} \\
&\sim \frac{(n + \sqrt{nz})^{n-1}}{\sqrt{2\pi(n-1)} \left(\frac{n-1}{e}\right)^{n-1}} \cdot \exp(-(n + \sqrt{nz})) \cdot \sqrt{n} \\
&= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{n-1}} \cdot \left(1 + \frac{1 + \sqrt{nz}}{n-1}\right)^{n-1} \cdot \exp(-(1 + \sqrt{nz})) \\
&\rightarrow \frac{1}{\sqrt{2\pi}} \exp(-z^2/2).
\end{aligned}$$

Here the convergence follows by letting $a_n \equiv 1 + \sqrt{n+1}z$ and noting that

$$\begin{aligned}
& \left(1 + \frac{1 + \sqrt{n+1}z}{n}\right)^n \cdot \exp(-(1 + \sqrt{n+1}z)) \\
&= \left(1 + \frac{a_n}{n}\right)^n \cdot \exp(-a_n) \\
&= \left\{\left(1 + \frac{a_n}{n}\right) \cdot \exp(-a_n/n)\right\}^n \\
&\approx \left\{\left(1 + \frac{a_n}{n}\right) \left(1 - \frac{a_n}{n} + \frac{1}{2} \frac{a_n^2}{n} + O(n^{-3/2})\right)\right\}^n \\
&= \left\{1 - \frac{1}{2} \frac{a_n^2}{n^2} + O(n^{-3/2})\right\}^n \\
&= \left\{1 - \frac{1}{2} \frac{(1 + \sqrt{n+1}z)^2}{n}\right\}^n \\
&\rightarrow \exp(-z^2/2).
\end{aligned}$$

Since $\bar{X}_n + n^{-1}\beta \xrightarrow{a.s.} \theta_0^{-1}$ and $\sqrt{n}(Y_0 - \alpha) \frac{1/n}{\bar{X}_n + n^{-1}\beta} \xrightarrow{a.s.} 0$, it follows (via the convolution formula) that the density of $\sqrt{n}(\theta - E(\theta|\underline{X}))$ converges pointwise to $\phi(z/\theta_0)/\theta_0$, the density of $N(0, \theta_0^2)$.

(b) In the present case Theorem 5.8.2 says that

$$\sqrt{n}(E(\theta|\underline{X}) - \theta_0) \rightarrow_d N(0, 1/I(\theta_0)) = N(0, \theta_0^2)$$

since $I(\theta_0) = 1/\theta_0^2$. This also follows from a direct argument since

$$\begin{aligned}
\sqrt{n}(E(\theta|\underline{X}) - \theta_0) &= \sqrt{n} \left(\frac{1 + \alpha/n}{\bar{X}_n + \beta/n} - \theta_0 \right) \\
&= \sqrt{n}(\alpha/n + 1 - \theta_0(\beta/n + \bar{X}_n))/(\beta/n + \bar{X}_n) \\
&= \left\{ -\theta_0\sqrt{n}(\bar{X}_n - 1/\theta_0) + n^{-1/2}(\alpha - \theta_0\beta) \right\} / (\beta/n + \bar{X}_n) \\
&\rightarrow_d \theta_0^2 N(0, \theta_0^{-2}) = N(0, \theta_0^2).
\end{aligned}$$

2. Specialize the decision rule in Theorem 5.2 of the course notes to the case when P_i is the normal distribution $N_d(\mu_i, I)$, $i = 1, \dots, k$ where μ_1, \dots, μ_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

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

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 $i = 1, \dots, k$, 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 $d_{\Sigma}^2(x, \mu) \equiv (x - \mu)^T \Sigma^{-1} (x - \mu)$.

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.)

Solution: (a) First,

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

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

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

Thus $R_{\alpha}(f, g) \geq \log 1/(\alpha - 1) = 0$ 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 $R_\alpha(f, g) \geq 0$ 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 $\alpha > 1$

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

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 $\alpha \rightarrow 1$, 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

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

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

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

4. **Optional bonus problem 1:** (Birgé). Let $X = (X_0, X_1, \dots, X_k)$ be a $(k + 1)$ -dimensional vector, and assume that $X \sim N_{k+1}(\theta, I_{k+1})$ where I_{k+1} denotes the $(k + 1) \times (k + 1)$ identity matrix. For any vector $\theta \in \mathbb{R}^{k+1}$, let θ' denote the projection of θ onto the k -dimensional linear space spanned by the k -last coordinates. Consider the subset Θ_0 of $\Theta = \mathbb{R}^{k+1}$ given by

$$\Theta_0 = \{\theta \in \mathbb{R}^{k+1} : |\theta_0| \leq k^{1/4} \text{ and } \|\theta'\| \leq 2(1 - k^{-1/4}|\theta_0|)\}.$$

- (a) Show that the MLE of θ over Θ_0 is given by $\hat{\theta}_0 = 0$ and $\hat{\theta}' = 2X'/\|X\|$ on the event

$$\Omega_0 \equiv \{\|X'\|^2 > 3k/4 \text{ and } |X_0| < k^{1/4} + 1.21\}.$$

- (b) Show that $P_\theta(\Omega_0) \geq 3/4$ for all $\theta \in \Theta_0$.
(c) Let $\tilde{\theta} = (X_0, \underline{0})$. Show that for $k \geq 128$ we have

$$\begin{aligned} \sup_{\theta \in \Theta_0} E_\theta \|\theta - \hat{\theta}\|^2 &\geq (3/4)\sqrt{k} + 3, \text{ and} \\ \sup_{\theta \in \Theta_0} E_\theta \|\theta - \tilde{\theta}\|^2 &\leq 5. \end{aligned}$$

Thus the maximal risk of the MLE may be much larger than the minimax risk when k is large.

Hint: A non-central χ_k^2 distribution is stochastically larger than a central χ_k^2 distribution; then use Lemma 1 of Laurent and Massart (2000) [Laurent, B.; Massart, P. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** (2000), 13021338].

Solution: Let $c = k^{1/4}$. Since $k \geq 128$

$$c \geq 3.36 \quad \text{and} \quad \sqrt{3k} - \sqrt{k} - 4 - 1.21k^{1/4} > 0. \quad (1)$$

It follows that, whatever the true value of $\theta \in \Theta_0$, $P_\theta(|X_0| \geq c + 1.21) \leq 0.114$ and that $\|X'\|^2$ has a non-central χ_k^2 distribution. Since a non-central χ^2 random variable is stochastically larger than a central one, it follows from Laurent and Massart (2000), Lemma 1 that

$$P_\theta(\|X'\|^2 \leq k - 2\sqrt{kx}) \leq e^{-x} \quad \text{for } x > 0.$$

Setting $x = k/64 \geq 2$, we conclude, since $e^{-x} < 0.136$, that

$$P_\theta(\Omega_0) > 3/4.$$

Now assume that the event Ω_0 holds. Since the MLE is the least squares estimator, $\hat{\theta}$ is the minimizer over Θ_0 of $(X_0 - \theta_0) + \|X' - \theta'\|^2$. On Ω_0 , $\|X'\| > \sqrt{3k}/2 > \|\theta'\|$ and, given θ_0 , the minimum with respect to θ' is obtained for $\theta' = 2X'(1 - |\theta_0|/c)/\|X'\|$ with value

$$f(\theta_0) = (X_0 - \theta_0)^2 + (\|X'\| - 2(1 - |\theta_0|/c))^2.$$

Since for $\theta_0 \neq 0$,

$$\begin{aligned} (c/2)\theta_0 f'(\theta_0) &= c(\theta_0 - X_0)\theta_0 + 2|\theta_0|(\|X'\| - 2(1 - |\theta_0|/c)) \\ &> |\theta_0|(2\|X'\| - 4 - c(c + 1.21)) \\ &\geq |\theta_0|(\sqrt{3k} - 4 - \sqrt{k} - 1.21k^{1/4}) \end{aligned}$$

is non-negative by (1), $f(\theta_0)$ is minimal when $\theta_0 = 0$. Therefore, if Ω_0 holds, $\hat{\theta}_0 = 0$ and $\hat{\theta}' = 2X'/\|X'\|$. This implies that the quadratic risk at $s = (s_0, 0)$ of the MLE is bounded from below by $(3/4)(\theta_0^2 + 4)$ with maximum value $(3/4)\sqrt{k} + 3$ when $|\theta_0| = c$. On the other hand, the estimator $\tilde{\theta}$ with $\tilde{\theta}_0 = X_0$ and $\tilde{\theta}' = 0$ has a quadratic risk which is uniformly bounded by 5.

5. **Optional bonus problem 2:** Problem 3.9, Lehmann and Casella, TPE, page 286.

(a) Suppose that $p_\eta(x) = \exp(\eta x - A(\eta))h(x)$ as in 4.3.18, Lehmann and Casella page 244. Show that $E_\eta(X) = A'(\eta)$ and $Var_\eta(X) = A''(\eta)$.

(b) Suppose that $\pi(\eta; k, \mu) = c(k, \mu) \exp(k\mu\eta - kA(\eta))$ as in 4.3.19, Lehmann and Casella page 244. Show that $EA'(\eta) = \mu$ and that $Var(A(\eta)) = EA''(\eta)/k$ where the expectations are with respect to the prior distribution.

Solution: (a) Now $\log p_\eta(x) = \eta x - A(\eta) + \log h(x)$ so

$$\dot{l}_\eta(x) = x - A'(\eta), \quad \text{and} \quad \ddot{l}_{\eta,\eta}(x) = -A''(\eta).$$

Thus $0 = E_\eta \dot{l}_\eta(X) = E_\eta(X) - A'(\eta)$, and hence $E_\eta(X) = A'(\eta)$. Similarly,

$$A''(\eta) = -E_\eta \ddot{l}_{\eta,\eta}(X) = E_\eta \dot{l}_\eta^2(X) = E_\eta(X - E_\eta X)^2 = Var_\eta(X).$$

(b) Now $\log \pi(\eta; k, \mu) = \log c(k, \mu) + k\mu\eta - kA(\eta)$, so

$$\begin{aligned} \frac{\partial}{\partial \eta} \log \pi(\eta; k, \mu) &= (k\mu - kA'(\eta)), & \frac{\partial}{\partial \eta} \pi(\eta; k, \mu) &= k(\mu - A'(\eta))\pi(\eta; k, \mu), \quad \text{and} \\ \frac{\partial^2}{\partial \eta^2} \pi(\eta; k, \mu) &= (k\mu - kA'(\eta))^2 \pi(\eta; k, \mu) - kA''(\eta)\pi(\eta; k, \mu). \end{aligned}$$

Thus we find that

$$kE(\mu - A'(\eta)) = \int \frac{\partial}{\partial \eta} \pi(\eta; k, \mu) d\eta = 0$$

by the fundamental theorem of calculus, and hence $EA'(\eta) = \mu$. Similarly,

$$k^2 E(\mu - A'(\eta))^2 - kEA''(\eta) = \int \frac{\partial^2}{\partial \eta^2} \pi(\eta; k, \mu) d\eta = 0,$$

and hence $Var(A'(\eta)) = k^{-1}EA''(\eta)$. (Note: these calculations generalize to the k -dimensional case; see Diaconis and Ylvisaker (1979), *Ann. Statist.* **7**, 269-281.)