**Statistics 582, Problem Set 7 Solutions**

Wellner; 2/23/2011

1. Suppose that $X \sim P_\theta$ for $\theta \in \Theta \subset R^k$ has well-defined Fisher information matrix $I(\theta)$ for $\theta$. The *Jeffreys prior* distribution $\Lambda_J$ has density $\lambda_J(\theta) = \det(I(\theta))^{1/2}$ with respect to Lebesgue measure on $\Theta$. Note that $\Lambda_J$ may not be a finite measure, and even if $\Lambda_J$ is a finite measure, it may not have total mass 1. If a prior distribution is a finite measure, then call it a *proper prior distribution*, and correspondingly if it is not a finite measure, call it an *improper prior distribution*. If the resulting posterior distribution is a finite measure, call it a *proper posterior distribution*, and (by convention) normalize it to have total mass 1. See Lehmann and Casella, TPE, pages 230, 234, 287, 305.
(a) Suppose that $X \sim$ Bernoulli$(\theta)$. Find the Jeffrey's prior density $\lambda_J$ for $\theta$. Is $\Lambda_J$ a finite measure? If it is finite, what is $\Lambda_J((0,1))$? Find the corresponding posterior distribution of $\Theta$ starting with the Jeffrey's prior.
(b) Suppose that $X \sim$ Poisson$(\theta)$ with $\theta \in (0,\infty)$. Find the Jeffrey's prior density $\lambda_J$ for $\theta$. Is $\Lambda_J$ a finite measure? If it is finite, what is $\Lambda_J((0,\infty))$? Find the corresponding posterior distribution of $\Theta$ starting with the Jeffrey's prior. Is it ever a proper posterior distribution?
(c) Suppose that $X \sim$ Geometric$(\theta)$, i.e. the number of trials until the first success in i.i.d. Bernoulli trials with probability $\theta$ of success for each trial – recall Chapter 1, section 1. Find the Jeffrey's prior density $\lambda_J$ for $\theta$. Is $\Lambda_J$ a finite measure? If it is finite, what is $\Lambda_J((0,1))$? Find the corresponding posterior distribution of $\Theta$ starting with the Jeffrey's prior. If we observe $X_1, \ldots, X_n$ i.i.d. Geometric$(\theta)$, so that $\sum X_i \sim$ Negative Binomial$(n,\theta)$ is the posterior distribution "proper" for some $n$?
(d) Suppose that $X \sim$ Weibull$(\theta)$ with $\theta = (\alpha,\beta) \in (0,\infty)\times(0,\infty)$ as in chapters 3 and 4. Find the Jeffrey's prior density $\lambda_J$ for $\theta$. Is $\Lambda_J$ a finite measure? If it is finite, what is $\Lambda_J((0,\infty)^2)$? Find the corresponding posterior distribution of $\Theta$ starting with the Jeffrey's prior.

**Solution:** (a) When $X \sim$ Bernoulli$(\theta)$, the Information for $\theta$ is $I(\theta) = \{\theta(1-\theta)\}^{-1}$, so the Jeffrey's prior for $\theta$ is has density

$$\lambda_J(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}}\,.$$

This density is proportional to the Beta$(1/2,1/2)$ density

$$\lambda(\theta) = \frac{\Gamma(1)}{\Gamma(1/2)\Gamma(1/2)}\theta^{1/2-1}(1-\theta)^{1/2-1}1_{(0,1)}(\theta)$$

(which is also known as the "arcsin" distribution because the corresponding distribution function is $\Lambda(\theta) = (2/\pi)\arcsin(\sqrt{\theta})$; it arises naturally as the limiting distribution of the proportion of time a random walk stays positive). Thus $\lambda_J((0,1)) = \Gamma(1/2)^2 = \pi$. The corresponding posterior distribution is proportional to

$$\theta^{x-1/2}(1-\theta)^{1-x-1/2} = \theta^{(x+1/2)-1}(1-\theta)^{(3/2-x)-1};$$

i.e. the posterior density is Beta$(x+1/2, 3/2-x)$ if $X = x$ is observed.

(b) When $X \sim$ Poisson$(\theta)$ with $P_\theta(X = x) = \theta^x e^{-\theta}/x!$ for $x = 0, 1, 2, \ldots$, the score function is

$$\dot{l}_\theta(x) = \frac{x}{\theta} - 1,$$

and the information for $\theta$ is $I(\theta) = 1/\theta$. Thus the Jeffrey's prior for $\theta$ has density $\lambda_J$ proportional to $\theta^{-1/2}$. Hence

$$\int_0^\infty \lambda_J(\theta)d\theta = \int_0^\infty \lambda^{-1/2}d\theta = \infty,$$

because the integral diverges at $\infty$. The posterior density is proportional to

$$p(x|\theta)\lambda_J(\theta) = \frac{\theta^x}{x!}\exp(-\theta)\cdot\theta^{-1/2} = \frac{\theta^{x-1/2}}{x!}\exp(-\theta),$$

which has a finite integral for every $x \in \{0, 1, 2, \ldots\}$, namely

$$\int_0^\infty \frac{\theta^{x-1/2}}{x!}\exp(-\theta)d\theta = \frac{\Gamma(x+1/2)}{x!}.$$

Thus the posterior density is Gamma $(x+1/2, 1)$.

(c) When $X \sim$ Geometric$(\theta)$ with $P_\theta(X = x) = (1-\theta)^{x-1}\theta$ for $x = 1, 2, \ldots$, the score function is

$$\dot{l}_\theta(x) = \frac{1}{1-\theta}\left(\frac{1}{\theta} - X\right)$$

and the information for $\theta$ is

$$I(\theta) = E_\theta(\dot{l}_\theta(X))^2 = -E_\theta\ddot{l}_{\theta\theta}(X) = \frac{1}{\theta^2(1-\theta)}.$$

Hence the Jeffrey's prior for $\theta$ has density

$$\lambda_J(\theta) = \frac{1}{\theta\sqrt{(1-\theta)}}.$$

Here we have

$$\int_0^1 \lambda_J(\theta)d\theta = \int_0^1 \frac{1}{\theta\sqrt{(1-\theta)}}d\theta = \infty$$

because the integral diverges at 0. Hence this density does not correspond to a finite measure. It corresponds in some sense to a Beta$(0, 1/2)$ density; but recall that the Beta$(\alpha, \beta)$ densities are defined for $\alpha, \beta > 0$. In spite of this the posterior density is proportional to

$$\theta^0(1-\theta)^{x-1-1/2} = (1-\theta)^{x-1/2-1}$$

which corresponds to Beta$(1, x - 1/2)$ if $X = x$ is observed. This is completely proper for any $x \in \{1, 2, \ldots\}$ since $x - 1/2 > 0$.

When $X_1, \ldots, X_n$ are i.i.d. Geometric$(\theta)$ so that $\sum_1^n X_i \sim$ Negative Binomial$(n, \theta)$, then the density of the data is proportional to $\theta^n(1-\theta)^{y-n}$ with $y \geq n$, and since $I_n(\theta) = n/[\theta^2(1-\theta)]$, the posterior is proportional to

$$\theta^{n-1}(1-\theta)^{y-n-1/2} = \theta^{n-1}(1-\theta)^{y-n+1/2-1}$$

which has finite mass for all $y \geq n$. Hence the posterior is proper for all $n \geq 1$.

(d) From Example 3.2.7 the information matrix for the Weibull density is

$$I(\theta) = \begin{pmatrix} \frac{\beta^2}{\alpha^2} & \frac{a}{\alpha} \\ \frac{a}{\alpha} & \frac{b^2}{\beta^2} \end{pmatrix}$$

where $a = -(1-\gamma)$ and $b^2 = \pi^2/6 + (1-\gamma)^2$. Therefore the Jeffrey's prior is given by

$$\lambda_J(\theta) = \det(I(\theta))^{1/2} = \frac{\pi/\sqrt{6}}{\alpha}\,.$$

This is not the density of a finite measure:

$$\int_0^\infty \int_0^\infty \lambda_J(\alpha, \beta) d\alpha d\beta = \frac{\pi}{\sqrt{6}} \int_0^\infty \int_0^\infty \frac{1}{\alpha} d\alpha d\beta = \infty\,.$$

The posterior density is proportional to

$$\begin{aligned} p_\theta(x)\lambda_J(\theta) &= \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} \exp(-(x/\alpha)^\beta) \frac{\pi/\sqrt{6}}{\alpha} 1_{(0,\infty)}(\alpha) 1_{(0,\infty)}(\beta) \\ &= \frac{\pi}{\sqrt{6}} \frac{\beta}{\alpha^2}\left(\frac{x}{\alpha}\right)^{\beta-1} \exp(-(x/\alpha)^\beta) 1_{(0,\infty)}(\alpha) 1_{(0,\infty)}(\beta)\,. \end{aligned}$$

When we try to compute the normalizing constant to form a density (namely the marginal density $p(x)$), we find, however, that (using the change of variables $u = \alpha^{-\beta}$ in the second line so that $du = -\beta\alpha^{-\beta-1}d\alpha$)

$$\begin{aligned} p(x) &= \frac{\pi}{\sqrt{6}} \int_0^\infty \int_0^\infty \frac{\beta}{\alpha^2}\left(\frac{x}{\alpha}\right)^{\beta-1} \exp(-(x/\alpha)^\beta) d\alpha d\beta \\ &= \frac{\pi}{\sqrt{6}} \int_0^\infty \frac{x^{\beta-1}}{x^\beta} \int_0^\infty x^\beta \exp(-x^\beta u) du d\beta \\ &= \frac{\pi}{\sqrt{6}} \int_0^\infty \frac{1}{x} d\beta = \infty\,. \end{aligned}$$

Hence the posterior (for one observation) is not proper for any $x$.

For $n$ observations we compute

$$p_\theta(\underline{x})\lambda_J(\theta) \;=\; \frac{\beta^n}{\alpha^n}\frac{(\prod x_i)^{\beta-1}}{\alpha^{n(\beta-1)}}\exp(-(\sum x_i^\beta/\alpha^\beta)\frac{\pi/\sqrt{6}}{\alpha}1_{(0,\infty)}(\alpha)1_{(0,\infty)}(\beta)\,.$$

Therefore, using the same change of variables as before, $u = \alpha^{-\beta}$,

$$\begin{aligned} p(\underline{x}) &= \frac{\pi}{\sqrt{6}}\int_0^\infty \beta^n(\prod x_i)^{\beta-1}\int_0^\infty \frac{1}{\alpha^{n+1}}\frac{1}{\alpha^{n(\beta-1)}}\exp(-\sum x_i^\beta/\alpha^\beta)d\alpha d\beta \\ &= \frac{\pi}{\sqrt{6}}\int_0^\infty \beta^{n-1}(\prod x_i)^{\beta-1}\int_0^\infty u^{n-1}\exp(-\sum x_i^\beta u)du d\beta \\ &= \frac{\pi}{\sqrt{6}}\Gamma(n)\int_0^\infty \beta^{n-1}\frac{(\prod x_i)^{\beta-1}}{(\sum x_i^\beta)^n}d\beta = \frac{\pi}{\sqrt{6}}\frac{\Gamma(n)}{\prod x_i}\int_0^\infty \beta^{n-1}\frac{(\prod x_i)^\beta}{(\sum x_i^\beta)^n}d\beta \end{aligned}$$

This converges if the $x_i$'s are not all equal. This can be seen by writing the last integral as follows:

$$\begin{aligned} \int_0^\infty \beta^{n-1}\frac{(\prod x_i)^\beta}{(\sum x_i^\beta)^n}d\beta &= \int_0^\infty \beta^{n-1}\frac{1}{\left(\frac{\sum_1^n x_i^\beta}{(\prod x_i^\beta)^{1/n}}\right)^n}d\beta \\ &= \int_0^\infty \beta^{n-1}\frac{1}{(\sum_1^n y_i^\beta)^n}d\beta \\ &= \int_0^1 \beta^{n-1}\frac{1}{(\sum_1^n y_i^\beta)^n}d\beta + \int_1^\infty \beta^{n-1}\frac{1}{(\sum_1^n y_i^\beta)^n}d\beta. \end{aligned}$$

where $y_i \equiv x_i/(\prod x_i)^{1/n}$, $i = 1, \ldots, n$. But for $\beta \geq 1$ we have

$$\left(\frac{1}{n}\sum_1^n y_i^\beta\right)^{1/\beta} \geq \frac{1}{n}\sum_1^n y_i,$$

by Liapunov's inequality, so

$$\sum_1^n y_i^\beta \geq n\left(\frac{1}{n}\sum_1^n y_i\right)^\beta$$

where $1 = (\prod_1^n y_i)^{1/n} \leq n^{-1}\sum_1^n y_i$ with strict inequality if the $y_i$'s are not all equal to 1, or, equivalently, if the $x_i$'s are not all equal. Thus the integrals in the last display are bounded above by

$$\begin{aligned} &\int_0^1 \beta^{n-1}\frac{1}{(n y_{(1)}^\beta)^n}d\beta + \int_1^\infty \beta^{n-1}\frac{1}{n^n\overline{y}^{n\beta}}d\beta \\ &= \int_0^1 \beta^{n-1}\frac{1}{(n y_{(1)}^\beta)^n}d\beta + \int_1^\infty \frac{\beta^{n-1}}{n^n}\exp(-n\beta\log(\overline{y}))d\beta \\ &< \infty \end{aligned}$$

where $\log(\overline{y}) > 0$ since $\overline{y} > 1$ if the $x_i$'s are not all equal. Thus the resulting posterior distribution is proper for $n \geq 2$ as long as all the observations are not all equal.

2. Continuation of problem 3, problem set 5: Suppose that $X_1, \ldots, X_n$ are i.i.d. Exponential($\theta$) (so the $X$'s have distribution $P_\theta$ 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 5 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 $X_1, \ldots, X_n$ are i.i.d. $P_{\theta_0} \equiv P$ with $\theta_0 \in (0, \infty)$)?

**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 $Y_i \sim \text{Gamma}(1, 1) = \text{Exp}(1)$, $i = 1, \ldots, n$ are all independent. Thus conditionally on the $X_i$'s we have, with $Z \sim N(0, 1)$ and with $\theta_0$ the true value of $\theta$,

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

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 $1/\overline{X}_n$.
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, $\sqrt{n}(\overline{Y}_n - 1)$: since $n\overline{Y}_n = \sum_1^n Y_i \sim \text{Gamma}(n, 1)$,

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

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

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

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

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

Since $\overline{X}_n + n^{-1}\beta \to_{a.s.} \theta_0^{-1}$ and $\sqrt{n}(Y_0 - \alpha)\frac{1/n}{\overline{X}_n + n^{-1}\beta} \to_{a.s.} 0$, 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

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

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

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

3. 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 $\Sigma$?

**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 $p_i \neq 0$ and $p_j \neq 0$,

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

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 $\mu_i$ to $\mu_j$. When we consider all the $k-1$ mean vectors $\mu_j$ 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 $\mu_i$ that is closest to $X$. When the identity covariance matrix $I$ is replaced by an arbitrary nonsingular covariance matrix $\Sigma$, this remains true with ordinary Euclidean distance replaced by $d_\Sigma^2(x, \mu) \equiv (x-\mu_i)^T \Sigma^{-1}(x-\mu_i)$.

4. Lehmann and Casella, TPE, Problem 5.17, page 293. (Also note Problems 5.18, 5.19, 5.20, page 293.)

**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 $\log\left(\int f^\alpha g^{1-\alpha} d\mu\right) \leq 0$, 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 $\alpha \to 1$, 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

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

5. **Optional bonus problem 1:** (Birgé). Let $X = (X_0, X_1, \ldots, X_k)$ 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 $\theta'$ denote the projection of $\theta$ onto the $k-$dimensional linear space spanned by the $k-$last coordinates. Consider the subset $\Theta_0$ of $\Theta = \mathbb{R}^{k+1}$ given by

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

(a) Show that the MLE of $\theta$ over $\Theta_0$ is given by $\widehat{\theta}_0 = 0$ and $\widehat{\theta'} = 2X'/\|X\|$ *on the event*

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

(b) Show that $P_\theta(\Omega_0) \ge 3/4$ for all $\theta \in \Theta_0$.

(c) Let $\widetilde{\theta} = (X_0, \underline{0})$. Show that for $k \ge 128$ we have

$$\sup_{\theta\in\Theta_0} E_\theta\|\theta - \hat{\theta}\|^2 \ge (3/4)\sqrt{k} + 3, \text{ and}$$
$$\sup_{\theta\in\Theta_0} E_\theta\|\theta - \widetilde{\theta}\|^2 \le 5.$$

Hint: A non-central $\chi^2_k$ distribution is stochastically larger than a central $\chi^2_k$ distribution; then use Lemma 1 of Laurent and Massart (2000).

6. **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,

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

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

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

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,

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

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.)