Statistics 582, Problem Set 7 Solutions

Wellner; 2/22/2018

1. Let $\underline{X} \sim N_k(\underline{\theta}, I)$ and suppose that $\underline{\theta} \sim N_k(\underline{\tau}, \Sigma)$ where Σ is non-singular. Find the posterior distribution of $\underline{\theta}$. Argue directly to show that the Bernstein-von Mises theorem holds in this case.

Solution: In sections 5.4 and 5.6 we showed that if $X_1, \ldots, X_n$ are i.i.d. $N(\theta, \sigma^2)$ and $\theta \sim N(\mu, \tau^2)$, then the posterior distribution of θ given $\overline{X}_n$ is Normal with

$$\begin{aligned} E(\theta|\underline{X}) &= \frac{1/\tau^2}{1/\tau^2 + n/\sigma^2}\mu + \frac{n/\sigma^2}{1/\tau^2 + n/\sigma^2}\overline{X}_n \\ Var(\theta|\underline{X}) &= \frac{1}{1/\tau^2 + n/\sigma^2}. \end{aligned}$$

Similarly, if $\underline{X}_1, \ldots, \underline{X}_n$ are i.i.d. $N_k(\theta, I)$ and $\theta \sim N_k(\tau, \Sigma)$, then the posterior distribution of θ given $\underline{X}_1, \ldots, \underline{X}_n$ is $k-$variate normal with

$$\begin{aligned} E(\theta|\underline{X}_1, \ldots, \underline{X}_n) &= (nI + \Sigma^{-1})^{-1}(\Sigma^{-1}\tau + n\overline{\underline{X}}_n), \\ Cov(\theta|\underline{X}_1, \ldots, \underline{X}_n) &= (nI + \Sigma^{-1})^{-1}. \end{aligned}$$

Thus we find that

$$\begin{aligned} &\sqrt{n}\,(E(\theta|\underline{X}) - \theta_0) \\ &= \sqrt{n}(nI + \Sigma^{-1})^{-1}(\Sigma^{-1}\tau + n\overline{\underline{X}}_n) - \sqrt{n}\theta_0 \\ &= \sqrt{n} \cdot n^{-1}(I + n^{-1}\Sigma^{-1})^{-1}\Sigma^{-1}\tau + (nI + \Sigma^{-1})^{-1}\sqrt{n}n\overline{\underline{X}}_n - (nI + \Sigma^{-1})^{-1}(nI + \Sigma^{-1})\sqrt{n}\theta_0 \\ &= \sqrt{n} \cdot n^{-1}(I + n^{-1}\Sigma^{-1})^{-1}\Sigma^{-1}(\tau - \theta_0) + (I + n^{-1}\Sigma^{-1})^{-1}\sqrt{n}(\overline{\underline{X}}_n - \theta_0) \\ &\to_d \; 0 + 1 \cdot N_k(0, I). \end{aligned}$$

On the other hand

$$Cov(\sqrt{n}(\theta - \theta_0)) = n(nI + \Sigma^{-1})^{-1} = (I + n^{-1}\Sigma^{-1})^{-1} \to I,$$

so it follows that

$$\mathcal{L}\left(\sqrt{n}(\theta - \theta_0 - (\overline{X}_n - \theta_0))\Big|\underline{X}\right) \to_d N_k(0, I).$$

Thus the Bernstein - von Mises theorem holds, at least in the sense of convergence in distribution of the centered and rescaled posterior distributions.

2. Suppose that $X \sim \text{Poisson}(\theta_0)$ for some $\theta_0 \in \Theta \equiv (0, \infty)$, and suppose that the prior distribution Λ of θ is absolutely continuous with a continuous positive density at θ_0. Verify the other hypotheses of van der Vaart's Bernstein- von Mises theorem 10.1, page 141. What does the theorem say in this case?

Solution: Since the Poisson family of densities is an exponential family (by writing

$$p_\theta(x) = e^{-\theta}\theta^x/x! = \exp(x \log \theta - \theta)/x!\,, x \in \{0, 1, 2, \ldots\}$$

with respect to counting measure, it is regular and differentiable in quadratic mean. By standard computations $\log p_\theta(x) = x \log \theta - \theta + \text{constant}$, so $\dot{l}_\theta(x) = x/\theta - 1$ and $I(\theta) = E_\theta(X - \theta)^2/\theta^2 = 1/\theta$. Thus $I(\theta_0) > 0$ for $\theta_0 \in (0, \infty)$.

To complete the verification of the hypotheses of van der Vaart's Bernstein-von Mises theorem we need to find a sequence of test statistics $\phi_n \equiv \phi_n(X_1, \ldots, X_n)$ satisfying the following: for every $\epsilon > 0$

$$P^n_{\theta_0}\phi_n \to 0 \quad \text{and} \quad \sup_{|\theta - \theta_0| \geq \epsilon} P^n_{\theta_0}(1 - \phi_n) \to 0. \tag{1}$$

Fix $\epsilon > 0$, let $T_n = \overline{X}_n$, and set $\phi_n = 1\{|\overline{X}_n - \theta_0| \geq \epsilon/2\}$. First, by Chebychev's inequality,

$$\begin{aligned} P^n_{\theta_0}\phi_n &= P^n_{\theta_0}(|T_n - \theta_0| \geq \epsilon/2) = P^n_{\theta_0}(|\overline{X}_n - \theta_0| \geq \epsilon/2) \\ &\leq \frac{Var_{\theta_0}(\overline{X}_n)}{\epsilon^2/4} = \frac{4\theta_0}{n\epsilon^2} \to 0. \end{aligned}$$

Also note that by the exponential bounds below we also have

$$\begin{aligned} P^n_{\theta_0}\phi_n &= P_{\theta_0}(|T_n - \theta_0| \geq \epsilon/2) = P^n_{\theta_0}(|n\overline{X}_n - n\theta_0| \geq n\epsilon/2) \\ &\leq 2\exp\left(-\frac{n\epsilon^2}{8\theta_0}\psi\left(\frac{\epsilon}{2\theta_0}\right)\right) \to 0 \end{aligned}$$

geometrically quickly. Now we verify that the second convergence in (1) holds. Writing P_θ for P^n_θ on the right side,

$$\begin{aligned} P^n_\theta(1 - \phi_n) &= P_\theta(|T_n - \theta_0| < \epsilon/2) = P_\theta(\theta_0 - \epsilon/2 < T_n < \theta_0 + \epsilon/2) \\ &\leq \min\{P_\theta(T_n > \theta_0 - \epsilon/2), P_\theta(T_n < \theta_0 + \epsilon/2)\} \end{aligned}$$

where each of the two terms in the second line of the last display is a monotone function of θ in view of the Karlin - Rubin theorem and the fact that the Poisson family of distributions has monotone likelihood ratio. The first term is increasing

(and the relevant set of θ's is $\theta \leq \theta_0 - \epsilon$), while the second term is decreasing (and the relevant set of θ's is $\theta \geq \theta_0 + \epsilon$). Thus it follows that

$$\begin{aligned}
&\sup_{|\theta-\theta_0|\geq\epsilon} P_\theta^n(1-\phi_n) \\
&\leq \min\left\{\sup_{\theta\leq\theta_0-\epsilon} P_\theta(T_n > \theta_0 - \epsilon/2),\ \sup_{\theta\geq\theta_0+\epsilon} P_\theta(T_n < \theta_0 + \epsilon/2)\right\} \\
&= \min\{P_{\theta_0-\epsilon}(T_n > \theta_0 - \epsilon/2),\ P_{\theta_0+\epsilon}(T_n < \theta_0 + \epsilon/2)\} \\
&\leq \min\left\{P_{\theta_0-\epsilon}(T_n - (\theta_0 - \epsilon) > \epsilon/2), P_{\theta_0+\epsilon}(-(T_n - (\theta_0 + \epsilon)) > \epsilon/2)\right\} \qquad (2)
\end{aligned}$$

where $nT_n = \sum_1^n X_i \sim \text{Poisson}(n\theta)$. But if $Y \sim \text{Poisson}(\nu)$, then

$$P(\pm(Y-\nu) > t) \leq \exp(-\nu h(1 \pm t/\nu)) = \exp\left(-\frac{t^2}{2\nu}\psi(\pm t/\nu)\right) \qquad (3)$$

where $h(v) \equiv v(\log v - 1) + 1$ and $\psi(v) \equiv 2v^{-2}h(1+v)$. Therefore, it follows that for $0 < \epsilon < \theta_0$ we have,

$$\begin{aligned}
P_{\theta_0-\epsilon}(T_n - (\theta_0 - \epsilon) > \epsilon/2) &= P_{\theta_0-\epsilon}(nT_n - n(\theta_0 - \epsilon) > n\epsilon/2) \\
&\leq \exp\left(-\frac{n\epsilon^2}{8(\theta_0-\epsilon)}\psi\left(\frac{\epsilon}{2(\theta_0-\epsilon)}\right)\right) \\
&\to 0 \quad \text{as} \quad n \to \infty.
\end{aligned}$$

Similarly, but more easily

$$\begin{aligned}
P_{\theta_0+\epsilon}(-(T_n - (\theta_0 + \epsilon) > \epsilon/2) &\leq \exp\left(-\frac{n\epsilon^2}{8(\theta_0+\epsilon)}\psi\left(\frac{\epsilon}{2(\theta_0+\epsilon)}\right)\right) \\
&\to 0 \quad \text{as} \quad n \to \infty.
\end{aligned}$$

Thus (1) holds and van der Vaart's Bernstein-von Mises theorem holds if the prior density λ is continuous in a neighborhood of θ_0 with $\lambda(\theta_0) > 0$.

We conclude that if Λ^* denotes the posterior distribution of

$$\sqrt{n}\left\{(\boldsymbol{\theta} - \theta_0) - (\overline{X}_n - \theta_0)\right\} = \sqrt{n}\left(\boldsymbol{\theta} - \overline{X}_n\right),$$

then under $P_{\theta_0}^n$,

$$d_{TV}\left(\Lambda^*(t|X_1, \ldots X_n), \Phi(t/\sqrt{\theta_0})\right) \to_p 0.$$

It remains to show that the exponential bounds (3) hold. First, for $r > 0$ and $t > 0$ we have

$$\begin{aligned}
P(Y - \nu \geq t) &= P(\exp(r(Y-\nu)) \geq \exp(rt) \leq \frac{Ee^{r(Y-\nu)}}{e^{rt}} \quad \text{by Markov's inequality} \\
&= e^{-r(\nu+t)}Ee^{r^Y} = \exp(\nu(e^r - 1 - r) - rt) \equiv \exp(-H(r;\nu,t))
\end{aligned}$$

since $Ee^{rY} = \exp(\nu(e^r - 1))$. Since this holds for all $r > 0$, we may minimize the bound with respect to r. Equivalently, we may maximize the quantity $H(r) \equiv H(r; \nu, t)$ in the exponential with respect to r. Now

$$H'(r) = t - \nu(e^r - 1) = 0$$

if $e^r = 1 + t/\nu$, or $r \equiv r_0 \equiv \log(1 + t/\nu)$. This yields a maximum since $H''(r) = -\nu e^r < 0$. Thus the bound becomes

$$P(Y - \nu \geq t) \leq e^{-H(r_0)} = \exp(-\nu h(1 + t/\nu)) = \exp\left(-\frac{t^2}{2\nu}\psi(t/\nu)\right)$$

where the second form of the bound follows immediately from the definition of ψ. The proof of the bound for $P(-(Y - \nu) \geq t)$ follows by a similar argument.

Simpler solution for the second convergence in (1): Note that from (2) it follows that

$$\begin{aligned}
\sup_{|\theta - \theta_0| \geq \epsilon} P_\theta^n(1 - \phi_n) &\leq P_{\theta_0 - \epsilon}(T_n - (\theta_0 - \epsilon) > \epsilon/2) + P_{\theta_0 + \epsilon}(-(T_n - (\theta_0 + \epsilon)) > \epsilon/2) \\
&\leq \frac{Var_{\theta_0 - \epsilon/2}(T_n)}{\epsilon^2/4} + \frac{Var_{\theta_0 + \epsilon/2}(T_n)}{\epsilon^2/4} \\
&= \frac{\theta_0 - \epsilon/2}{n\epsilon^2/4} + \frac{\theta_0 + \epsilon/2}{n\epsilon^2/4} = \frac{4\theta_0}{n\epsilon^2} \\
&\to 0
\end{aligned}$$

as $n \to \infty$. While this simple argument gives the desired uniform convergence to 0 of the type 2 error, it does not capture the true speed of convergence, which is, in fact, exponentially fast rather than just $O(1/n)$ as obtained via the simple Chebychev argument.

3. 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($\underline{\alpha}$):

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

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: (a) If $\underline{\theta} \sim$ Dirichlet($\underline{\alpha}$) then $\theta_j \sim$ Beta($\alpha_j, \sum_{j' \neq j} \alpha_{j'}$), 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 $\underline{\theta}$ is Dirichlet($\underline{\alpha} + \underline{X}$), the posterior mean is

$$d_\Lambda(\underline{X}) = E(\underline{\theta}|\underline{X}) = (\underline{\alpha} + \underline{X})/(\sum_i \alpha_i + n).$$

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

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

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

$$\frac{\alpha_j + X_j - 1}{\theta_j} + c = 0, \qquad j = 1, \ldots, k. \tag{4}$$

and

$$\sum_{j=1}^k \theta_j = 1. \tag{5}$$

The first equation yields

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

substitution of this into (5) yields

$$1 = \sum_{j=1}^k \theta_j^{mode} = \frac{1}{-c} \left\{ \sum_{j=1}^k \alpha_j + n - k \right\},$$

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 $\hat{\underline{\theta}} = \underline{X}/n$. Note that $\underline{\alpha} = \underline{1}$ yields a uniform prior over $\underline{\theta}$.
(c) As shown in class, 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 $\alpha_1 = \cdots = \alpha_k = \alpha$, this yields the Bayes estimator

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

Note that $d_{\Lambda,i}(\underline{X}) = (\alpha + X_i)/(k\alpha + n)$ has

$$\begin{aligned} Var_{\underline{\theta}}(d_{\Lambda,i}(X) &= \frac{n\theta_i(1-\theta_i)}{(k\alpha+n)^2}, \\ E_{\underline{\theta}}(d_{\Lambda,i}(X)) &= \frac{\alpha + n\theta_i}{k\alpha + n}, \\ \text{bias}_{\underline{\theta}}(d_{\Lambda,i}(X)) &= \frac{\alpha - k\alpha\theta_i}{k\alpha + n}. \end{aligned}$$

Thus the risk is

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

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

$$\begin{aligned} d_\Lambda(\underline{X}) &= \frac{\sqrt{n}}{\sqrt{n}+n}\frac{1}{k} + \frac{n}{\sqrt{n}+n}\frac{\underline{X}}{n} \\ &= (1-\lambda_n)\frac{\underline{1}}{k} + \lambda_n \underline{\hat{p}}_n \end{aligned}$$

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

4. Find the limit distribution of the minimax estimator d_M in problem 3 (i.e. $\sqrt{n}(d_M(X_n) - p) \to_d$ "something" and find "something"). Is d_M a regular estimator of p?

Solution: Note that $\sqrt{n}(1-\lambda_n) = \lambda_n \to 1$. Hence

$$\begin{aligned}
\sqrt{n}(d_M(\underline{X}_n) - \underline{\theta}) &= \sqrt{n}\{\lambda_n \underline{\hat{p}}_n + (1-\lambda_n)\frac{\underline{1}}{k} - (\lambda_n + 1 - \lambda_n)\underline{\theta}\} \\
&= \lambda_n \sqrt{n}(\underline{\hat{\theta}}_n - \underline{\theta}) + \sqrt{n}(1-\lambda_n)(\frac{\underline{1}}{k} - \underline{\theta}) \\
&\to_d N_k(0, \Sigma) + \frac{\underline{1}}{k} - \underline{\theta} \\
&= N_k\left(\frac{\underline{1}}{k} - \underline{\theta}, \Sigma\right)
\end{aligned}$$

where $\Sigma = \text{diag}(\underline{\theta}) - \underline{\theta}\underline{\theta}^T$. To see that $d_M(\underline{X}_n)$ is a regular estimator of θ, let $\theta_n = \theta_0 + tn^{-1/2}$ where $1't = 0$ (so that $1'\theta_n = 1$). Then since $\hat{p}_n$ is a regular estimator of θ with

$$\sqrt{n}(\hat{p}_n - \theta_n) \to_d Z \sim N_k(0, \text{diag}(\theta_0) - \theta_0\theta_0')$$

under P_{θ_n} (which follows from the Liapunov CLT together with the Cram'er-Wold device, or from contiguity theory), it follows that

$$\begin{aligned}
\sqrt{n}(d_M(\underline{X}_n) - \theta_n) &= \sqrt{n}((1-\lambda_n)(1/k) + \lambda_n \hat{p}_n - \theta_n) \\
&= \lambda_n \sqrt{n}(\hat{p}_n - \theta_n) + \sqrt{n}(1-\lambda_n)(1/k - \theta_n) \\
&\to_d 1 \cdot Z + 1 \cdot (1/k - \theta_0) \\
&\sim N_k((1/k - \theta_0), \text{diag}(\theta_0) - \theta_0\theta_0'),
\end{aligned}$$

where we used $\sqrt{n}(1-\lambda_n) = \lambda_n \to 1$ and $\theta_n \to \theta_0$. Since this limiting distribution does not depend on t, $d_M(\underline{X}_n)$ is regular.