Statistics 583, Problem Set 2 Solutions

Wellner; 4/14/2016

1. Van der Vaart (1998), problem 25, page 84.
(i) Verify the conditions of Wald's theorem for m_θ the log-likelihood function of the $N(\mu, \sigma^2)$ distribution if the parameter space for $\theta = (\mu, \sigma^2)$ is a compact subset of $\mathbb{R} \times \mathbb{R}^+$.
(ii) Extend m_θ by continuity to the compactification of $\mathbb{R} \times \mathbb{R}^+$. Show that the conditions of Wald's theorem fail at the points $(\mu, 0)$.
(iii) Replace m_θ by the log-likelihood function of a pair of two independent observations from the $N(\mu, \sigma^2)$ distribution, say (X, Y). Show that Wald's theorem now does apply, also with a compactified parameter set.

Solution: In van der Vaart's notation, $m_\theta(x) = \log p(x, \theta)$. I will work with $f(x, \theta) \equiv \log p(x, \theta) - \log p(x, \theta_0)$ as we discussed in class.
(i) If $p(x, \theta) = \phi((x - \mu)/\sigma)/\sigma$ where $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty) \equiv \Theta$, and Θ_c is a compact subset of Θ, then

$$f(x, \theta) = \log\left(\frac{\sigma_0}{\sigma}\right) - \frac{(x - \mu)^2}{2\sigma^2} + \frac{(x - \mu_0)^2}{2\sigma_0^2}$$

is a continuous function of (μ, σ^2), and hence achieves its supremum at some point $(\mu_*, \sigma_*^2) \in \Theta_c$. But since $f(x, (\mu, \sigma^2)) \le f(x, (x, \sigma^2))$ and since $f(x, (x, \sigma^2))$ is bounded by $f(x, (x, \sigma^2_{c_*}))$ where $c_* \equiv \inf\{c : \ (\mu, c) \in \Theta_c\}$, The function $f(x, (x, \sigma_c^2)) = \log\left(\frac{\sigma_0}{\sigma_c}\right) + (x - \mu_0)^2/\sigma_0^2 \equiv F(x)$ is an integrable envelope for the class of functions $\{f(\cdot, \theta) : \ \theta \in \Theta_c)$. The other conditions of Wald's theorem clearly hold in this case, and the Theorem applies to the smaller model in which Θ is replaced by Θ_c.
(ii) If $\Theta = \mathbb{R} \times (0, \infty)$ is compactified by identifying it with $[-\pi/2, \pi/2] \times [0, \pi/2] \equiv \tilde{\Theta}$ by way of the transformation $\nu \equiv \arctan\mu$ and $\tilde{\sigma}^2 = \arctan(\sigma^2)$, then points of the form $(\mu, \sigma^2) = (\mu, 0)$ correspond to points $(\nu, \tilde{\sigma}^2) = (\arctan(\mu), 0)$, and we see that

$$\begin{aligned} \sup_{\tilde{\theta} \in \tilde{\Theta}} f(x, \tilde{\theta}) &\ge \sup_{\mu \in \mathbb{R},\ \sigma^2 > 0} f(x, (\mu, \sigma^2)) \\ &= \sup_{\sigma^2 > 0} f(x, (x, \sigma^2)) = \sup_{\sigma^2 > 0} \log\left(\frac{\sigma_0}{\sigma}\right) + \frac{(x - \mu_0)^2}{2\sigma_0^2} = +\infty. \end{aligned}$$

Thus there is no integrable envelope function in this case, and Wald's theorem fails via this approach.

(iii) If we replace $f(\cdot, \theta)$ in (i) and (ii) by the corresponding log density for two (independent) observations, then we find that

$$\begin{aligned}
f((x,y),\theta) &= \log p((x,y),\theta) - \log p((x,y),\theta_0) \\
&= \log\left(\frac{\sigma_0^2}{\sigma^2}\right) - \frac{(x-\mu)^2+(y-\mu)^2}{2\sigma^2} + \frac{(x-\mu_0)^2+(y-\mu_0)^2}{2\sigma_0^2} \\
&\le f((x,y),((x+y)/2,\sigma^2) \text{ by maximizing with respect to } \mu \\
&= \log\left(\frac{\sigma_0^2}{\sigma^2}\right) - \frac{2(x-y)^2/4}{2\sigma^2} + \frac{(x-\mu_0)^2+(y-\mu_0)^2}{2\sigma_0^2} \\
&\le f((x,y),((x+y)/2,((x-y)/2)^2)) \text{ by maximizing with respect to } \sigma^2, \\
&= \log\frac{\sigma_0^2}{((x-y)/2)^2} - 1 + \frac{(x-\mu_0)^2+(y-\mu_0)^2}{2\sigma_0^2} \\
&\equiv F(x,y)
\end{aligned}$$

where $E_{\theta_0}(-1 + \{(X-\mu_0)^2 + (Y-\mu_0)^2\}/(2\sigma_0^2)) = 0$.
Moreover, since $X - Y \sim N(0, 2\sigma_0^2)$, it follows that $(X-Y)/2 \sim N(0, \sigma_0^2/2)$ and hence that $((X-Y)/2)^2/(\sigma_0^2/2) \sim \chi_1^2$. Thus we find that

$$E_{\theta_0} F(X,Y) = E_{\theta_0} \log\left(\frac{2}{\chi_1^2}\right) = \log(2) + E\log\frac{1}{\chi_1^2} < \infty$$

since

$$E\log\frac{1}{\chi_1^2} = \int_0^\infty \log(1/v)\frac{v^{1/2-1}}{2\Gamma(1/2)}e^{-v/2}dv = \gamma + \log(2) < \infty$$

where $\gamma \approx .577216...$ is Euler's gamma. Thus for pairs of independent observations we have an integrable envelope function and Wald's theorem can be applied "blockwise" as discussed in van der Vaart (1998), pages 48-49.

2. Consider the symmetric location family based on a fixed density f_0 symmetric at 0 defined as follows:

$$\mathcal{P} = \{p_\theta : \ p_\theta(x) = f_0(x-\theta), \quad \theta \in \mathbb{R}\}.$$

Particular cases of interest include:
(a) $f_0 = \phi$, the standard normal density (magenta);
(b) $f_0(x) = 2^{-1}\exp(-|x|)$, the Laplace density (blue);
(c) $f_0(x) = \pi^{-1}(1+x^2)^{-1}$, the standard Cauchy density (red);
(d) $f_0(x) = e^{-x}(1+e^{-x})^{-2}$, the logistic density (green).
(i) For each of these densities compute and plot f_0, $\varphi \equiv -\log f_0$, and the corresponding score functions for location $-f_0'/f_0$. A density f is *log-concave* if

$\varphi \equiv -\log f$ is a convex function. Which of the densities (a) - (d) are log-concave?
(ii) Suppose that $X_1, \ldots, X_n$ are i.i.d. P_0. In which of these cases is the MLE explicitly calculable? In which cases is it unique?
(iii) Suppose that $P_0 = P_{\theta_0} \in \mathcal{P}$. In which cases is the MLE consistent?
(iv) Suppose that $P_0 \notin \mathcal{P}$. For which case or cases is it possible that

$$\Theta_0 \equiv \{\theta_0 \in \Theta : \ P_0 m_{\theta_0} = \sup_{\theta \in \Theta} P_0 m_\theta\}$$

contains more than one point? [Hint: see Freedman and Diaconis (1982), *Ann. Statist.* **10**, 454-461, especially page 455.]

Solution: (i) Plots of the densities are shown in Figure 1; plots of the functions $\varphi \equiv -\log f$ are shown in Figure 2; plots of the score functions for location $-f'/f = \varphi'$ are shown in Figure 3; and the second derivative functions φ'' are shown in Figure 4.

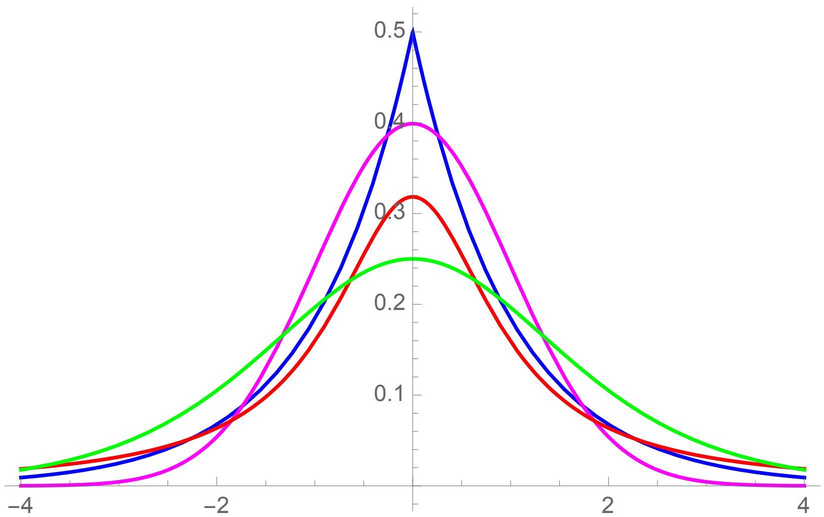

Figure 1: The four densities: Gaussian, Laplace, Cauchy, and Logistic).

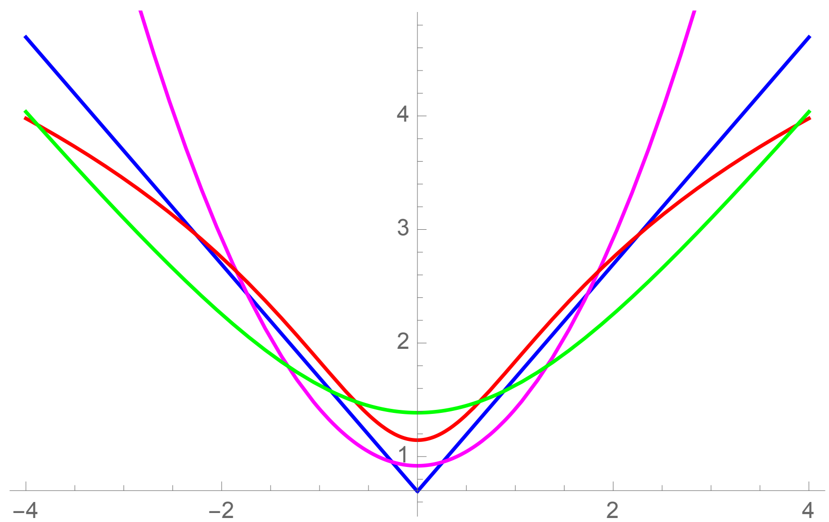

Figure 2: $\varphi \equiv -\log f$ for four densities: Gaussian, Laplace, Cauchy, and Logistic).

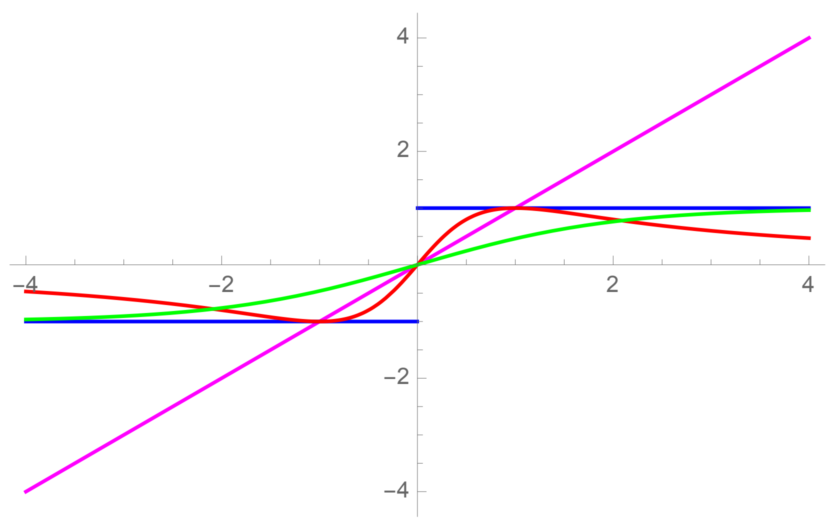

Figure 3: $\varphi' \equiv (-\log f)'$ for four densities: Gaussian, Laplace, Cauchy, and Logistic).

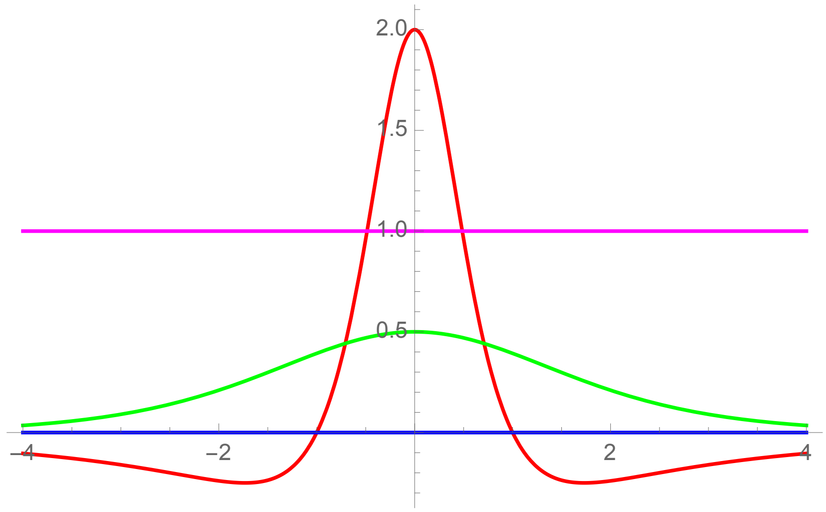

Figure 4: $\varphi'' \equiv (-\log f)''$ for four densities: Gaussian, Laplace, Cauchy, and Logistic).

The functions φ are convex in cases (a), (b), and (d), but not convex in (c): note that φ' is monotone non-decreasing in these first three cases, but not monotone in case (c). Furthermore the second derivatives are all non-negative functions in cases (a), (b), and (d), but in case (c) the second derivative function takes on negative values when $|x| > 1$.
(ii) The MLE in case (a) is, of course, $\hat{\theta} = \overline{X}_n$, and is unique. The MLE in case (b) is any sample median, e.g. $\mathbb{F}_n^{-1}(1/2)$ which is unique when the sample size is odd, but is not unique when the sample size is even: one common convention is to take it to be $(X_{(m)} + X_{(m+1)})/2$ where $2m = n$ and $X_{(1)} \le \cdots \le X_{(n)}$ are the order statistics. In case (c), the likelihood equation for the Cauchy family may have multiple solutions, but the MLE is unique with probability one. In case (d) the likelihood equation has a unique solution and yields the unique maximum of the likelihood function.
(iii) The MLE is consistent in all four cases when $P_0 \in \mathcal{P}$. This follows from either direct arguments in cases (a) and (b), and from Wald's theorem in cases (c) and (d).
(iv) When φ is not convex (i.e. $\varphi'' \ge 0$ fails to hold), then it is often possible to construct a symmetric distribution with (symmetric) density p_0 such that the population version of the function to be minimized has a local *maximum* at 0 (rather than a local minimum as expected), and equal local minima at two (or more) points away from 0. Thus the set

$$\Theta_0 \equiv \text{argmin}_\theta P_0(-\log f_0(X-\theta)) = \text{argmin}_\theta P_0(\varphi(X-\theta)) \supset \{\pm\theta^*\}$$

and the estimator $\widehat{\theta}_n = \text{argmin}_\theta \mathbb{P}_n \varphi(X-\theta)$ oscillates between $+\theta^*$ and $-\theta^*$. This is what Freedman and Diaconis (1981) do explictly in the case when f_0 is the

Cauchy density as well as two other proposed f_0's and corresponding φ's with φ not convex. [My φ is their M. Their terminology concerning unimodality and convexity differs slightly from current standard usage. It would be useful to update their paper using standard terminology.]

3. Suppose that $X_1, \ldots, X_n$ are i.i.d. P on $\mathbb{R}$ with $E|X_1| = \int |x| dP(x) < \infty$. Consider the following measure of dispersion:

$$D_n \equiv n^{-1} \sum_{i=1}^{n} |X_i - \overline{X}_n| = \mathbb{P}_n |X - \overline{X}_n|.$$

The corresponding measure in the population is $d \equiv P|X - PX| = \int |x - \mu(P)| dP(x)$ where $\mu(P) = P(X) = \int x dP(x)$. Use a Glivenko-Cantelli theorem to show that $D_n \rightarrow_p d$. [Hint: Let $\delta > 0$. Since $Pr(|\overline{X}_n - \mu| \le \delta) \rightarrow 1$ as $n \rightarrow \infty$, it becomes useful to consider the class of functions $\mathcal{F}_\delta \equiv \{f(x,t) \equiv |x - t| : |t - \mu| \le \delta\}$.]

Solution: Let $\delta > 0$. Note that $D_n = \mathbb{P}_n f(X, \overline{X}_n)$ and $d = P|X - \mu| = Pf(X, \mu)$ where $f(x,t) \equiv |x - t|$. Then we can write

$$\begin{aligned} |D_n - d| &= |\mathbb{P}_n f(X, \overline{X}_n) - Pf(X, \overline{X}_n) + Pf(X, \overline{X}_n) - Pf(X, \mu)| \\ &\le |(\mathbb{P}_n - P) f(X, \overline{X}_n)| + |Pf(X, \overline{X}_n) - Pf(X, \mu)|. \end{aligned} \tag{1}$$

First note that $|f(x,t) - f(x,s)| \le |t - s|$ for all $x \in \mathbb{R}$ and $s, t \in \mathbb{R}$, so the second term in the last display is bounded by $|\overline{X}_n - \mu| \rightarrow_{a.s.} 0$ by the SLLN. Since $G_n \equiv \{\overline{X}_n \in [\mu - \delta, \mu + \delta]\}$ has probability converging to 1 as $n \rightarrow \infty$, we see that the first term in the last display is bounded, on the event G_n, by

$$\sup_{t \in [\mu - \delta, \mu + \delta]} |(\mathbb{P}_n - P) f(X, t)| = \sup_{f \in \mathcal{F}_\delta} |\mathbb{P}_n(f) - P(f)| \equiv \|\mathbb{P}_n - P\|_{\mathcal{F}_\delta} \rightarrow_{a.s.} 0$$

if $\mathcal{F}_\delta$ is a $P-$Glivenko-Cantelli class of functions. But this follows from our first Wald type Glivenko-Cantelli theorem since: (i) we can take $\Theta \equiv [\mu - \delta, \mu + \delta]$ which is compact; (ii) the functions $t \mapsto f(x,t)$ are continuous in t for each $x \in \mathbb{R}$; and (iii) the functions $f(x,t)$ satisfy $|f(x,t)| \le F(x,t) \equiv \max\{f(x, \mu + \delta), f(x, \mu - \delta)\}$ where we note that $E|X_1 - (\mu - \delta)| \le E|X_1| + |\mu - \delta| < \infty$ and $E|X_1 - (\mu + \delta)| \le E|X_1| + (\mu + \delta) < \infty$ by our assumption that $E|X_1| < \infty$. Thus we conclude that the right side of (1) converges in probability to 0 and hence $D_n \rightarrow_p d$.

4. Suppose that P_j and Q_j have densities p_j and q_j respectively with respect to μ_j for $j = 1, \ldots, n$, and suppose that X_j has distribution either P_j or Q_j for each $j \le n$. Let $\prod_{j=1}^n P_j$ and $\prod_{j=1}^n Q_j$ denote the product measures.
(i) Show that $\rho(\prod_{j=1}^n P_j, \prod_{j=1}^n Q_j) = \prod_{j=1}^n \rho(P_j, Q_j)$.
(ii) Use the result of (i) to find a formula for $H^2(\prod_{j=1}^n P_j, \prod_{j=1}^n Q_j)$ in terms of

$H^2(P_j, Q_j)$, $j = 1, \ldots, n$.
(iii) Specialize the results of (i) and (ii) to the setting in which $P_1 = \cdots = P_n \equiv P$ and $Q_1 = \cdots = Q_n \equiv Q$ where we now write $\prod_{j=1}^n P_j = P^n$ and $\prod_{j=1}^n Q_j = Q^n$.
(iv) Suppose that $\rho(P, Q) < 1$. Using (iii) what can you say about $\lim_n \rho(P^n, Q^n)$ and $\lim_n H(P^n, Q^n)$?

Solution: (i) Let $p_j \equiv dP_j/d\mu_j$ and $q_j \equiv dQ_j/d\mu_j$ where μ_j is a dominating measure for P_j and Q_j for $j = 1, \ldots, n$. Then

$$\begin{aligned}
\rho\left(\prod_{j=1}^n P_j, \prod_{j=1}^n Q_j\right) &= \int \cdots \int \sqrt{\prod_{j=1}^n p_j(x_j) \prod_{j=1}^n q_j(x_j)} d\mu_1(x_1) \cdots d\mu(x_n) \\
&= \int \cdots \int \prod_{j=1}^n \sqrt{p_j(x_j) q_j(x_j)} d\mu_1(x_1) \cdots d\mu_n(x_n) \\
&= \prod_{j=1}^n \int \sqrt{p_j q_j} d\mu_j = \prod_{j=1}^n \rho(P_j, Q_j).
\end{aligned}$$

(ii) Note that

$$\begin{aligned}
H^2\left(\prod_{j=1}^n P_j, \prod_{j=1}^n Q_j\right) &= 1 - \rho\left(\prod_{j=1}^n P_j, \prod_{j=1}^n Q_j\right) \\
&= 1 - \prod_{j=1}^n \rho(P_j, Q_j) = 1 - \prod_{j=1}^n (1 - H^2(P_j, Q_j)).
\end{aligned}$$

(iii) When $P_1 = \cdots = P_n \equiv P$ and $Q_1 = \cdots = Q_n \equiv Q$, the formula in (i) becomes $\rho(P^n, Q^n) = \prod_{j=1}^n \rho(P, Q) = \rho(P, Q)^n$. Then the formula in (ii) becomes $H^2(P^n, Q^n) = 1 - (1 - H^2(P, Q))^n$.
(iv) If $\rho(P, Q) < 1$, then $\rho(P^n, Q^n) = \rho(P, Q)^n \to 0$. Thus $H(P^n, Q^n) = 1 - \rho(P^n, Q^n) = 1 - \rho(P, Q)^n \to 1$.