Statistics 582, Midterm Exam Solutions

Wellner; 2/9/2007

1. (24 points) **Define** *any three* of the following terms. In each case, provide an appropriate context for your definition.
   (a) The *risk function* $R(\theta, d)$, $\theta \in \Theta$, of a decision rule d.
   (b) The *Bayes risk* $\mathcal{R}(\Lambda, d)$ of a decision rule d with respect to a prior distribution Λ.
   (c) An *inadmissible* decision rule d.
   (d) An *envelope function* F of a class of functions $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$ in the context of a Uniform Strong Law of Large Numbers (or Glivenko-Cantelli theorem).

   Solution: See Chapter 4 and 5 course notes.

2. (24 points) **State** *any two* of the following results:
   (a) A uniform strong law of large numbers (or Glivenko - Cantelli theorem).
   (b) Wald's theorem on strong consistency of maximum likelihood estimators.
   (c) An identity expressing the survival function $1 - F(t)$ of a random variable X with distribution function $F(t) = P(X \le t)$ on $[0, \infty)$ in terms of the hazard function $\Lambda(t) \equiv \int_{[0,t]} (1 - F(s-))^{-1} dF(s)$.
   (d) Huber's $Z-$theorem for solutions $\hat{\theta}_n$ of $\Psi_n(\hat{\theta}_n) = 0$.

   Solution: See Chapter 4 and 5 course notes plus the $M-$ and $Z-$ theorem handout.

3. (40 points) Suppose that P and $Q_1, \ldots, Q_r$ are probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$, and $\lambda_1, \ldots, \lambda_r$ satisfy $0 \le \lambda_j \le 1$ and $\sum_{j=1}^r \lambda_j = 1$. Let $K(P, Q)$ be the Kullback-Leibler information (or divergence) between P and Q.
   (a) Define $K(P, Q)$ in terms of P and dP/dQ.
   (b) Show that $K(P, \sum_{j=1}^r \lambda_j Q_j) \le \sum_{j=1}^r \lambda_j K(P, Q_j)$.
   (c) Show that $g(x) = x \log x$ is a convex function for $x > 0$.
   (d) Show that $K(\sum_{j=1}^r \lambda_j Q_j, P) \le \sum_{j=1}^r \lambda_j K(Q_j, P)$.
   Hint: note that $K(Q, P) = E_Q \log(dQ/dP) = E_P\{(dQ/dP) \log(dQ/dP)\}$.

   Solution: (a) Since $K(P, Q) = E_P \log(dP/dQ) = -E_P \log(dQ/dP)$ where

$-\log(x)$ is convex,

$$\begin{aligned}
K(P, \sum_{j=1}^r \lambda_j Q_j) &= -E_P \log\left(\frac{d(\sum_{j=1}^r \lambda_j Q_j)}{dP}\right) \\
&= E_P\left\{-\log\left(\sum_{j=1}^r \lambda_j \frac{dQ_j}{dP}\right)\right\} \\
&\leq E_P \sum_{j=1}^r \lambda_j \left(-\log\frac{dQ_j}{dP}\right) \\
&= \sum_{j=1}^r \lambda_j E_P\left(-\log\frac{dQ_j}{dP}\right) \\
&= \sum_{j=1}^r \lambda_j K(P, Q_j).
\end{aligned}$$

(b) $g'(x) = 1 + \log x$, and hence $g''(x) = 1/x > 0$. Thus g is convex.
(c) Let $\overline{Q} \equiv \sum_1^r \lambda_j Q_j$. Then

$$\begin{aligned}
K(\overline{Q}, P) &= E_{\overline{Q}} \log(d\overline{Q}/dP) \\
&= E_P\left\{\frac{d\overline{Q}}{dP}\log\left(\frac{d\overline{Q}}{dP}\right)\right\} \\
&= E_P g(\sum_{j=1}^r \lambda_j dQ_j/dP) \\
&\leq E_P \sum_{j=1}^r \lambda_j g(dQ_j/dP) \\
&= \sum_{j=1}^r \lambda_j E_P g(dQ_j/dP) \\
&= \sum_{j=1}^r \lambda_j E_{Q_j}\log(dQ_j/dP) = \sum_{j=1}^r \lambda_j K(Q_j, P).
\end{aligned}$$

This second inequality plays an important role in certain arguments about hierarchical Bayes procedures; see e.g. Theorem 4.5.7, page 260, Lehmann and Casella, and its proof.

4. (40 points) Suppose that $X_1, \ldots, X_n$ are i.i.d. with $E|X_1| < \infty$, and let $D_n \equiv n^{-1}\sum_{i=1}^n |X_i - \overline{X}_n|^{3/4}$. Use a uniform law of large numbers to show that $D_n \rightarrow_{a.s.} d \equiv E\{|X_1 - \mu|^{3/4}\}$ with $\mu = E(X_1)$.

Solution: Let $\delta > 0$. Then $\overline{X}_n \in [\mu - \delta, \mu + \delta]$ for n sufficiently large with probability 1, so we can consider the collection of functions $f(x,t) \equiv f_t(x) \equiv |x - t|^{3/4}$ with $t \in [\mu - \delta, \mu + \delta]$, namely:

$$\mathcal{F} = \{f_t(x) = |x - t|^{3/4} : \quad \mu - \delta \le t \le \mu + \delta\}.$$

Note that: (a) $T = [\mu - \delta, \mu + \delta]$ is compact; (b) $t \mapsto f(x,t) = f_t(x)$ is continuous in t for all x; and (c) there is a function $F(x)$ such that $EF(X) < \infty$ and $|f(x,t)| \le F(x)$ for all $x \in \mathbb{R}$ and $t \in T$: namely

$$\begin{aligned}
|f(x,t)| &\le |x - (\mu+\delta)|^{3/4} \vee |x - (\mu - \delta)|^{3/4} \\
&\le (|x|^{3/4} + |\mu + \delta|^{3/4}) \vee (|x|^{3/4} + |\mu - \delta|^{3/4}) \\
&\le 2|x|^{3/4} + |\mu + \delta|^{3/4} + |\mu - \delta|^{3/4} \equiv F(x).
\end{aligned}$$

Note that

$$EF(X) = 2E|X|^{3/4} + C(\mu,\delta) \le 2(E|X|)^{3/4} + C(\mu, \delta) < \infty$$

by concavity of $x \mapsto x^{3/4}$, $x \ge 0$, and $E|X| < \infty$. Thus the uniform strong law of large numbers (or Glivenko-Cantelli theorem) holds for $\mathcal{F}$:

$$\sup_{\mu-\delta \le t \le \mu+\delta} |\mathbb{P}_n f_t - P f_t| \to_{a.s.} 0. \tag{1}$$

But then, for $n > N_\omega$ (so that $\overline{X}_n(\omega) \in [\mu - \delta, \mu + \delta]$),

$$\begin{aligned}
|D_n - d| &= |n^{-1} \sum_{i=1}^n |X_i - \overline{X}_n|^{3/4} - E|X_1 - \mu|^{3/4}| \\
&= |\mathbb{P}_n f_{\overline{X}_n} - P f_{\overline{X}_n} + P f_{\overline{X}_n} - P f_\mu| \\
&\le \sup_{\mu-\delta \le t \le \mu+\delta} |\mathbb{P}_n f_t - P f_t| + |P f_{\overline{X}_n} - P f_\mu| \\
&\to_{a.s.} 0 + 0 = 0
\end{aligned}$$

where the first convergence holds by the uniform strong law of large numbers (1) and the second convergence holds by $\overline{X}_n \to_{a.s.} \mu$, continuity of $t \mapsto f_t(x)$, and the dominated convergence theorem.

Do **either** problem 3 **or** problem 4.

5. (40 points). Suppose that $X \equiv X_1, \ldots, X_n$ are i.i.d. F and $Y \equiv Y_1, \ldots, Y_n$ are i.i.d. G, and that we observe (Z_i, Δ_i), $i = 1, \ldots, n$, where $Z_i = X_i \wedge Y_i$, $\Delta_i = 1\{X_i \le Y_i\}$.
(a) Show how the cumulative hazard function $\Lambda = \Lambda_F$ corresponding to F can

be expressed in terms of the sub-distribution functions

$$H^{uc}(t) = P(Z \le t, \Delta = 1), \qquad H^{c}(t) = P(Z \le t, \Delta = 0),$$

and the survival function $1 - H(t) = P(Z > t) = 1 - (H^{uc}(t) + H^{c}(t))$.
(b) Show how the result of (a) leads to a natural nonparametric maximum likelihood estimator of Λ, and hence of F via problem 2(c) above.

Solution: (a) First, $H^{uc}(t) = P(Z \le t, \Delta = 1) = \int_{[0,t]}(1 - G(s-))dF(s)$, $H^{c}(t) = P(Z \le t, \Delta = 0) = \int_{[0,t]}(1 - F(s))dG(s)$, and $1 - H(t) = P(Z > t) = (1 - F(t))(1 - G(t))$, it follows that

$$\begin{aligned} \Lambda(t) &= \int_{[0,t]} \frac{1}{1 - F(s-)} dF(s) = \int_{[0,t]} \frac{1 - G(s-)}{(1 - F(s-))(1 - G(s-))} dF(s) \\ &= \int_{[0,t]} \frac{1}{1 - H(s-)} dH^{uc}(s). \end{aligned}$$

(b) Since nonparametric maximum likelihood estimators of H^{uc}, H^{c}, and $1 - H$ are given by

$$\begin{aligned} &\mathbb{H}_n^{uc}(t) = n^{-1} \sum_{i=1}^{n} \Delta_i 1\{Z_i \le t\}, \\ &\mathbb{H}_n^{c}(t) = n^{-1} \sum_{i=1}^{n} (1 - \Delta_i) 1\{Z_i \le t\}, \qquad \text{and} \\ &1 - \mathbb{H}_n(t) = n^{-1} \sum_{i=1}^{n} 1\{Z_i > t\}, \end{aligned}$$

it follows from (a) that the nonparametric MLE of Λ is given by

$$\widehat{\Lambda}_n(t) = \int_{[0,t]} \frac{1}{1 - \mathbb{H}_n(s-)} d\mathbb{H}_n^{uc}(s).$$

Since

$$1 - F(t) = \exp(-\Lambda_c(t)) \prod_{s \le t} (1 - \Delta\Lambda(s))$$

for any distribution function F with corresponding hazard function Λ and $\Lambda_c(t) \equiv \Lambda(t) - \sum_{s \le t} \Delta\Lambda(s)$, it follows that the nonparametric MLE of F is given by

$$1 - \widehat{F}_n(t) = \prod_{s \le t} (1 - \Delta\widehat{\Lambda}_n(s)).$$

6. (40 points) Suppose that $X_1, \ldots, X_n$ are i.i.d. with mixture density

$$p(x; \lambda, \mu, \theta) = \left(\theta \frac{\lambda(\lambda x)^{r-1}}{\Gamma(r)} e^{-\lambda x} + (1-\theta) \frac{\mu(\mu x)^{s-1}}{\Gamma(s)} e^{-\mu x} \right) 1_{(0,\infty)}(x) \qquad (2)$$

where $0 < \theta < 1$, $0 < \lambda \neq \mu < \infty$, and $r, s > 0$ are known. In other words, p is the mixture of two Gamma densities with parameters (r, λ) and (s, μ) respectively.
(a) Describe an EM - algorithm for estimation of (λ, μ, θ).
(b) What is a natural corresponding nonparametric model for the data which were modeled with the parametric mixture distribution in (a)? What is a natural nonparametric maximum likelihood estimator here?

Solution: (a) Here it is natural to let the "complete data" $\underline{X}$ be $(X_1, \delta_1), \ldots, (X_n, \delta_n)$ where $\delta_i \in \{0, 1\}$ and (X_i, δ_i) are i.i.d. with density

$$p(x, \delta; \theta, \lambda, \mu) = \left(\theta \frac{\lambda(\lambda x)^{r-1}}{\Gamma(r)} e^{-\lambda x} \right)^{\delta} \left((1-\theta) \frac{\mu(\mu x)^{s-1}}{\Gamma(s)} e^{-\mu x} \right)^{1-\delta}$$

for $(x, \delta) \in (0, \infty) \times \{0, 1\}$. Then the incomplete $\underline{Y}$ is $X_1, \ldots, X_n$, which are iid with the mixture density given in (2). It follows that conditional on $X = x$, δ is Bernoulli($\theta(x)$) where

$$\theta(x) \equiv \theta(x; \theta, \lambda, \mu) = \frac{\theta \frac{\lambda(\lambda x)^{r-1}}{\Gamma(r)} e^{-\lambda x}}{\left(\theta \frac{\lambda(\lambda x)^{r-1}}{\Gamma(r)} e^{-\lambda x} + (1-\theta) \frac{\mu(\mu x)^{s-1}}{\Gamma(s)} e^{-\mu x} \right)} . \qquad (3)$$

Hence $E(\delta | X) = \theta(X)$; this is the basis of the E - step of an EM algorithm.

To find the M - step, note that the contribution of each (X, δ) pair to the complete-data log-likelihood is

$$\begin{aligned} l(\theta, \lambda, \mu | X, \delta) &= \delta\{\log \theta + r \log \lambda - \lambda X\} + (1-\delta)\{\log(1-\theta) + s \log \mu - \mu X\} \\ &\quad + \text{constant} , \end{aligned}$$

so that the scores (for a sample of size one) are

$$\dot{\mathbf{l}}_\theta(X, \delta) = \frac{\delta}{\theta} - \frac{1-\delta}{1-\theta} ,$$

$$\dot{\mathbf{l}}_\lambda(X, \delta) = \delta \left\{ \frac{r}{\lambda} - X \right\} ,$$

$$\dot{\mathbf{l}}_\mu(X, \delta) = (1-\delta) \left\{ \frac{s}{\mu} - X \right\} .$$

Thus the score equations for the complete data are solved by

$$\widehat{\lambda}_n^{-1} = \frac{\sum \delta_i X_i}{r \sum \delta_i}, \qquad \widehat{\mu}_n^{-1} = \frac{\sum (1-\delta_i) X_i}{s \sum (1-\delta_i)}, \qquad \widehat{\theta}_n = \frac{\sum \delta_i}{n}.$$

This is the basis of an M - step.

Set $\theta^{(0)} = 1/2$, $(\widehat{\lambda}^{(0)})^{-1} = (\widehat{\mu}^{(0)})^{-1} = \overline{X}$. Then, for $m = 0, 1, \ldots$, define

$$\widehat{\delta}_i^{(m)} \equiv \theta(X_i; \widehat{\theta}^{(m)}, \widehat{\lambda}^{(m)}, \widehat{\mu}^{(m)}) \tag{4}$$

where $\theta(x; \theta, \lambda, \mu)$ is given by (3), and

$$(\widehat{\lambda}^{(m+1)})^{-1} = \frac{\sum \widehat{\delta}_i^{(m)} X_i}{r \sum \widehat{\delta}_i^{(m)}}, \tag{5}$$

$$(\widehat{\mu}^{(m+1)})^{-1} = \frac{\sum (1-\widehat{\delta}_i^{(m)}) X_i}{s \sum (1-\widehat{\delta}_i^{(m)})}, \tag{6}$$

$$\widehat{\theta}^{(m+1)} = \frac{\sum \widehat{\delta}_i^{(m)}}{n}. \tag{7}$$

Iteration of (4) and (5,6,7) yields an EM algorithm for estimation of (θ, λ, μ).

(b) A natural nonparametric model for this data would be

$$\mathcal{P} = \{P : \quad P \text{ an arbitrary distribution on } \mathbb{R}^+ \equiv [0, \infty)\}.$$

In this case the nonparametric maximum likelihood estimator is just $\hat{P}_n = \mathbb{P}_n$ where

$$\mathbb{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i};$$

or, equivalently $\mathbb{F}_n(x) = \mathbb{P}_n((-\infty, x])$.