Statistics 582, Problem Set 2 Solutions

Wellner; 1/18/2006

1. 1. Suppose, as in Example 4.3.10, that $\underline{X}_1, \ldots, \underline{X}_n$ are i.i.d. $\text{Mult}_k(1, \underline{p})$ so that $\underline{N}_n = \sum_{i=1}^n \underline{X}_i \sim \text{Mult}_k(n, \underline{p})$.
(a) Use Jensen's inequality to show that the log-likelihood

$$l_n(\underline{p}|\underline{X}) = \sum_{j=1}^k N_j \log p_j + \sum_{i=1}^n \log\left(\frac{1!}{X_{i1}! \cdots X_{ik}!}\right)$$

is maximized by $\hat{\underline{p}} = \underline{N}_n/n$. [Hint: write the first term of $l_n(\underline{p}|\underline{X})$ as $n \sum_{j=1}^k \hat{p}_j \log p_j$.]
(b) Relate $l_n(\underline{p})$ to $K(\hat{\underline{p}}, \underline{p})$ and hence show again that the maximizing value of $\underline{p}$ is $\hat{p}$.

Solution: (a) Our goal is to show that

$$n \sum_{j=1}^k \hat{p}_j \log p_j \le n \sum_{j=1}^k \hat{p}_j \log \hat{p}_j$$

with equality if and only if $\underline{p} = \hat{\underline{p}}$. Subtracting the right side from the the left side and dividing by n, we see that we want to show that

$$\sum_{j=1}^k \hat{p}_j \log\left(\frac{p_j}{\hat{p}_j}\right) \le 0.$$

But since log is a concave function, Jensen's inequality yields

$$\begin{aligned} \sum_{j=1}^k \hat{p}_j \log\left(\frac{p_j}{\hat{p}_j}\right) &\le \log\left(\sum_{j=1}^k \hat{p}_j \left(\frac{p_j}{\hat{p}_j}\right)\right) \\ &= \log\left(\sum_{j=1}^k p_j\right) = \log(1) = 0. \end{aligned}$$

(b) Note that in the above argument we have shown that

$$l_n(\underline{p}) - l_n(\hat{\underline{p}}) = -nK(\hat{\underline{p}}, \underline{p}) \le 0$$

since $K(P, Q) \ge 0$ for all P, Q. Thus $l_n(\underline{p})$ is maximized by $\underline{p} = \hat{\underline{p}}$.

2. Human beings can be classified into one of four blood groups (phenotypes) O, A, B, AB. The inheritance of blood groups is controlled by three genes, O, A, B, of which O is recessive to A and B. If r, p, q are the gene probabilities in the population of O, A, B respectively, then the probabilities of the six possible combinations (genotypes) in random mating (where two individuals drawn at random from the population contribute one gene each) are shown in the following table:

Phenotype	Genotype	probability
O	OO	r^2
A	AA	p^2
A	AO	$2rp$
B	BB	q^2
B	BO	$2rq$
AB	AB	$2pq$

We observe among N individuals the phenotype frequencies N_O, N_A, N_B, N_{AB}, and wish to estimate the gene probabilities from such data. A simple approach is to regard the observations as incomplete, the complete data set being the genotype frequencies N_{OO}, N_{AA}, N_{AO}, N_{BB}, N_{BO}, N_{AB}.

A. Derive the EM algorithm for estimation of (p, q, r).

B. Estimate (p, q, r) from $N_O = 176$, $N_A = 182$, $N_B = 60$, $N_{AB} = 17$.

C. Estimate the covariance matrix of the estimator $(\hat{p}, \hat{q}, \hat{r})$.

Solution: A. The complete data is $\underline{N} \equiv (N_{OO}, N_{AA}, N_{AO}, N_{BB}, N_{BO}, N_{AB})$ with multinomial distribution $\text{Mult}_6(N; (r^2, p^2, 2rp, q^2, 2rq, 2pq))$. Thus

$$P(\underline{N} = \underline{n}) = \frac{N!}{n_{OO}!n_{AA}!n_{AO}!n_{BB}!n_{BO}!n_{AB}!} \cdot p^{2n_{AA}+n_{AO}+n_{AB}}\, q^{2n_{BB}+n_{BO}+n_{AB}}\, r^{2n_{OO}+n_{AO}+n_{BO}}\, 2^{n_{AO}+n_{BO}+n_{AB}}.$$

This is proportional to a $\text{Mult}_3(2N; (p, q, r))$ distribution, and hence the MLE's based on the complete data are

$$(\hat{p}, \hat{q}, \hat{r}) = \frac{1}{2N}(2N_{AA} + N_{AO} + N_{AB}, 2N_{BB} + N_{BO} + N_{AB}, 2N_{OO} + N_{AO} + N_{BO})\,.$$

This forms the basis of the "M - step"of an E-M algorithm. The incomplete data Y is $(N_A, N_B, N_O, N_{AB}) = (N_{AA} + N_{AO}, N_{BB} + N_{BO}, N_{OO}, N_{AB})$; thus

$$(N_{AA}|Y) = (N_{AA}|N_A) \sim \text{Binomial}(N_A, \frac{p^2}{p^2 + 2rp}), \qquad E(N_{AA}|Y) = N_A \frac{p}{p + 2r},$$

$$(N_{AO}|Y) = (N_{AO}|N_A) \sim Binomial(N_A, \frac{2rp}{p^2+2rp}), \qquad E(N_{AO}|Y) = N_A\frac{2r}{p+2r},$$

$$(N_{BB}|Y) = (N_{BB}|N_B) \sim Binomial(N_B, \frac{q^2}{q^2+2rq}), \qquad E(N_{BB}|Y) = N_B\frac{q}{q+2r},$$

$$(N_{BO}|Y) = (N_{BO}|N_B) \sim Binomial(N_B, \frac{2rq}{q^2+2rq}), \qquad E(N_{BO}|Y) = N_B\frac{2r}{q+2r}.$$

This gives the basis of the "E - step" for an E - M algorithm. Hence, starting from $(\hat{p}^{(0)}, \hat{q}^{(0)}, \hat{r}^{(0)}) = (1/3, 1/3, 1/3)$ say, we take

$$(\hat{p}^{(m+1)}, \hat{q}^{(m+1)}) = \frac{1}{2N}(2\hat{N}_{AA}^{(m)} + \hat{N}_{AO}^{(m)} + N_{AB}, 2\hat{N}_{BB}^{(m)} + \hat{N}_{BO}^{(m)} + N_{AB}),$$

$$\hat{r}^{(m+1)} = 1 - \hat{p}^{(m+1)} - \hat{q}^{(m+1)}$$

where

$$\hat{N}_{AA}^{(m)} \equiv N_A\frac{\hat{p}^{(m)}}{\hat{p}^{(m)}+2\hat{r}^{(m)}}, \qquad \hat{N}_{AO}^{(m)} \equiv N_A - \hat{N}_{AA}^{(m)},$$

$$\hat{N}_{BB}^{(m)} \equiv N_B\frac{\hat{q}^{(m)}}{\hat{q}^{(m)}+2\hat{r}^{(m)}}, \qquad \hat{N}_{BO}^{(m)} \equiv N_B - \hat{N}_{BB}^{(m)}.$$

B. For the given data, the E - M algorithm in A yields:

Iteration	$\hat{p}^{(m)}$	$\hat{q}^{(m)}$
0	.333	.333
1	.298	.111
2	.271	.094
3	.266	.093
4	.265	.093
5	.264	.093
6	.264	.093

Thus the estimator is $(\hat{p}, \hat{q}, \hat{r}) = (.264, .093, .642)$.

C. The likelihood of the observations $(N_A, N_B, N_O, N_{AB}) = (N_{AA} + N_{AO}, N_{BB} + N_{BO}, N_{OO}, N_{AB})$ is

$$\begin{aligned} l_N(p,q) &= N_A\log(p^2+2p(1-p-q)) \\ &\quad + N_B\log(q^2+2q(1-p-q)) \\ &\quad + N_O\log(1-p-q)^2 + N_{AB}\log(2pq). \end{aligned}$$

Thus

$$\begin{aligned}
-\frac{\partial^2}{\partial p^2} l_N(p,q) &= 2N_A\{\frac{1}{2p-p^2-2pq}+\frac{2(1-p-q)^2}{(2p-p^2-2pq)^2}\} \\
&\quad + N_B\frac{4q^2}{(2q-q^2-2pq)^2} \\
&\quad + \frac{N_{AB}}{p^2}+\frac{2N_O}{(1-p-q)^2}\,,
\end{aligned}$$

$$\begin{aligned}
-\frac{\partial^2}{\partial p\partial q} l_N(p,q) &= 2N_A\{\frac{1}{2p-p^2-2pq}-\frac{2p(1-p-q)}{(2p-p^2-2pq)^2}\} \\
&\quad + 2N_B\{\frac{1}{2q-q^2-2pq}-\frac{4q^2}{(2q-q^2-2pq)^2}\} \\
&\quad + \frac{2N_O}{(1-p-q)^2}\,,
\end{aligned}$$

$$\begin{aligned}
-\frac{\partial^2}{\partial q^2} l_N(p,q) &= N_A\frac{4p^2}{(2p-p^2-2pq)^2} \\
&\quad + 2N_B\{\frac{1}{2q-q^2-2pq}+\frac{2(1-p-q)^2}{(2q-q^2-2pq)^2}\} \\
&\quad + \frac{N_{AB}}{q^2}+\frac{2N_O}{(1-p-q)^2}\,.
\end{aligned}$$

Since

$$\begin{aligned}
E(N_A) &= N(p^2+2p(1-p-q))\,, \\
E(N_B) &= N(2q-q^2-2pq)\,, \\
E(N_{AB}) &= N(2pq),
\end{aligned}$$

and

$$E(N_O) = N(1-p-q)^2\,,$$

it follows that

$$I_{11}(p,q) = 2N\{1+\frac{2r^2}{2p-p^2-2pq}-\frac{2q^2}{2q-q^2-2pq}+\frac{q}{p}+1\},$$

$$I_{12}(p,q) = 2N\{2-\frac{2p(1-p-q)}{(2p-p^2-2pq)^2}-\frac{2q(1-p-q)}{(2q-q^2-2pq)^2}+1\}\,,$$

$$I_{22}(p,q) = 2N\{1+\frac{2r^2}{2q-q^2-2pq}-\frac{2p^2}{2p-p^2-2pq}+\frac{p}{q}+1\}$$

and hence the estimated Fisher information matrix is

$$\hat{I}(p,q) = \begin{pmatrix} 5.063 & 1.793 \\ 1.793 & 12.182 \end{pmatrix}$$

so that

$$\hat{I}^{-1}(p,q) = \frac{1}{2N} \begin{pmatrix} .208 & -.003 \\ -.003 & .087 \end{pmatrix}.$$

Furthermore, since $\hat{r} = 1 - \hat{p} - \hat{q}$,

$$Var(\hat{r}) = Var(\hat{p}) + Var(\hat{q}) + 2Cov(\hat{p}, \hat{q}),$$

$$Cov(\hat{p}, \hat{r}) = -Var(\hat{p}) - Cov(\hat{p}, \hat{q}),$$

$$Cov(\hat{q}, \hat{r}) = -Var(\hat{q}) - Cov(\hat{p}, \hat{q});$$

and hence we estimate $Cov(\hat{p}, \hat{q}, \hat{r})$ by

$$\widehat{Cov}(\hat{p}, \hat{q}, \hat{r}) = \begin{pmatrix} .000240 & -.000035 & -.000205 \\ -.000035 & .000095 & -.000060 \\ -.000205 & -.000060 & .000265 \end{pmatrix}.$$

3. Suppose that the "complete data" X is given by three independent multinomial random vectors,

$$N(1) \equiv (N_{ij}(1) : i = 1, \ldots, r; j = 1, \ldots, s) \sim \text{Mult}_{rs}(n_1; p = (p_{ij}, i = 1, \ldots, r, j = 1, \ldots, s),$$

$$N(2) \equiv (N_{ij}(2) : i = 1, \ldots, r; j = 1, \ldots, s) \sim \text{Mult}_{rs}(n_2; p = (p_{ij}, i = 1, \ldots, r, j = 1, \ldots, s),$$

$$N(3) \equiv (N_{ij}(3) : i = 1, \ldots, r; j = 1, \ldots, s) \sim \text{Mult}_{rs}(n_3; p = (p_{ij}, i = 1, \ldots, r, j = 1, \ldots, s).$$

Suppose that the "incomplete data" Y consists of $N(1), (N_{i\cdot}(2) : 1 \leq i \leq r), (N_{\cdot j}(3) : 1 \leq j \leq s)$.
A. What are the distributions of $N(1), (N_{i\cdot}(2) : 1 \leq i \leq r)$ and $(N_{\cdot j}(3) : 1 \leq j \leq s)$?
B. Find the conditional distribution(s) of X given Y.
C. Suggest an EM - algorithm for estimation of p.

Solution: A. By elementary considerations,

$$(N_{i\cdot}(2) : 1 \leq i \leq r) \sim Mult_r(n_2; (p_{i\cdot} : 1 \leq i \leq r))$$

and

$$(N_{\cdot j}(3) : 1 \leq j \leq s) \sim Mult_s(n_3; (p_{\cdot j} : 1 \leq j \leq s)).$$

B. First note that if

$$(N_{ij}) \sim \text{Mult}_{rs}(n; (p_{ij})),$$

then

$$(N_{i\cdot}) \sim \text{Mult}_r(n; (p_{i\cdot}))$$

as in A (since the components of $(N_{i\cdot})$ give the number of times outcome i occurred in n independent trials with probability $p_{i\cdot}$ on each trial). Furthermore

$$((N_{ij})|(N_{i\cdot})) \sim \prod_{i=1}^{r} Mult_s(N_{i\cdot}; (p_{ij}/p_{i\cdot})) \,. \tag{1}$$

(1) can be proved most easily by direct calculation of the conditional distribution:

$$\begin{aligned} & P(N_{ij} = k_{ij}, i = 1, \ldots, r, j = 1, \ldots, s \Big| N_{i\cdot} = k_{i\cdot}, i = 1, \ldots, r) \\ = \;& n! \prod_{i=1}^{r} \prod_{j=1}^{s} \frac{p_{ij}^{k_{ij}}}{k_{ij}!} / n! \prod_{i=1}^{r} \frac{p_{i\cdot}^{k_{i\cdot}}}{k_{i\cdot}!} \\ = \;& \prod_{i=1}^{r} \left\{ k_{i\cdot}! \prod_{j=1}^{s} \frac{(p_{ij}/p_{i\cdot})^{k_{ij}}}{k_{ij}!} \right\} \end{aligned}$$

on the set $k_{i\cdot} = \sum_{j=1}^{s} k_{ij}$, $i = 1, \ldots, r$. The terms inside the first product are just the $\text{Mult}_s(k_{i\cdot}; (p_{ij}/p_{i\cdot}))$ probabilities.

Hence conditional on $(N_{i\cdot}(2) : 1 \le i \le r)$ the vectors $(N_{ij}(2) : 1 \le j \le s)$, $i = 1, \ldots, r$ are independent with $(N_{ij}(2) : 1 \le j \le s)|N_{i\cdot} \sim \text{Mult}_s(N_{i\cdot}; (p_{ij}/p_{i\cdot}; j = 1, \ldots, s))$. Similarly, conditional on $(N_{\cdot j}(3) : 1 \le j \le s)$ the vectors $(N_{ij}(3) : 1 \le i \le r)$, $j = 1, \ldots, s$ are independent with $(N_{ij}(3) : 1 \le i \le r)|N_{\cdot j} \sim \text{Mult}_r(N_{\cdot j}; (p_{ij}/p_{\cdot j}; i = 1, \ldots, r))$.

C. If we had the complete data $N_{ij}(1), N_{ij}(2), N_{ij}(3)$ for all i, j, then $N_{ij} \equiv N_{ij}(1) + N_{ij}(2) + N_{ij}(3)$ has a multinomial distribution with number of trials $n \equiv n_1 + n_2 + n_3$, and hence the MLE $\underline{\hat{p}} = (\hat{p}_{ij})$ of $\underline{p} = (p_{ij})$ is given by

$$\hat{p}_{ij} = \frac{N_{ij}}{n} = \frac{N_{ij}(1) + N_{ij}(2) + N_{ij}(3)}{n_1 + n_2 + n_3}.$$

This is the basis of the "M - step" of an E-M algorithm. But from B it follows that

$$E(N_{ij}(2)|N_{i\cdot}(2)) = N_{i\cdot}(2)\frac{p_{ij}}{p_{i\cdot}} \,, \qquad E(N_{ij}(3)|N_{\cdot j}(3)) = N_{\cdot j}(3)\frac{p_{ij}}{p_{\cdot j}} \,.$$

This is the basis of the "E - step" of an E-M algorithm. Thus, for some reasonable preliminary estimator like $\underline{\hat{p}}^{(0)} \equiv (\hat{p}_{ij}^{(0)}) = (N_{ij}(1)/n)$, a natural E - M algorithm is defined by

$$\hat{p}_{ij}^{(m+1)} = \frac{N_{ij}(1) + \widehat{N}_{ij}^{(m)}(2) + \widehat{N}_{ij}^{(m)}(3)}{n_1 + n_2 + n_3}$$

where

$$\widehat{N}_{ij}^{(m)}(2) \equiv N_{i\cdot}(2)\frac{\hat{p}_{ij}^{(m)}}{\hat{p}_{i\cdot}^{(m)}}\,, \qquad \widehat{N}_{ij}^{(m)}(3) \equiv N_{\cdot j}(3)\frac{\hat{p}_{ij}^{(m)}}{\hat{p}_{\cdot j}^{(m)}}\,.$$

4. (Right censored data). Suppose that $X, X_1, \ldots, X_n$ are i.i.d. survival times with unknown distribution function F, that $Y, Y_1, \ldots, Y_n$ are i.i.d. censoring times with unknown distribution function G, assumed to be independent of the X_i's, and that we can observe only the iid pairs $(Z_1, \delta_1), \ldots, (Z_n, \delta_n)$ where $Z_i \equiv X_i \wedge Y_i$ and $\delta_i \equiv 1_{[X_i \le Y_i]}$; also let $Z \equiv X \wedge Y$ and $\delta = 1_{[X \le Y]}$.
A. Show that the joint distribution of (Z, δ) is given by

$$H^{(uc)}(z) = P(Z \le z, \delta = 1) = \int_{[0,z]} (1 - G(x-))dF(x)$$

where $G(x-) \equiv \lim_{y \uparrow x} G(y)$, and

$$H^{(c)}(z) = P(Z \le z, \delta = 0) = \int_{[0,z]} (1 - F(y))dG(y).$$

Furthermore, show that the survival function $1 - H(z) = P(Z > z)$ is given by $1 - H(z) = (1 - F(z))(1 - G(z))$ and also $H(z) = H^{(uc)}(z) + H^{(c)}(z)$.
B. Suppose that the cumulative hazard function corresponding to F is defined by

$$\Lambda_F(x) = \int_{[0,x]} \frac{1}{1 - F(y-)} dF(y).$$

Show that this can be expressed in terms of H and H_{uc} as

$$\Lambda_F(x) = \int_{[0,x]} \frac{1}{1 - H(y-)} dH^{(uc)}(y).$$

C. If $\mathbb{H}_n^{(uc)}(z) = n^{-1} \sum_{i=1}^n \delta_i 1\{Z_i \le z\}$ and $\mathbb{H}_n(z) = n^{-1} \sum_{i=1}^n 1\{Z_i \le z\}$, suggest an estimator of Λ_F based on the observed (Z_i, δ_i)'s.

Solution: A. First,

$$\begin{aligned} P(Z \le z, \delta = 1) &= P(X \le z, X \le Y) = E\{1_{[X \le z]} 1_{[X \le Y]}\} \\ &= E\{1_{[X \le z]} E(1_{[X \le Y]} | X)\} = E\{1_{[X \le z]}(1 - G(X-))\} \\ &= \int_{[0,z]} (1 - G(x-))dF(x). \end{aligned}$$

Similarly,

$$\begin{aligned} P(Z \le z, \delta = 0) &= P(Y \le z, Y < X) = E\{1_{[Y \le z]} 1_{[Y \le X]}\} \\ &= E\{1_{[Y \le z]} E(1_{[Y < X]} | Y)\} = E\{1_{[Y \le z]}(1 - F(Y))\} \\ &= \int_{[0,z]} (1 - F(y)) dG(y)\,. \end{aligned}$$

Also note that, using integration by parts,

$$\begin{aligned} H(z) &= P(Z \le z) = \int_{(0,z]} (1 - G(x-)) dF(x) + \int_{[0,z]} (1 - F(y)) dG(y) \\ &= (1 - G)F|_{[0,z]} - \int_{[0,z]} F d(1 - G) + \int_{(0,z]} (1 - F) dG \\ &= (1 - G(z))F(z) + G(z) - \int_{[0,z]} (1 - F) dG + \int_{[0,z]} (1 - F) dG \\ &= 1 - (1 - F(z))(1 - G(z)). \end{aligned}$$

B. Using $H^{(uc)}(z) = \int_{[0,z]} (1 - G(x-)) dF(x)$ and $1 - H(z) = (1 - F(z))(1 - G(z))$ we compute

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

C. Since the nonparametric MLE's of $H^{(uc)}$ and H are $\mathbb{H}_n^{(uc)}$ and $\mathbb{H}_n$, it follows that the nonparametric MLE of $\Lambda = \Lambda_F$ is

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

As we discussed in class on 1/18, this is the *Nelson-Aalen* estimator of the cumulative hazard function Λ.