Statistics 582, Problem Set 2 Solutions

Wellner; 1/21/2009

1. 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

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

and

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

it follows that

$$
\begin{aligned}
I_{11}(p,q) &= 2N\{1 + \frac{2r^2}{2p-p^2-2pq} - \frac{2q^2}{2q-q^2-2pq} + \frac{q}{p} + 1\}, \\
I_{12}(p,q) &= 2N\{2 - \frac{2p(1-p-q)}{(2p-p^2-2pq)^2} - \frac{2q(1-p-q)}{(2q-q^2-2pq)^2} + 1\}\,, \\
I_{22}(p,q) &= 2N\{1 + \frac{2r^2}{2q-q^2-2pq} - \frac{2p^2}{2p-p^2-2pq} + \frac{p}{q} + 1\}
\end{aligned}
$$

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

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

2. Lehmann and Casella, TPE, Problem 4.9, page 504.

Solution: (a) The density of a bivariate normal random vector (X, Y) with $\mu_1 = \mu_2 = 0$, variances $\sigma_1^2 \equiv \sigma^2$, $\sigma_2^2 \equiv \tau^2$, and correlation ρ (so that $\theta = (\sigma, \tau, \rho)$) is given by

$$p_\theta(x,y) = \frac{1}{2\pi\sqrt{\sigma^2\tau^2(1-\rho^2)}} \exp\left(-\frac{\frac{x^2}{\sigma^2} - \frac{2\rho xy}{\sigma\tau} + \frac{y^2}{\tau^2}}{2(1-\rho^2)}\right),$$

and the marginal densities of X and Y respectively are given by

$$p_{1,\theta}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right),$$

$$p_{2,\theta}(y) = \frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{y^2}{2\tau^2}\right).$$

Thus the contributions to the log-likelihood are of the form

$$\log p_\theta(x,y) = -\log\sigma - \log\tau - \frac{1}{2}\log(1-\rho^2) - \frac{\frac{x^2}{\sigma^2} - \frac{2\rho xy}{\sigma\tau} + \frac{y^2}{\tau^2}}{2(1-\rho^2)},$$

and $-\log\sigma - x^2/(2\sigma^2)$, $-\log\tau - x^2/(2\tau^2)$, respectively. Thus for the given data the log-likelihood is given by

$$\begin{aligned} l_n(\theta) \;&=\; -4\log\sigma - 4\log\tau - 2\log(1-\rho^2) \\ &\quad -\frac{1}{2(1-\rho^2)}\left\{\frac{1}{\sigma^2} - \frac{2\rho}{\sigma\tau} + \frac{1}{\tau^2}\right. \end{aligned}$$

$$
\begin{aligned}
&\qquad + \frac{1}{\sigma^2} + \frac{2\rho}{\sigma\tau} + \frac{1}{\tau^2} \\
&\qquad + \frac{1}{\sigma^2} + \frac{2\rho}{\sigma\tau} + \frac{1}{\tau^2} \\
&\qquad + \left. \frac{1}{\sigma^2} - \frac{2\rho}{\sigma\tau} + \frac{1}{\tau^2} \right\} \\
&\quad - 4\log\sigma - 4\log\tau - \frac{8}{\sigma^2} - \frac{8}{\tau^2} \\
= &\quad -8\log\sigma - 8\log\tau - 2\log(1-\rho^2) - \frac{1}{1-\rho^2}\left\{\frac{2}{\sigma^2} + \frac{2}{\tau^2}\right\} - \frac{8}{\sigma^2} - \frac{8}{\tau^2}.
\end{aligned}
$$

We compute

$$
\begin{aligned}
\frac{\partial}{\partial\sigma} l_n(\theta) &= -\frac{8}{\sigma} + \frac{4}{(1-\rho^2)\sigma^3} + \frac{16}{\sigma^3} = -\frac{1}{\sigma}\left\{8 - \frac{4}{(1-\rho^2)\sigma^2} - \frac{16}{\sigma^2}\right\}, \\
\frac{\partial}{\partial\tau} l_n(\theta) &= -\frac{8}{\tau} + \frac{4}{(1-\rho^2)\tau^3} + \frac{16}{\tau^3} = -\frac{1}{\tau}\left\{8 - \frac{4}{(1-\rho^2)\tau^2} - \frac{16}{\tau^2}\right\}, \\
\frac{\partial}{\partial\rho} l_n(\theta) &= \frac{4\rho}{1-\rho^2} - \frac{2\rho}{(1-\rho^2)^2}\left\{\frac{2}{\sigma^2} + \frac{2}{\tau^2}\right\} = \frac{2\rho}{(1-\rho^2)}\left\{2 - \frac{1}{1-\rho^2}\left\{\frac{2}{\sigma^2} + \frac{2}{\tau^2}\right\}\right\}.
\end{aligned}
$$

It is easily seen that these scores are zero at both $\theta = (\sqrt{8/3}, \sqrt{8/3}, \pm 1/2)$ and at $\theta = (\sqrt{5/2}, \sqrt{5/2}, 0)$. Furthermore $l_n(\sqrt{8/3}, \sqrt{8/3}, \pm 1/2) = -15.2713...$ while $l_n(\sqrt{5/2}, \sqrt{5/2}, 0) = -15.3303....$ Thus it seems that the first pair of points, $\theta = (\sqrt{8/3}, \sqrt{8/3}, \pm 1/2)$, yield a (non-unique) maximum, and that $\theta = (\sqrt{5/2}, \sqrt{5/2}, 0)$ corresponds to a saddle point. The plot below shows the (exponential of the) likelihood function $(\sigma, \rho) \mapsto \exp[l_n(\sigma, \sigma, \rho)]$.

(b) A natural EM - algorithm for estimation of θ proceeds as follows. Let the complete data X be

$$X = ((X_1, Y_1), \ldots (X_n, Y_n)) \qquad \text{with} \quad n = 12,$$

and let the incomplete data be

$$Y = ((X_1, Y_1), \ldots, (X_4, Y_4), X_5, \ldots, X_8, Y_9, \ldots, Y_{12}).$$

Then, since

$$\begin{array}{lll} E(Y_j|X_j) = \rho\tau X_j/\sigma, & E(Y_j^2|X_j) = \tau^2(1-\rho^2) + (\rho\tau X_j/\sigma)^2, & j = 5, \ldots, 8, \\ E(X_j|Y_j) = \rho\sigma Y_j/\tau, & E(X_j^2|Y_j) = \sigma^2(1-\rho^2) + (\rho\sigma Y_j/\tau)^2, & j = 9, \ldots, 12, \end{array}$$

the conditional expectation of the complete data log-likelihood given Y is given by

$$\begin{aligned} & E\left\{\log p_\theta(X)|Y\right\} \\ &= -12\log\{\sigma\tau(1-\rho^2)^{1/2}\} \\ &\qquad - \frac{1}{2(1-\rho^2)}\left\{\frac{E(\sum_1^{12} X_i^2|Y)}{\sigma^2} - \frac{2\rho E(\sum_1^{12} X_iY_i|Y)}{\sigma\tau} + \frac{E(\sum_1^{12} Y_i^2|Y)}{\tau^2}\right\} \\ &= -12\log\{\sigma\tau(1-\rho^2)^{1/2}\} - \frac{1}{2(1-\rho^2)}\left\{\frac{\hat{T}_{1,1}(Y)}{\sigma^2} - \frac{2\rho\hat{T}_{1,2}(Y)}{\sigma\tau} + \frac{\hat{T}_{2,2}(Y)}{\tau^2}\right\} \end{aligned}$$

where

$$\begin{aligned} \hat{T}_{1,1}(Y) &\equiv \hat{T}_{1,1}(Y,\theta) \equiv E\left(\sum_1^{12} X_i^2|Y\right) \\ &= \sum_{i=1}^{8} X_i^2 + \sum_{i=9}^{12} E(X_i^2|Y_i) \\ &= \sum_{i=1}^{8} X_i^2 + \sum_{i=9}^{12}\{\sigma^2(1-\rho^2) + (\rho\sigma Y_i/\tau)^2\}, \\ \hat{T}_{1,2}(Y) &\equiv \hat{T}_{1,2}(Y,\theta) \equiv E\left(\sum_1^{12} X_iY_i|Y\right) \\ &= \sum_{i=1}^{4} X_iY_i + \sum_{i=5}^{8} X_iE(Y_i|X_i) + \sum_{i=9}^{12} Y_iE(X_i|Y_i) \\ &= \sum_{i=1}^{4} X_iY_i + \sum_{i=5}^{8} X_i(\rho\tau X_i/\sigma) + \sum_{i=9}^{12} Y_i(\rho\sigma Y_i/\tau), \\ \hat{T}_{2,2}(Y) &\equiv \hat{T}_{2,2}(Y,\theta) \equiv E(\sum_1^{12} Y_i^2|Y) \\ &= \sum_{i=1}^{4} Y_i^2 + \sum_{i=5}^{8} E(Y_i^2|X_i) + \sum_{i=9}^{12} Y_i^2 \\ &= \sum_{i=1}^{4} Y_i^2 + \sum_{i=5}^{8}\{\tau^2(1-\rho^2) + (\rho\tau X_i/\sigma)^2\} + \sum_{i=9}^{12} Y_i^2\,. \end{aligned}$$

Furthermore, the MLE's $\hat{\theta} \equiv \hat{\theta}(X) = (\hat{\sigma}, \hat{\tau}, \hat{\rho})$ of $\theta = (\sigma, \tau, \rho)$ for the complete data are given by

$$\hat{\sigma}^2 = n^{-1}T_{1,1}(X) \equiv n^{-1}\sum_{i=1}^{n} X_i^2,$$
$$\hat{\tau}^2 = n^{-1}T_{2,2}(X) \equiv n^{-1}\sum_{i=1}^{n} Y_i^2,$$
$$\hat{\rho} = n^{-1}T_{1,2}(X)/(\hat{\sigma}\hat{\tau}) \equiv n^{-1}\sum_{i=1}^{n} X_iY_i/(\hat{\sigma}\hat{\tau}).$$

We find that the $E-$step of an EM - algorithm is given by

$$\hat{T}^{(m)} \equiv (\hat{T}_{1,1}(Y,\hat{\theta}^{(m)}), \hat{T}_{1,2}(Y,\hat{\theta}^{(m)}), \hat{T}_{2,2}(Y,\hat{\theta}^{(m)})) \equiv (\hat{T}_{1,1}^{(m)}, \hat{T}_{1,2}^{(m)}, \hat{T}_{2,2}^{(m)}).$$

Here $\hat{\theta}^{(0)} = (\hat{\sigma}^{(0)}, \hat{\tau}^{(0)}, \hat{\rho}^{(0)})$ is an initial point to start the algorithm, and, for $m \geq 0$,

$$\hat{\theta}^{(m+1)} = \hat{\theta}(\hat{T}^{(m)}) \equiv \left(n^{-1}\hat{T}_{1,1}^{(m)}, n^{-1}\hat{T}_{2,2}^{(m)}, n^{-1}\hat{T}_{1,2}^{(m)}/(\hat{\sigma}^{(m)}\hat{\tau}^{(m)})\right)$$

gives the M-step.
Note that when $\hat{\rho}^{(0)} = 0$ we have $\hat{T}_{1,2}^{(m)} = \sum_{i=1}^{4} X_iY_i = 0$ for all $m \geq 1$, and hence $\hat{\rho}^{(m)} = 0$ for all $m \geq 0$.
(c) To show that if an EM sequence starts with ρ bounded away from zero, it converges to one of the two maximizing points $\hat{\theta}_{\pm}^{(\infty)} \equiv (\sqrt{8/3}, \sqrt{8/3}, \pm 1/2)$, note that if we start with $\hat{\rho}^{(0)} > 0$, then the sequence $\hat{\rho}^{(m)}$ stays positive for all m. This follows because

$$\hat{T}_{1,2}^{(m)} = 0 + 16\hat{\rho}^{(m)}\frac{\hat{\tau}^{(m)}}{\hat{\sigma}^{(m)}} + 16\hat{\rho}^{(m)}\frac{\hat{\sigma}^{(m)}}{\hat{\tau}^{(m)}} > 0.$$

Furthermore, if we start at $\hat{\theta}^{(0)}$ with $\hat{\sigma}^{(0)} = \hat{\tau}^{(0)}$, then by symmetry of the data, the whole sequence $\hat{\theta}^{(m)}$ satisfies $\hat{\sigma}^{(m)} = \hat{\tau}^{(m)}$. Thus

$$\begin{aligned}
\hat{T}_{1,1}^{(m)} &= 20 + 4(\hat{\sigma}^{(m)})^2(1 - (\hat{\rho}^{(m)})^2) + 16(\hat{\rho}^{(m)})^2 \\
&= 20 + 4(\hat{\tau}^{(m)})^2(1 - (\hat{\rho}^{(m)})^2) + 16(\hat{\rho}^{(m)})^2 = \hat{T}_{2,2}^{(m)}
\end{aligned}$$

and

$$\hat{T}_{1,2} = 16\hat{\rho}^{(m)} + 16\hat{\rho}^{(m)} = 32\hat{\rho}^{(m)}.$$

Since

$$\hat{\rho}^{(m+1)} = \frac{32\hat{\rho}^{(m)}/12}{\hat{\sigma}^{(m)}\hat{\tau}^{(m)}} = \frac{32\hat{\rho}^{(m)}/12}{(\hat{\sigma}^{(m)})^2}$$
$$(\hat{\sigma}^{(m+1)})^2 = \frac{20 + 4(\hat{\sigma}^{(m)})^2(1 - (\hat{\rho}^{(m)})^2) + 16(\hat{\rho}^{(m)})^2}{12},$$

it follows that any limiting point $(\sigma_\infty, \tau_\infty, \rho_\infty)$ must satisfy

$$\rho_\infty = \frac{32}{12}\frac{\rho_\infty}{\sigma_\infty^2}, \qquad \text{and}$$
$$\sigma_\infty^2 = \frac{20}{12} + \frac{1}{3}\sigma_\infty^2(1-\rho_\infty^2) + \frac{4}{3}\rho_\infty^2.$$

The first of these implies that $\sigma_\infty^2 = 8/3$, and plugging this into the second relation we find that $\rho_\infty^2 = 1/4$, or $\rho_\infty = \pm 1/2$. The resulting two points $\theta_\pm^{(\infty)} = (\sqrt{8/3}, \sqrt{8/3}, \pm 1/2)$ are exactly the points of maximum of the incomplete data log-likelihood. This argument extends to the case in which $\hat\sigma^{(0)} \neq \hat\tau^{(0)}$.

It is straightforward to implement the algorithm in Mathematica or R, and numerical experimentation confirms these conclusions.

3. Lehmann and Casella, TPE, Problem 4.12, page 505.

Solution: Suppose X_i, $i = 1, \ldots, n$ are i.i.d. with density $p_\theta(x) = \theta g(x) + (1-\theta)h(x)$ where g and h are known densities with respect to some dominating measure μ. Suppose that Z_i are i.i.d. Bernoulli(θ) and that conditionally on Z_i, $(X_i|Z_i)$ has conditional density $g(x)^{Z_i}h(x)^{1-Z_i}$.
(a) It follows immediately that the joint density q_θ of (X_i, Z_i) is

$$q_\theta(x_i, z_i) = g(x_i)^{z_i}h(x_i)^{1-z_i}\theta^{z_i}(1-\theta)^{1-z_i},$$

and the joint density of the complete data is

$$q_\theta(\underline{x}, \underline{z}) = \prod_{i=1}^n (\theta g(x_i))^{z_i}((1-\theta)h(x_i))^{1-z_i} = \prod_{i=1}^n \{z_i g(x_i) + (1-z_i)h(x_i)\}\theta^{z_i}(1-\theta)^{z_i}.$$

Thus the complete data likelihood is given by

$$L(\theta|\underline{X}, \underline{Z}) = \prod_{i=1}^n (\theta g(X_i))^{Z_i}((1-\theta)h(X_i))^{1-Z_i}.$$

(b) To show that

$$E(Z_i|\theta, X_i) = \frac{\theta g(X_i)}{\theta g(X_i) + (1-\theta)h(X_i)}, \tag{1}$$

it suffices to show that

$$E\left\{1_B(X_i)\frac{\theta g(X_i)}{\theta g(X_i) + (1-\theta)h(X_i)}\right\} = E\{1_B(X_i)Z_i\}$$

for all Borel sets B. But since the marginal density of X_i is given by p_θ, we see that the left side in the last display is equal to $\theta \int 1_B(x)g(x)d\mu(x)$, and this equals the right side by direct calculation. Thus (1) holds.

For the complete data $(X_1, Z_1), \ldots, (X_n, Z_n)$, the MLE of θ is $\widehat{\theta}_n = \overline{Z}_n$. Hence the EM algorithm for estimation of θ is given by

$$\begin{aligned}
\widehat{\theta}_n^{(j+1)} &= \frac{1}{n} \sum_{i=1}^n \widehat{Z}_i(\widehat{\theta}_n^{(j)}, X_i) && (2)\\
&= \frac{1}{n} \sum_{i=1}^n \frac{\widehat{\theta}_n^{(j)} g(X_i)}{\widehat{\theta}_n^{(j)} g(X_i) + (1-\widehat{\theta}_n^{(j)}) h(X_i)}. && (3)
\end{aligned}$$

(c) The expected complete data log-likelihood given the observed (or incomplete) data is

$$\begin{aligned}
&Q(\theta|\theta_0, \underline{X})\\
&= E_{\theta_0}\left(\sum_{i=1}^n \{Z_i(\log\theta + \log g(X_i)) + (1-Z_i)(\log(1-\theta) + \log h(X_i))\} \Big| \underline{X} \right)\\
&= \sum_{i=1}^n \left\{ \frac{\theta_0 g(X_i)}{\theta_0 g(X_i) + (1-\theta_0) h(X_i)} (\log\theta + \log g(X_i)) \right.\\
&\qquad\qquad \left. + \left(1 - \frac{\theta_0 g(X_i)}{\theta_0 g(X_i) + (1-\theta_0) h(X_i)}\right) (\log\theta + \log g(X_i)) \right\}.
\end{aligned}$$

This is a continuous function of θ_0 and $\theta \in (0,1)$. Thus by Theorem 4.12, Lehmann and Casella page 460, all limit points of an EM sequence $\{\widehat{\theta}^{(j)}\}$ are stationary points of $L(\theta|\underline{X})$, and $L(\widehat{\theta}^{(j)}|\underline{X})$ converges monotonically to $L(\widehat{\theta}|\underline{X})$ for some stationary point $\widehat{\theta}$.

In this case, the log-likelihood function for the incomplete data is (strictly) concave: note that

$$l_n(\theta|\underline{X}) = \sum_{i=1}^n \log\{\theta g(X_i) + (1-\theta) h(X_i)\}$$

is a concave function of θ since the individual terms in the sum are logarithms of linear functions. Here we compute

$$\begin{aligned}
\dot{l}_n(\theta) &= \sum_{i=1}^n \frac{g(X_i) - h(X_i)}{\theta g(X_i) + (1-\theta) h(X_i)}\\
\ddot{l}_n(\theta) &= -\sum_{i=1}^n \frac{(g(X_i) - h(X_i))^2}{(\theta g(X_i) + (1-\theta) h(X_i))^2} < 0
\end{aligned}$$

with strict inequality if $g(X_i) \neq h(X_i)$ for some i. In the latter case, the MLE $\widehat{\theta}$ exists and is the unique stationary point of the log-likelihood. Note that if $\widehat{\theta}^{(j)}$ converges to, say $\widehat{\theta}^{(\infty)}$, then by (3), $\widehat{\theta}^{(\infty)}$ satisfies

$$\widehat{\theta}_n^{(\infty)} = \frac{1}{n} \sum_{i=1}^n \frac{\widehat{\theta}_n^{(\infty)} g(X_i)}{\widehat{\theta}_n^{(\infty)} g(X_i) + (1-\widehat{\theta}_n^{(\infty)}) h(X_i)}.$$

But this is just another way of writing the score equation $\dot{l}_n(\widehat{\theta}^{(\infty)}) = 0$. Thus $\widehat{\theta}^{(j)} \to \widehat{\theta}$, the MLE of θ based on $\underline{X}$.

4. Lehmann and Casella, TPE, Problem 4.16, page 506. (It seems to me that ζ_i in the third line of the problem statement should be just ζ.)

Solution: Here $Z_i \sim N(\zeta, \sigma^2)$, $i = 1, \ldots, n$ are i.i.d., and then $X_i = 1_{(u,\infty)}(Z_i)$ for $i = 1, \ldots, n$. Thus $X_i \sim \text{Bernoulli}(p)$ with

$$\begin{aligned} p &\equiv p(\zeta, \sigma) = P_{\zeta,\sigma}(Z > u) = 1 - \Phi\left(\frac{u - \zeta}{\sigma}\right) \\ &= \Phi\left(\frac{\zeta - u}{\sigma}\right). \end{aligned}$$

(a) Thus the likelihood of the X_i's (the incomplete data) is

$$L(\zeta, \sigma|\underline{X}) = p^{\sum_1^n X_i}(1-p)^{n - \sum_1^n X_i}.$$

(b) The likelihood of the Z_i's (the complete data) is

$$L(\zeta, \sigma|\underline{Z}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Z_i - \zeta)^2\right),$$

and the expected complete data log-likelihood given the observed (or incomplete data) is, with $\theta = (\zeta, \sigma)$, $\theta_0 = (\zeta_0, \sigma_0)$,

$$Q(\theta|\theta_0, \underline{X}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_1^n \left\{E_{\theta_0}(Z_i^2|X_i) - 2\zeta E_{\theta_0}(Z_i|X_i) + \zeta^2\right\}.$$

(c) Since the MLE's for the complete data (the Z_i's) are the usual

$$\begin{aligned} &\hat{\zeta} = \overline{Z} \qquad \text{and} \\ &\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (Z_i - \overline{Z})^2 = n^{-1}\sum_{i=1}^n Z_i^2 - \overline{Z}_n^2, \end{aligned}$$

it follows that the EM sequence is given by

$$\begin{aligned} \hat{\zeta}^{(j+1)} &= \frac{1}{n}\sum_{i=1}^n t(\hat{\zeta}^{(j)}, \hat{\sigma}^{(j)}, X_i), \\ (\hat{\sigma}^{(j+1)})^2 &= \frac{1}{n}\sum_{i=1}^n v(\hat{\zeta}^{(j)}, \hat{\sigma}^{(j)}, X_i), \end{aligned}$$

where

$$\begin{aligned} t(\zeta, \sigma, X) &= E(Z|X, \zeta, \sigma), \\ v(\zeta, \sigma, X) &= E(Z^2|X, \zeta, \sigma). \end{aligned}$$

(d) To find explicit expressions for the conditional expectations in the last display, we proceed via the following two claims:
Claim 1. Let $W \sim N(0,1)$ and $Y \equiv 1_{(t,\infty)}(W)$. Then

$$E(W|Y) = Y\frac{\phi(t)}{1-\Phi(t)} + (1-Y)\frac{-\phi(t)}{\Phi(t)} \equiv H(t,Y),$$
$$E(W^2|Y) = 1 + tE(W|Y).$$

Claim 2. If $Z \sim N(\zeta, \sigma^2)$ and $X \equiv 1_{(u,\infty)}(Z)$, then

$$E(Z|X,\zeta,\sigma^2) = \zeta + \sigma H\left(\frac{u-\zeta}{\sigma}, X\right),$$
$$E(Z^2|X,\zeta,\sigma^2) = \zeta^2 + \sigma^2 + \sigma(u+\zeta)H\left(\frac{u-\zeta}{\sigma}, X\right).$$

Proof of Claim 1: To prove the first part of Claim 1 we need to show that

$$E\left\{1_B(Y)\left(Y\frac{\phi(t)}{1-\Phi(t)} + (1-Y)\frac{-\phi(t)}{\Phi(t)}\right)\right\} = E\{1_B(Y)W\}$$

for all Borel sets B; see e.g. Lehmann and Romano, TSH, pages 36 - 37, or Shorack, *Probability for Statisticians*, page 158. Since Y takes values in $\{0,1\}$, it suffices to show this for $B = \{0\}$ and for $B = \{1\}$. For $B = \{1\}$, the left side equals

$$E\left\{Y\frac{\phi(t)}{1-\Phi(t)}\right\} = \phi(t),$$

while the right side equals

$$E\{1_B(Y)W\} = E\{W1_{[W>t]}\} = \int_t^\infty z\phi(z)dz = -\int_t^\infty \phi'(z)dz = \phi(t),$$

so the required identity holds. For $B = \{0\}$ the left side equals

$$E\left\{(1-Y)\frac{-\phi(t)}{\Phi(t)}\right\} = -\phi(t),$$

and the right side equals

$$E\{W1_{[W\le t]}\} = \int_{-\infty}^t z\phi(z)dz = -\int_{-\infty}^t \phi'(z)dz = -\phi(t),$$

so the identity holds in this case as well, and this completes the proof of the first part of the claim. To prove the second part of claim 1, we need to show that

$$E\left\{1_B(Y)(1+tE(W|Y))\right\} = E\{1_B(Y)W^2\}$$

for all Borel sets B. As before, since Y takes values in $\{0,1\}$ it suffices to consider $B=\{0\}$ and $B=\{1\}$. For $B=\{1\}$, we use the calculations above to see that the left side equals

$$p + t\phi(t)$$

while the right side equals

$$\begin{aligned}
E(W^2 1_{[W>t]}) &= \int_t^\infty z^2\phi(z)dz = -\int_t^\infty z\phi'(z)dz \equiv -\int_t^\infty u dv \\
&= -\left\{ uv|_t^\infty - \int_t^\infty v du \right\} \qquad \text{with } u=z, \quad v=\phi(z), \\
&= t\phi(t) + 1 - \Phi(t) = t\phi(t) + p,
\end{aligned}$$

so the required identity holds. The verification for $B=\{0\}$ is similar, and this completes the proof of Claim 1.

Proof of Claim 2: This proceeds by reduction to Claim 1: note that $(Z-\zeta)/\sigma =_d W \sim N(0,1)$, and $X = 1_{[Z>u]} = 1_{[(Z-\zeta)/\sigma > (u-\zeta)/\sigma]} =_d Y$ with $t=(u-\zeta)/\sigma$. Thus

$$\begin{aligned}
t(\zeta,\sigma,X) &\equiv E(Z|X,\zeta,\sigma) \\
&= \zeta + \sigma E\left(\frac{Z-\zeta}{\sigma}|X = 1\{(Z-\zeta)/\sigma > (u-\zeta)/\sigma\}\right) \\
&= \zeta + \sigma H\left(\frac{u-\zeta}{\sigma}, X\right),
\end{aligned}$$

and (with $t=(u-\zeta)/\sigma$),

$$\begin{aligned}
v(\zeta,\sigma,X) &\equiv E(Z^2|X,\zeta,\sigma) = E((\zeta+\sigma W)^2|Y) \\
&= \zeta^2 + 2\sigma\zeta E(W|Y) + \sigma^2 E(W^2|Y) \\
&= \zeta^2 + 2\sigma\zeta H\left(\frac{u-\zeta}{\sigma}, X\right) + \sigma^2\left(1 + \frac{u-\zeta}{\sigma} H\left(\frac{u-\zeta}{\sigma}, X\right)\right) \\
&= \zeta^2 + \sigma^2 + \sigma(u+\zeta) H\left(\frac{u-\zeta}{\sigma}, X\right).
\end{aligned}$$

(e) To see that the EM iterates converge to the ML estimates $\hat{\zeta}$ and $\hat{\sigma}$ of ζ and σ, note that $t(\zeta,\sigma,X)$ and $v(\zeta,\sigma,X)$ are continuous functions of ζ and σ, and hence the expected complete data log-likelihood $Q(\theta|\theta_0,\underline{X})$ is continuous in both $\theta=(\zeta,\sigma)$ and $\theta_0=(\zeta_0,\sigma_0)$. It follows from Theorem 4.12 of TPE page 460 that all the limit points of an EM iteration sequence are stationary points of $L(\zeta,\sigma|\underline{X})$. Unfortunately, it seems to me that ζ and σ are not identifiable in the current version of this model: Note that if $p(\zeta,\sigma)=p(\gamma,\tau)$, then we have

$$\Phi\left(\frac{\zeta-u}{\sigma}\right) = \Phi\left(\frac{\gamma-u}{\tau}\right),$$

and hence

$$\frac{\zeta - u}{\sigma} = \frac{\gamma - u}{\tau},$$

or $\zeta = u + (\sigma/\tau)(\gamma - u)$. Thus the likelihood is constant along this surface, and gives no information about ζ or σ separately. This is consistent with what we get by writing down the score equations for the incomplete data version of the model: In this case

$$l(\zeta, \sigma | \underline{X}) = \sum_1^n X_i \log p(\zeta, \sigma) + (n - \sum_1^n X_i) \log(1 - p(\zeta, \sigma)),$$

so, with $T \equiv \sum_1^n X_i$,

$$\begin{aligned} \dot{l}_\zeta(\zeta, \sigma | \underline{X}) &= \frac{T}{p} \frac{\partial p}{\partial \zeta} - \frac{n - T}{1 - p} \frac{\partial p}{\partial \zeta}, \\ &= \frac{T - np}{p(1-p)} \frac{\partial p}{\partial \zeta} \\ \dot{l}_\sigma(\zeta, \sigma | \underline{X}) &= \frac{T}{p} \frac{\partial p}{\partial \sigma} - \frac{n - T}{1 - p} \frac{\partial p}{\partial \sigma} \\ &= \frac{T - np}{p(1-p)} \frac{\partial p}{\partial \sigma}, \end{aligned}$$

which shows that the score equations degenerate to one equation yields estimation of $p = P(Z > u)$ by $\hat{p} = T/n$, but not ζ and σ.

It seems to me that a natural modification of the problem which does have a non-trivial solution is obtained by letting the cut-off level u change with i: then $p_i = P(Z_i > u_i) = \Phi((\zeta - u_i)/\sigma)$, and we should regard the incomplete data data as $(u_1, X_1), \ldots, (u_n, X_n)$. Note the close correspondence with the current status or interval censoring model we have discussed from a nonparametric perspective.

5. 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 $N(1), (N_{i\cdot}(2) : 1 \le i \le r), (N_{\cdot j}(3) : 1 \le j \le s)$. Thus $N(1)$ gives cell counts for a two-way table, while $(N_{i\cdot}(2) : 1 \le i \le r)$ and $(N_{\cdot j}(3) : 1 \le j \le s)$ give additional information on the marginal distributions of the table. (If n_2 and n_3 are very large relative to n_1, we might regard the marginal distributions as "known".)

A. What are the distributions of $N(1), (N_{i\cdot}(2) : 1 \le i \le r)$ and $(N_{\cdot j}(3) : 1 \le j \le$

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,

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

and

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

B. First note that if

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

then

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

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

$$((N_{ij})|(N_{i\cdot})) \sim \prod_{i=1}^{r} Mult_s(N_{i\cdot}; (p_{ij}/p_{i\cdot}))\,. \tag{4}$$

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

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

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

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

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

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

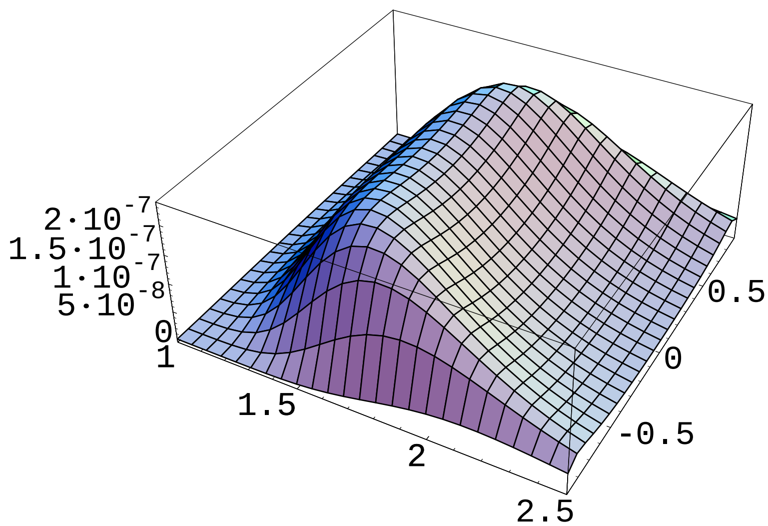

Figure 1: Plot of $(\sigma, \rho) \mapsto \exp[l_n(\sigma, \sigma, \rho)]$.

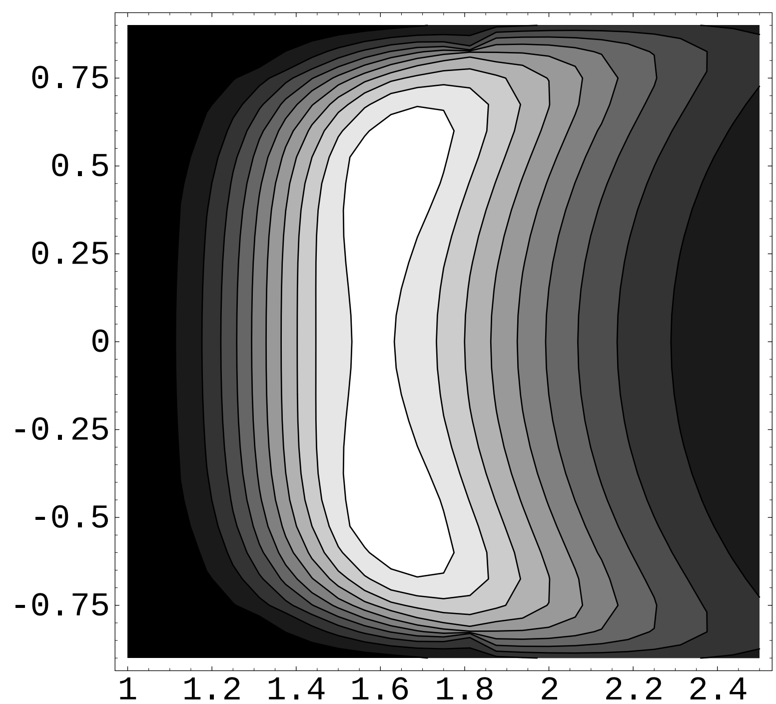

Figure 2: Contour plot of $(\sigma, \rho) \mapsto \exp[l_n(\sigma, \sigma, \rho)]$.