STAT 613 Midterm (1 hour 15 minutes) April 13th, 2012

Marks will be given for clarity of the solution.

Good Luck!

(1) The object of this question is to use the log-likelihood ratio test to derive the χ-squared test for independence (in the case of two by two tables). In other words, derive the distribution of the test statistic

$$T = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4},$$

under the null that there is no association between the categorical variables C and R, where and $E_1 = n_3 \times n_1/N$, $E_2 = n_4 \times n_1/N$, $E_3 = n_3 \times n_2/N$ and $E_2 = n_4 \times n_2/N$.

<table>
<thead>
<tr>
<th></th>
<th>C₁</th>
<th>C₂</th>
<th>Subtotal</th>
</tr>
</thead>
<tbody>
<tr>
<td>R₁</td>
<td>O₁</td>
<td>O₂</td>
<td>n₁</td>
</tr>
<tr>
<td>R₂</td>
<td>O₃</td>
<td>O₄</td>
<td>n₂</td>
</tr>
<tr>
<td>Subtotal</td>
<td>n₃</td>
<td>n₄</td>
<td>N</td>
</tr>
</tbody>
</table>

State all results you use. (hint: It may be useful to use the Taylor approximation $x \log(x/y) \approx (x - y) + \frac{1}{2}(x - y)^2/y$).

(2) Consider the following shifted exponential mixture distribution

$$f(x; \lambda_1, \lambda_2, p, a) = p \frac{1}{\lambda_1} \exp(-x/\lambda_1) I(x \geq 0) + (1-p) \frac{1}{\lambda_2} \exp(-(x-a)/\lambda_2) I(x \geq a),$$

where $p, \lambda_1, \lambda_2$ and $a$ are unknown.

(i) Make a plot of the above mixture density.

Considering the cases $x \geq a$ and $x < a$ separately, calculate the probability of belonging to each of the mixtures, given the observation $X_i$. (i.e. Define the variable $\delta_i$, where $P(\delta_i = 0) = p$, $f(x|\delta_i = 0) = \frac{1}{\lambda_1} \exp(-x/\lambda_1)$ etc. and calculate $P(\delta_i = 0|X_i = x)$ and $P(\delta_i = 0|X_i = x)$).

(ii) Show how the EM-algorithm can be used to estimate $a, p, \lambda_1, \lambda_2$. At each iteration you should be able to obtain explicit solutions for most of the parameters, give as many details as you can.

Hint: It may be beneficial for you to use profiling too.

(iii) From your knowledge of estimation of these parameters, what do you conjecture the rates of convergence to be? Will they all be the same, or possibly different?

(iv) Not part of the exams: code the estimator. Through simulations try to verify your conjecture in (iii).
Here we have the $2 \times 2$ table:

\[
\begin{array}{ccc}
 & Y_1 & Y_2 \\
X_1 & P_1 & P_2 \\
X_2 & P_3 & P_4 \\
\end{array}
\]

$H_0$: There is no association in which case

\[P_1 = q_1 q_2, \quad P_2 = (1-q_1) q_2\]
\[P_3 = q_1 (1-q_2), \quad P_4 = (1-q_1)(1-q_2)\]

$H_A$: There is an association in which case

\[P_1 + P_2 + P_3 + P_4 = 1\]

We observe

\[
\begin{array}{c|c|c|c}
 & y_1 & y_2 \\
X_1 & n_{11} & n_{12} \\
X_2 & n_{21} & n_{22} \\
\end{array}
\]

\[n = n_{11} + n_{12} + n_{21} + n_{22}\]

In other words, the number $n_{11}$ is the number of people out of $n$ who are in the $(Y_1, X_1)$ category.

eetc.

We model the observations $(n_{11}, n_{12}, n_{21}, n_{22})$ using a multinomial distribution.

\[L(\Theta) = \left( \frac{n}{n_{11} n_{12} n_{21} n_{22}} \right)^n \prod_{i=1}^{2} \prod_{j=1}^{2} n_{ij}^{n_{ij}}\]
Now we want to evaluate estimates of $\pi_{11}, \ldots, \pi_{22}$ under both the null and the alternative.

For the alternative (it is easier!):

$$\log L \propto n \log \pi_1 + n \log \pi_2 + n \log \pi_3 + n \log \pi_4$$

Differentiating with respect to $\pi_1, \pi_2, \pi_3,$ and $\pi_4$ and setting to zero gives

$$\hat{\pi}_1 = \frac{n_{11}}{n}, \quad \hat{\pi}_2 = \frac{n_{21}}{n}, \quad \hat{\pi}_3 = \frac{n_{12}}{n}, \quad \hat{\pi}_4 = \frac{n_{22}}{n}$$

(Quite obvious!)

Next, for the null:

$$\log L(\pi) = \log \left( \frac{n}{n_{11} n_{21} n_{12} n_{22}} \right) + n_{11} \log \hat{q}_1 q_2 + n_{21} \log \hat{q}_1 (1-q_2) + n_{12} \log (1-q_1) q_2 + n_{22} \log (1-q_1) (1-q_2)$$

Differentiating with respect to $q_1$ and $q_2$ and setting to zero gives

$$\hat{q}_1 = \frac{n_{11} + n_{21}}{n}, \quad \hat{q}_2 = \frac{n_{11} + n_{12}}{n}$$
To do the CRT we substitute these estimates into the full and restricted likelihoods.

\[ \mathcal{L}_F(\pi) - \mathcal{L}_R(\pi) \]

\[ = 2 \left[ n_{11} \left( \log \hat{p}_1 - \log \hat{q}_1 \hat{q}_2 \right) + n_{12} \left( \log \hat{p}_2 - \log (1-\hat{q}_1) \hat{q}_2 \right) + n_{21} \left( \log \hat{p}_3 - \log \hat{q}_1 \hat{q}_3 \right) + n_{22} \left( \log \hat{p}_4 - \log (1-\hat{q}_1)(1-\hat{q}_3) \right) \right] \]

After lots of cancellations will and using the fact that

\[ \begin{align*}
O_{11} &= n_{11} \\
O_{12} &= n_{12} \\
O_{21} &= n_{21} \\
O_{22} &= n_{22} \\
E_{11} &= \left( \frac{n_{11} + n_{12}}{n} \right) \times \left( \frac{n_{11} + n_{21}}{n} \right) \\
E_{22} &= \frac{(n_{22} + n_{21})(n_{22} + n_{12})}{n}
\end{align*} \]

we have

\[ 2 \left( \mathcal{L}_F(\pi) - \mathcal{L}_R(\pi) \right)^2 = 2 \left\{ \log O_{11} \log \frac{O_{11}}{E_{11}} + \log O_{12} \log \frac{O_{12}}{E_{12}} + \log O_{21} \log \frac{O_{21}}{E_{21}} + \log O_{22} \log \frac{O_{22}}{E_{22}} \right\} \]

Noting that under the null hypothesis we can see that

\[ \begin{align*}
& \frac{D}{\chi^2_{1}} \\
& \chi^2_{3-2=1} \alpha.
\end{align*} \]

To obtain the form of the \( \chi^2 \)-test for independence
we use the expression
\[ \log \frac{O}{E} \approx \frac{O - E + \frac{1}{2} (O - E)^2}{E} \]

This gives
\[ \sum_{i,j} \left( \frac{O_{ij} - E_{ij}}{E_{ij}} \right)^2 \rightarrow \chi^2_1. \]
2) Mixture model:

\[ f(x; \theta) = \frac{p}{a_1} e^{-x/a_1} I(x > 0) + \frac{(1-p)}{a_2} e^{-(x-a)/a_2} I(x \geq a) \]

where \( \theta = (p, a_1, a_2, a) \) are unknown parameters.

Define the variable \( S = \begin{cases} 0 & \text{if } X < a \\ 1 & \text{if } X \geq a \end{cases} \)

where

\[ P(X|S=0) = \frac{1}{a_1} e^{-x/a_1} \]

\[ P(X|S=1) = \frac{1}{a_2} e^{-(x-a)/a_2} \]

\[ P(S=0) = p \text{ and } P(S=1) = 1-p \]

a) Let us consider what happens to \( S \) when \( x < a \) and \( x \geq a \).

If \( x < a \), it is clear that we are in mixture 1, hence

\[ P(S=0|X=x) = p \text{ and } P(S=1|X=x) = 0 \]

If \( x \geq a \), then \( S \) can be either in mixture 1 or 2.

\[ P(S=0|X=x) = \frac{p d_1^{-1} e^{-x/a_1}}{pd_1^{-1} e^{-x/a_1} + (1-p) d_2^{-1} e^{-(x-a)/a_2}} \]

\[ P(S=1|X=x) = \frac{(1-p) d_2^{-1} e^{-(x-a)/a_2}}{pd_1^{-1} e^{-x/a_1} + (1-p) d_2^{-1} e^{-(x-a)/a_2}} \]
Now to construct the EM-algorithm.

At the $k$th step we have the parameters

$$\alpha_k, \beta_1, \beta_2, \kappa_k$$

and we want to obtain estimates of the parameters

$$\alpha_{k+1}, \beta_{1k+1}, \beta_{2k+1}, \kappa_{k+1}$$

at the next stage.

The full likelihood is the likelihood of $E(X_c, S_c)$

$$L_c(\theta) = \sum_{i:S_i} \left\{ -\log \alpha_{1} - \alpha_{1} X_i X_3 \right\}_i$$

$$+ \sum_{i:S_i} \left\{ -\log \alpha_{2} - \alpha_{2} (X_i - a) + \log I(X_i > a) \right\}_i$$

$$+ \sum_{i:S_i} \log p + \sum_{i:S_i} \log (1-p)$$

Of course since $S_c$ is not observed the above likelihood cannot be maximised. Instead we look at the best approximation of $L_c(\theta)$ given what we do observe.

$$\mathbb{E} \theta(\theta, \sigma^2) = \mathbb{E}(L_c(\theta) \mid X, \sigma^2)$$

$$= \sum_{i} \mathbb{E}(S_i \mid X_i, \sigma^2) \left\{ -\log \alpha_1 - \alpha_1 X_i X_3 \right\}_i$$

$$+ \sum_{i} \mathbb{E}(S_i \mid X_i, \sigma^2) \left\{ -\log \alpha_2 - \alpha_2 (X_i - a) + \log I(X_i > a) \right\}_i$$

$$+ \sum_{i} \mathbb{E}(S_i \mid X_i, \sigma^2) \log p + \sum_{i} \mathbb{E}(S_i \mid X_i, \sigma^2) \log (1-p).$$
Let \( \pi_i^* = P(\delta_l | x_l = x, \theta^*) \)
\[ 1 - \pi_i^* = P(\delta_l = 1 | x_l = x, \theta^*) \]
how we see from pot (2) that the value of \( \pi_i^* \)
depends on where \( x \) was used with respect to \( a^*_l \)
(the shift estimator at the previous iteration).

If \( x < a^*_l \) then \( \delta_l = 0 \), hence \( \pi_i^* = 0 \), and \( 1 - \pi_i^* = 0 \)
If \( x \geq a^*_l \), then \( \pi_i^* \) and \( 1(1 - \pi_i^*) \) will take different values.

\[
\begin{align*}
\text{must be in} \\
\text{estimated mixture!}
\end{align*}
\]

So we want to maximize
\[
Q(\Theta, \Theta^*) = \sum \pi_i^* \left[ -\log d_1 - d_1^T x_l \right] + \sum (1 - \pi_i^*) \left[ -\log d_2 - d_2^T (x_l - a) \right] \\
+ \log I(x_l > a)^2 + \sum \pi_i^* \log \pi + \sum (1 - \pi_i^*) \log(1 - \pi).
\]

at the \( (k+1)^{th} \) iteration.

It is straightforward to show that
\[
\hat{a}^*_{k+1} = \frac{\sum \pi_i^* x_l}{\sum \pi_i^*} \\
\hat{\theta} = \frac{\sum \pi_i^* \theta}{\sum \pi_i^*} \\
\hat{p} = \frac{\sum \pi_i^*}{n}
\]
The tricky part is dealing with estimators of \( \theta \) and \( \theta_2 \).

Here I use profiling. I assume that \( \theta_1 \) is known.

If \( \theta_1 \) is 'known', then my estimate \( \hat{\theta}_2 \) is standard (by differentiating \( \psi \)).

The profile estimate \( \hat{\theta}_2 \) is

\[
\hat{\theta}_{2, k+1}(a) = \frac{\sum_i (1 - \pi_i^x) (x_i - a)}{\sum (1 - \pi_i^x)}
\]

Of course \( \theta_1 \) is not known and we have to estimate this. Putting \( \hat{\theta}_{2, k+1}(a) \) back into \( \psi(\theta, \theta_2) \) gives

\[
\psi(\theta, \theta_2) |_{\hat{\theta}} = \sum_i \pi_i^x \left[ -\log \lambda_1 - \lambda_1' x_i \right] + \sum (1 - \pi_i^x) \right] - \log d_2(a)
\]

These parts do not contain \( \theta_1 \) and so we consider the only part which matter.

\[
\psi(\theta, \theta_2) |_{\hat{\theta}} = \sum \pi_i^x \log p + \sum (1 - \pi_i^x) \log (1 - p)
\]

\[
P(a) = \sum (1 - \pi_i^x) \left[ -\log d_2(a) = d_2(a) (x_i - a) + \log I(x_i > a) \right]
\]

It is difficult to maximize this. So we consider
the exponential of the likelihood
\[
\exp \{ P(a) \} = \prod_{l=1}^{n} \left( \hat{d}_l(a) e^{-\frac{\hat{d}_l(a) [X_l - a]}{\Sigma_l(1-\pi_i^*) (X_l - a) \Sigma_l(1-\pi_i^*)^2}} \right) I(X_l > a) \]

where \[ 1 - \pi_i^* = 0 \quad \forall \quad X_l < a^*_K, \]
thus we only consider those \( X_l \) where \( X_l > a^*_K \). To do this we order \( \{X_l\} \rightarrow \{X_{\omega}\} \)

\[
\exp \{ P(a) \} = \prod_{l, X_{\omega} > a^*} \left( \hat{d}_l(a) e^{\frac{-\hat{d}_l(a) [X_{\omega} - a]}{\Sigma_l(1-\pi_i^*) (X_{\omega} - a) \Sigma_l(1-\pi_i^*)^2}} \right) I(X_{\omega} > a)^{1-\pi_i^*}
\]

\[
= \prod_{l, X_{\omega} > a^*} \sum_{l, X_{\omega} \geq a} \frac{\hat{d}_l(a)}{\Sigma_l(1-\pi_i^*) (X_{\omega} - a) \Sigma_l(1-\pi_i^*)^2}
\]

\[
\exp \{ P(a) \} = \left\{ \prod_{l, X_{\omega} > a^*} \frac{\sum (1-\pi_i^*)}{\Sigma_l(1-\pi_i^*) (X_{\omega} - a) \Sigma_l(1-\pi_i^*)^2} \right\} \times \left\{ \prod_{l, X_{\omega} > a^*} e^{-\frac{\hat{d}_l(a) [X_{\omega} - a]}{\Sigma_l(1-\pi_i^*) (X_{\omega} - a) \Sigma_l(1-\pi_i^*)^2}} \right\}
\]

This puts a constant does not depend on \( \pi_i^* \)
Thus
\[ \exp \{ p(a) \} \propto \sum_{j} \frac{1}{\sum (1 - \pi_{j})} \left( x_{j} - a \right) \int_{x_{j} > a} \left( I \right) \]

What does this function look like? (As a function of a, go on a remember).

Now \[ \hat{a}_{k+1} = \arg \max \exp \{ p(a) \} \]

Further analysis:

It is not clear how the curve behaves. This will depend on. However, it will be zero after \( x_{cu} \), which is the smallest \( x_{c} \) greater than \( a \). Thus \( \hat{a}_{k+1} \) must be < \( x_{cu} \).

How this curve looks like, I am not sure.
So this is the algorithm.

How sensitive it is to ethical values, is not clear.

(iii) It is not clear what the rate of convergence is. However we do know (see notes), if $X$ has the shifted exponential distribution (not mixture), i.e., \[ \frac{1}{\lambda} \cdot e^{-\frac{x-a}{\lambda}} I(x\geq a) \]. The the estimator of $a$ is $\hat{a} = \min X_i$. Furthermore $\text{var}(\hat{a}) = O\left(\frac{1}{n^2}\right)$ [very fast], whereas $\text{var}(d) = O\left(\frac{1}{n}\right)$ → slower.

Does a similar situation arise here?

Note: For most estimates it is impossible to calculate the variance and one uses the inverse Fisher information etc. These are not exactly the variance of the estimates. Similar to how when we use the word biased we don't always consider the expectation of an estimator and only consider the probabilistic limit.