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 $E_2 = n_4 \times n_2/N$.

	C_1	C_2	Subtotal
R_1	O_1	O_2	n_1
R_2	O_3	O_4	n_2
Subtotal	n_3	n_4	N

State all results you use. (hint: It may be useful to use the Taylor approximation $x \log(x/y) \approx (x - y) + \frac{1}{2}(x - y)^2/y$). [10]

(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, λ_1, λ_2 and a are unknown.

[15]

(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 (ie. Define the variable δ_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).

Stat 613 2012 Midterm Solution 1

1) Here we have the 2×2 table:

		Y: Y_1	Y_2	
X	X_1	p_1	p_2	q_2
	X_2	p_3	p_4	$1-q_2$
		q_1	$(1-q_1)$	

H_0: There is no association in which case

$p_1 = q_1 q_2$, $p_2 = (1-q_1)q_2$

$p_3 = q_1(1-q_2)$, $p_4 = (1-q_1)(1-q_2)$

H_A: There is an association. in which case

$p_1+p_2+p_3+p_4 = 1$.

We observe

	Y_1	Y_2
X_1	n_{11}	n_{12}
X_2	n_{21}	n_{22}

$n = n_{11}+n_{12}+n_{21}+n_{22}$

In otherwords ~~the number~~ n_{11} = number of people out of n who are in the (Y_1, X_1) category.

etc.

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

$$\mathcal{L}(\pi) = \binom{n}{n_{11}\, n_{12}\, n_{21}\, n_{22}} \pi_{11}^{n_{11}} \pi_{12}^{n_{21}} \pi_{12}^{n_{12}} \pi_{22}^{n_{22}}.$$

Now we want to evaluate estimates of $\pi_{11}, \ldots \pi_{22}$ under both the null and the alternative.

First the alternative (it is easier!).

$$\log \mathcal{L}_n^E(\pi) = \log \binom{n}{n_{11} \ldots n_{22}} + n_{11} \log \pi_{11} + n_{12} \log \pi_{12}$$
$$+ n_{21} \log \pi_{21} + n_{22} \log (1 - \pi_{11} - \pi_{12} - \pi_{21})$$

Diff. wrt $\pi_{11}, \pi_{21}, \pi_{12}$ we see that mle estimates of $p_1, \ldots, p_4$ are

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

(quite obvious!).

Now for the null:

$$\log \mathcal{L}(\pi) = \log \binom{n}{n_{11}, n_{21}, n_{12}, n_{22}} + n_{11} \log q_1 q_2 + n_{21} \log q_1 (1 - q_2)$$
$$+ n_{12} \log (1 - q_1) q_2 + n_{22} \log (1 - q_1)(1 - q_2).$$

Differentiating wrt q_1 and q_2 and equating to zero gives

$$\hat{q}_1 = \frac{n_{11} + n_{21}}{n} \qquad \hat{q}_2 = \frac{n_{11} + n_{12}}{n}$$

To do the LRT we substitute these estimates into the Full and restricted likelihoods.

$$2\{\mathcal{L}_F(\overset{H_A}{\pi}) - \mathcal{L}_R(\overset{H_0}{\pi})\}$$

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

After lots of cancellations and using the fact that

$O_{11} = n_{11}$ $O_{12} = n_{12}$ $O_{21} = n_{21}$ $O_{22} = n_{22}$

$E_{11} = \left(\frac{n_{11}+n_{12}}{n}\right) \times (n_{11}+n_{21})$ ..... $E_{22} = \frac{(n_{22}+n_{21})(n_{22}+n_{12})}{n}$

we have

$$2\{\mathcal{L}_F(\pi) - \mathcal{L}_R(\pi)\} = 2\Big\{\cancel{\log}\, O_{11}\log\frac{O_{11}}{E_{11}} + O_{12}\log\frac{O_{12}}{E_{12}}$$
$$+ O_{21}\log\frac{O_{21}}{E_{21}} + O_{22}\log\frac{O_{22}}{E_{22}}\Big\}$$

Under the null hypothesis we can see that

the above $\xrightarrow{D} \chi^2_{3-2=1}$.

To obtain the form of the χ^2-test for independence

④

we use the ~~expression~~ approximation $O \log \frac{O}{E} \approx O - E + \frac{1}{2} \frac{(O-E)^2}{E}$

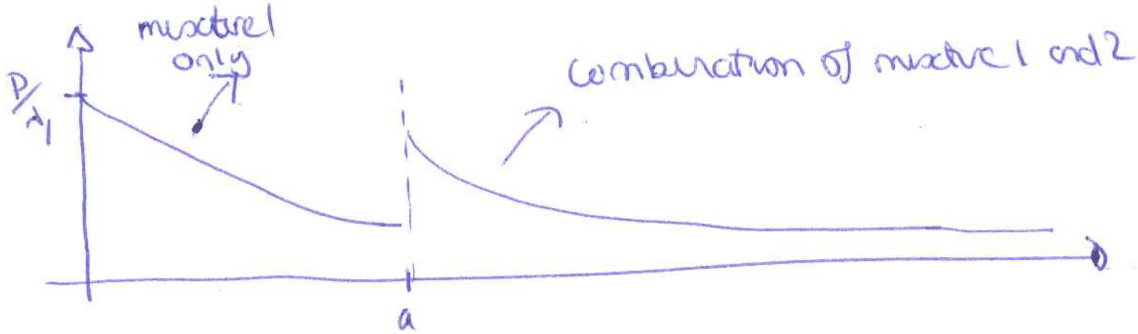
This gives $$\sum_{i,j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \to \chi^2_1 .$$

ℓ

2) Mixture model:

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

where $\theta = (p, \lambda_1, \lambda_2, a)$ are unknown parameters.



Define the variable $\delta = \begin{cases} 0 \\ 1 \end{cases}$, where $P(X|\delta=0) = \lambda_1^{-1} e^{-x/\lambda_1}$

$$P(x|\delta=1) = \lambda_2^{-1} e^{-x/\lambda_2}$$

$$P(\delta=0) = p \text{ and } P(\delta=1) = 1-p.$$

ii) Let us consider what happens to δ when $X<a$ and $X \geq a$.

If $X<a$ it is clear that we are in mixture 1 hence

$$P(\delta=0 \mid X=x) = 1 \text{ and } P(\delta=1 \mid X=x) = 0$$

If $X \geq a$, then δ can be either in mixture 1 or 2.

$$P(\delta=0 \mid X=x) = \frac{p\lambda_1^{-1} e^{-x/\lambda_1}}{p\lambda_1^{-1} e^{-x/\lambda_1} + (1-p)\lambda_2^{-1} e^{-(x-a)/\lambda_2}}$$

$$P(\delta=1 \mid X=x) = \frac{(1-p)\lambda_2^{-1} e^{-(x-a)/\lambda_2}}{p\lambda_1^{-1} e^{-x/\lambda_1} + (1-p)\lambda_2^{-1} e^{-(x-a)/\lambda_2}}$$

(ii) Now to construct the EM-algorithm.

At the kth step we have the parameters

$a_k^*, \alpha_{1k}^*, \alpha_{2k}^*, p_k^*$

and we want to obtain estimate of the parameters

$a_{k+1}^*, \alpha_{1k+1}^*, \alpha_{2k+1}^*, p_{k+1}^*$ at the next stage.

unobserved

The 'full' loglikelihood is the likelihood of $\{(X_i, \delta_i)\}$

$$\mathcal{L}_c(\theta) = \sum_i (1-\delta_i)\{-\log \alpha_1 - \alpha_1^{-1} x_i\}$$
$$+ \sum \delta_i \{-\log \alpha_2 - \alpha_2^{-1}(x_i - a) + \log I(x_i > a)\}.$$
$$+ \sum_i (1-\delta_i)\log p + \sum \delta_i \log(1-p)$$

Of course since δ_i is not observed the above likelihood cannot be maximised. Instead we look at the best approximation of $\mathcal{L}_c(\theta)$ given what we do observe

$$Q(\theta, \theta^*) = \mathbb{E}(\mathcal{L}_c(\theta) \mid \underline{X}, \theta^*)$$

$$= \sum_i \mathbb{E}(1-\delta_i \mid X_i, \theta^*)\{-\log \alpha_1 - \alpha_1^{-1} x_i\}$$
$$+ \sum_i \mathbb{E}(\delta_i \mid X_i, \theta^*)\{-\log \alpha_2 - \alpha_2^{-1}(X_i - a) + \log I(X_i > a)\}$$
$$+ \sum_i \mathbb{E}(1-\delta_i \mid X_i^*, \theta^*)\log p + \sum_i \mathbb{E}(\delta_i \mid X_i, \theta^*)\log(1-p).$$

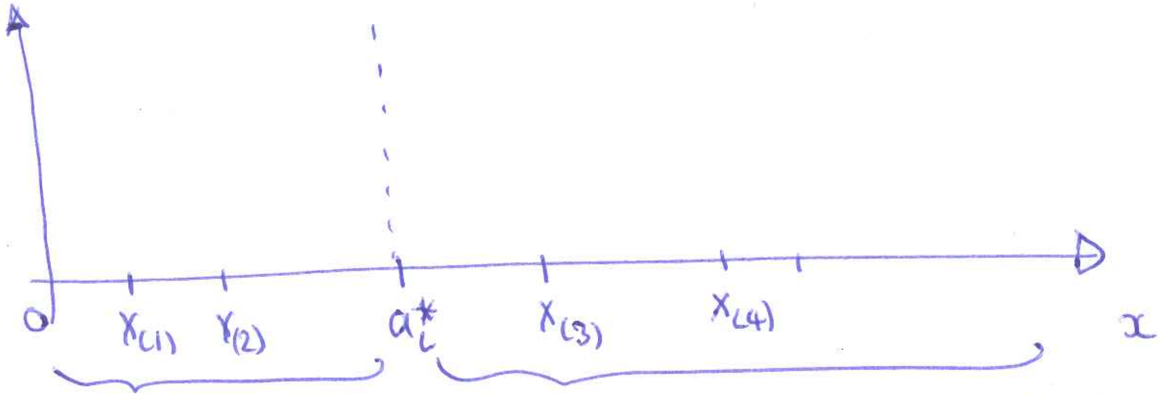
Let $\pi_i^* = P(\delta_i = 0 \mid X_i = x, \theta^*)$

$1 - \pi_i^* = P(\delta_i = 1 \mid X_i = x, \theta^*)$

now we see from part (i) that the value of π_i^* depends on where x lies with respect to a_i^* (the shift estimator at the previous iteration).

If $x < a_i^*$ then $\delta_i = 0$, hence $\pi_i^* = 1$ and $1 - \pi_i^* = 0$

If $x \geq a_i^*$, then π_i^* and $(1-\pi_i^*)$ will take different values.



must be in estimated mixture 1

can be in either mixture.

So we want to maximise

$$Q(\theta, \theta^*) = \sum \pi_i^* \{-\log \lambda_1 - \lambda_1^{-1} X_i\} + \sum (1-\pi_i^*)\{-\log \lambda_2 - \lambda_2 (X_i - a) + \log I(X_i > a)\} + \sum \pi_i^* \log p + \sum (1-\pi_i^*) \log(1-p).$$

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

It is straightforward to show that

$$\hat{\lambda}^*_{1\,k+1} = \frac{\sum_i \pi_i^*}{\sum_i \pi_i^* X_i}$$

swap round

$$\hat{p} = \frac{\sum \pi_i^*}{n}$$

sample size

The tricky part is dealing with estimators of a and λ_2.

Here I use profiling. I assume that a is known

If a is 'known', then my estimator of λ_2 is standard (by differenting Q).

The profile estimate of λ_2 is

$$\hat{\lambda}_{2k+1}(a) = \frac{\sum_i (1-\pi_i^*)(X_i - a)}{\sum (1-\pi_i^*)}$$

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

$$Q(\theta, \theta^*)\big|_a = \sum_i \pi_i^* \{-\log \lambda_1 - \lambda_1^{-1} X_i\} + \sum (1-\pi_i^*)\{-\log \lambda_2(a)$$
$$- \lambda_2(X_i - a) + \log I(X_i > a)\}$$
$$+ \sum \pi_i^* \log p + \sum (1-\pi_i^*) \log(1-p)$$

These parts do not contain a

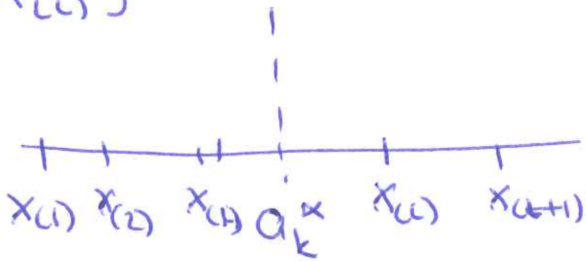
So we consider the only part which matter

$$P(a) = \sum_i (1-\pi_i^*)\{-\log \lambda_2(a) - \lambda_2(a)(X_i - a)$$
$$+ \log I(X_i > a)\}$$

It is difficult to maximise this. So we consider the exponential of this (the likelihood)

$$\exp\{P(a)\} = \prod_{i=1}^{n} \left[\hat{\lambda}_2(a)\, e^{-\hat{\lambda}_2(a)[X_i - a]}\, I(X_i > a) \right]^{1-\pi_i^*}$$

recall that $1-\pi_i^* = 0$ if $X_i < a_k^*$, thus we only consider those X_i where $X_i \geq a_k^*$. To do this we order $\{X_i\} \longrightarrow \{X_{(i)}\}$



$$\exp\{P(a)\} = \prod_{i;\, X_{(i)} \geq a^*}^{n} \left\{ \hat{\lambda}_2(a)^{-1} e^{-\hat{\lambda}_2(a)^{-1}[X_{(i)} - a]}\, I(X_{(i)} > a) \right\}^{1-\pi_i^*}$$

replace these by $\dfrac{\sum_i (1-\pi_i^*)(X_i - a)}{\sum_i (1-\pi_i^*)}$

$$= \cancel{\left\{ \prod_{i;\, X_{(i)} > a}^{n} \frac{\sum_{i;\, X_{(i)} > a} (1-\pi_i^*)}{\sum_{i;\, X_i \geq a} (1-\pi_i^*)(X_{(i)} - a)} \right.}$$

$$\exp[P(a)] = \left\{ \prod_{i;\, X_{(i)} > a} \frac{\sum_i' (1-\pi_i^*)}{\sum'_{i;\, X_{(i)} > a} (1-\pi_i^*)(X_{(i)} - a)} \right\} \times \left\{ \prod_{i;\, X_{(i)} > a^*} I(X_{(i)} > a) \right\}$$

$$\times \underbrace{\left\{ \prod_{i;\, X_{(i)} > a^*} e^{-\lambda_2(a)^{-1}(1-\pi_i^*)[X_{(i)} > a]} \right\}}$$

$$e^{-\sum_{i;\, X_{(i)} > a^*} (1-\pi_i^*)}$$

This part a constant does not depend on a.

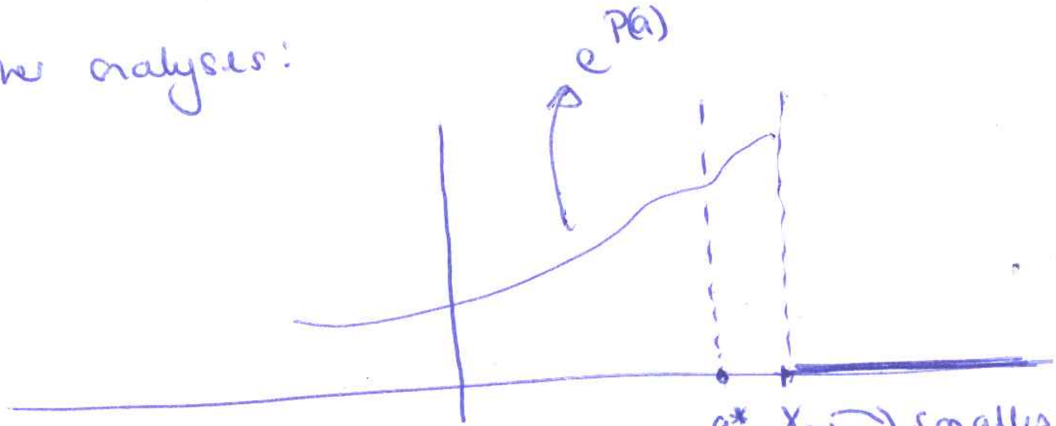
Thus ⑥

$$\exp\{P(a)\} \propto \left\{ \prod_{i;\, X_{(i)} > a^*} \frac{\sum (1-\pi_j^*)}{\sum (1-\pi_j^*)(x_{(j)}-a)} \right\} \left\{ \prod_{i;\, X_{(i)} > a^*} I(X_{(i)} > a) \right\}$$

What does this function look like (its a function of a remember).

Now $\hat{a}^*_{K+1} = \arg\max \exp\{P(a)\}$.

Further analyses:



$X_{(i)} \to$ smallest X_i greater than a^*

It is not clear how the curve behave, this will depend on.

However it will be zero after $X_{(i)}$ [which is the smallest X_i greater than a]. Thus $\hat{a}^*_{K+1}$ must be $< X_{(i)}$.

~~How sensitive~~

How this curve looks like, I am not sure.

So this is the algorithm.

How sensitive it is to initial 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), $\frac{1}{d} e^{-\frac{1}{d}(x-a)} I(x \geq a)$. 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 $\text{var}[\hat{d}] = O\left(\frac{1}{n}\right) \rightarrow$ slower.

Does a similar situation arise here?

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