UNIVERSITY OF KENT AT CANTERBURY

FACULTY OF SCIENCE, TECHNOLOGY AND MEDICAL STUDIES

PART II EXAMINATION

APPLIED STOCHASTIC MODELLING AND DATA ANALYSIS

Thursday, 2nd May 2002: 09.30 a.m. to 11.30 a.m.

This paper contains FIVE *questions.*
Candidates should not attempt more than THREE *questions.*
Copies of the New Cambridge Elementary Statistical Tables
are provided.
Approved calculators may be used.
Each question will be marked out of 40.

1. (a) Discuss model construction and fitting for *either* fecundity data *or* polyspermy data, both of which have been encountered in the lecture course. [10 marks]

(b) For the period 1920-1979. lengths of 'very warm spells' (periods of three or more consecutive days with maximum temperature more than $4°C$ above the long-term mean) have been recorded at Edgbaston, Birmingham. The results are as follows:

Length of spell (days):	3	4	5	6	7	8	9	10	11	12	> 12
Number of spells:	149	78	49	20	17	7	4	2	4	3	1

The total number of warm spells is 334, and the single spell of > 12 days actually lasted 17 days.

(i) Explain the assumptions you have to make in order to model warm spell data using a geometric variable X, with probability function,

$$p_X(j) = p(1-p)^j, \quad j = 0, 1, 2, \ldots, \quad 0 < p < 1.$$

[6 marks]

(ii) Define Y as the random variable that results when X is truncated, so that only values of $X \geq 3$ are recorded. Show that Y has probability function,

$$p_Y(j) = p(1-p)^{j-3}, \quad j = 3, 4, \ldots$$

[10 marks]

(iii) By fitting the probability function of Y to the warm spell data, obtain the maximum likelihood estimate of p, and an appropriate estimate of standard error. Test whether or not the distribution provides a good fit to the data. [14 marks]

2. (a) Explain the δ-method, and provide an example of its use. [10 marks]

(b) The data below summarise the daily mortality in groups of fish subjected to three levels of zinc concentration. Half the fish at each concentration level received one week's acclimatisation to the test aquaria, and half received two weeks' acclimatisation. There were therefore six treatment groups, and 50 fish were randomized to each group, resulting in 300 fish in total in the experiment.

		Acclimatisation time					
		One week			Two weeks		
Day	log zinc concentration	0.205	0.605	0.852	0.205	0.605	0.852
1		0	0	0	0	0	0
2		3	3	2	0	1	3
3		12	17	22	13	21	24
4		11	16	15	8	8	10
5		3	5	7	0	5	4
6		0	1	1	0	0	1
7		0	0	2	0	0	0
8		0	1	0	0	0	0
9, 10		0	0	0	0	0	0

Question continued on opposite page

The following *mixture* model is proposed for the data, in which a fish is classified as either a 'long-term' survivor or a 'short-term' survivor. If $\mathbf{x}$ denotes the two-dimensional vector indicating acclimatisation time and zinc concentration, the probability that a fish with covariate $\mathbf{x}$ is a 'short-term' survivor is:

$$p(\mathbf{x}) = (1 + e^{-\boldsymbol{\beta}'\mathbf{x}})^{-1}.$$

'Long-term' survivors can be assumed not to die during the course of the experiment. For a short-term survivor the time of death, measured from the start of the experiment, will have a Weibull distribution, with probability density function

$$f(t|\mathbf{x}) = \delta\lambda(\lambda t)^{\delta-1}\exp\{-(\lambda t)^{\delta}\}, \quad t \geq 0,$$

where $\lambda = \exp(-\boldsymbol{\gamma}'\mathbf{x})$. This model contains five parameters $\theta = (\delta, \boldsymbol{\gamma}, \boldsymbol{\beta})$, the vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ containing two elements each.

(i) Derive an expression for the Weibull survivor function. Write down an expression, involving $p(\mathbf{x})$ and the Weibull survivor function, for the probability that a fish with covariates $\mathbf{x}$ does not die during the course of the experiment. [7 marks]

(ii) Hence write down a general expression for the likelihood for a set of data of the form shown above. (You are not expected to incorporate the particular illustrative data above.) Indicate, in outline only, how you would proceed to obtain the maximum-likelihood estimate of θ and associated measures of standard error. [9 marks]

(iii) Draw conclusions from the following estimates obtained for the above data: $\widehat{\delta} = 3.47$.

Covariate	$\widehat{\beta}$	Estimated standard error	$\widehat{\gamma}$	Estimated standard error
Acclimatisation	-0.94	0.30	-0.13	0.04
Concentration	3.59	0.56	0.17	0.08

[10 marks]

(iv) Discuss any further analyses which you would carry out. [4 marks]

3. (a) Compare and contrast the simplex method and simulated annealing as methods for likelihood maximisation. You may find that simple sketch graphs would enhance your solution.
[8 marks]

(b) A MATLAB file with the name, likelihood.m contains code specifying a two-parameter likelihood surface. The MATLAB program below is designed to maximise the likelihood.

(i) Explain how the program works, and fill-in the gaps indicated by the two *s.

$xo = [.7\ .5]'$;

$g = [1,\ 1]$;

while norm $(g) > 0.00001$

$g =$ grad ('likelihood', xo);

$h =$ hessian ('likelihood', xo);

$xn = xo -$ *

*

end
[8 marks]

(ii) Explain why it is useful to display the eigenvalues of h evaluated at the maximum-likelihood estimates.
[3 marks]

(c) An electrical component has failure time given by the gamma, $\Gamma(2, \lambda)$, pdf:

$$f(t) = \lambda^2 t e^{-\lambda t} \quad \text{for} \quad t > 0, \quad f(t) = 0 \quad \text{for} \quad t \leq 0,$$

where $\lambda > 0$.
The corresponding cdf is $F(t) = 1 - e^{-\lambda t}(1 + \lambda t)$ for $t > 0$,

$= 0$ for $t \leq 0$.

For m components the failure times are known exactly, and given by $\{t_i,\ \ 1 \leq i \leq m\}$. The failure times of n components are all right-truncated, with truncation times, $\{t_i,\ \ m+1 \leq i \leq m+n\}$.

(i) Show that apart from an additive constant, the log-likelihood of λ may be expressed as

$$\ell(\lambda) = 2m \log \lambda - \lambda \sum_{i=1}^{m+n} t_i + \sum_{i=m+1}^{m+n} \log(1 + \lambda t_i).$$

[10 marks]

(ii) The Newton-Raphson iterative method is to be used to find the maximum likelihood estimator of λ. Let $\lambda^{(r)}$ be the value of λ at the rth iteration. Show that

$$\lambda^{(r+1)} = \lambda^{(r)} \left[1 + \frac{2m - \lambda^{(r)} \sum_{i=1}^{m+n} t_i + \lambda^{(r)} \sum_{i=m+1}^{m+n} z_i^{(r)}}{2m + (\lambda^{(r)})^2 \sum_{i=m+1}^{m+n} (z_i^{(r)})^2} \right],$$

where

$$z_i^{(r)} = \frac{t_i}{1 + \lambda^{(r)} t_i}.$$

[11 marks]

4. (a) Describe the EM algorithm for likelihood maximisation. [8 marks]

(b) The EM algorithm is to be used as an alternative to the Newton-Raphson method for maximising the likelihood of question 3(c), by imputing the values of the censored observations.

(i) Show that the E-step of the algorithm involves the term

$$E(T_i|T_i > t_i,\ \lambda^{(r)}) \quad \text{for} \quad m+1 \leq i \leq m+n,$$

where $\lambda^{(r)}$ is the current estimate of λ. [8 marks]

(ii) Prove that

$$E(T_i|T_i > t_i,\ \lambda^{(r)}) = \frac{2}{\lambda^{(r)}} + \frac{\lambda^{(r)} t_i^2}{1+\lambda^{(r)} t_i}. \qquad (*)$$

[You may assume that for $\theta > 0$

$$\int_s^\infty \frac{\theta^n y^{n-1} e^{-\theta y}}{(n-1)!}\ dy = \sum_{j=0}^{n-1} \frac{s^j \theta^j e^{-\theta s}}{j!}. \,]$$

[12 marks]

(iii) Show how the M-step is used to produce the next estimate of λ, denoted by $\lambda^{(r+1)}$. Give an intuitive explanation of your result. [8 marks]

(iv) Suppose that you had been unable to derive the formula for the expectation in $(*)$ above. How else could you have calculated $E(T_i|T_i > t_i,\ \lambda^{(r)})$? [4 marks]

5. Consider a long document that consists of $n(> 2)$ pages. It is known that one secretary started typing the document but it was completed by another. However, exactly where in the document the second secretary took over from the first is unknown. It is proposed to model this by supposing that the first r pages were typed by the first secretary and the final $n - r$ pages by the second where r is an unknown integer between 1 and $n - 1$ inclusive.

The two secretaries make typing errors at different rates θ_1 and θ_2 per page. Let X represent the number of errors on the ith page and assume that X has a Poisson distribution, $i = 1, 2, \ldots, n$.

Then for $1 \leq i \leq r$

$$P(X_i = x_i) = \frac{\theta_1^{x_i} e^{-\theta_1}}{x_i!} \quad x_i = 0, 1, 2, \ldots$$

and for $r + 1 \leq i \leq n$

$$P(X_i = x_i) = \frac{\theta_2^{x_i} e^{-\theta_2}}{x_i!} \quad x_i = 0, 1, 2, \ldots$$

It is proposed to use a Bayesian approach to make inferences about r. Prior information about θ_i is modelled by a gamma, $\Gamma(\alpha_i, \beta_i)$ distribution. Little is known about when the secretaries changed so that the prior for r is taken to be uniform over the integers $1, 2, \ldots, n - 1$.

(i) Show that the joint posterior distribution of θ_1, θ_2 and r is proportional to

$$\theta_1^{S_r + \alpha_1 - 1} \theta_2^{T_r + \alpha_2 - 1} e^{-\theta_1(\beta_1 + r)} e^{-\theta_2(\beta_2 + n - r)}, \quad \text{for } 1 \leq r \leq n - 1,$$

where

$$S_r = \sum_{i=1}^{r} x_i \quad \text{and} \quad T_r = \sum_{i=r+1}^{n} x_i.$$

[11 marks]

(ii) Hence identify the conditional posterior distributions of each of the three parameters, conditional on the remaining two. [9 marks]

(iii) Outline how you would sample from the full conditional posterior probability function of r. [8 marks]

(iv) Explain how you would use Gibbs sampling in this situation. [5 marks]

(v) Describe how you would estimate the posterior marginal probability function of r, its mean and variance. [4 marks]

(vi) How would you adapt your algorithm if it were known that the second secretary was more error prone than the first? [3 marks]

[Note: If Y has a $\Gamma(\alpha, \beta)$ distribution then its probability density function is given by

$$f(y) = \begin{cases} \dfrac{\beta^\alpha y^{\alpha-1} e^{-\beta y}}{\Gamma(\alpha)} & 0 < y, \\ 0 & y \leq 0. \end{cases}]$$