Data provided: Neave's Tables

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2012–2013**

MAS472 Computational Inference **2 hours**

Restricted Open Book Examination.
Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator which conforms to University regulations.
*Marks will be awarded for your best **three** answers. Total marks 90.*

Please leave this exam paper on your desk
Do not remove it from the hall

Registration number from U-Card (9 digits)
to be completed by student

Blank

1 (i) Let Y be a truncated exponential random variable with density function

$$f_Y(y) = \begin{cases} K \exp\left(-\frac{y}{2}\right) & \text{for } 0 \leq y \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the value of K. ***(3 marks)***

(b) Explain how to generate a random value of Y given a uniform random number using the inversion method. You may leave you answer in terms of K. ***(6 marks)***

(ii) Suppose we wish to sample from a standard normal random variable, truncated to $(0, 2)$ with density function

$$f_X(x) = \begin{cases} \frac{A}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) & \text{for } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the value of A. ***(2 marks)***

(b) Explain how we could use rejection sampling to sample from the distribution of X using the truncated exponential density function of Y defined in part (i). ***(6 marks)***

(c) Calculate the expected number of random variables drawn from Y required to produce one random value of X.

Hint: $\dfrac{2(1-e^{-1})e^{\frac{1}{8}}}{\sqrt{2\pi}(\Phi(2)-\frac{1}{2})} \approx 1.198$ ***(2 marks)***

(d) Compare the efficiency of this method to generate values of X with an alternative rejection sampling scheme using the non-truncated standard normal density as an envelope function. ***(3 marks)***

(iii) Suppose importance sampling, using a normal approximation as the importance density, is to be used to sample from a $Beta(4, 3)$ distribution with density

$$f_\theta(\theta) = \begin{cases} 60\,\theta^3(1-\theta)^2 & \text{for } \theta \in (0, 1), \\ 0 & \text{otherwise.} \end{cases}$$

By considering a Taylor series expansion of $\log f(\theta)$ about the mode of θ, obtain the mean and variance of the importance density. ***(8 marks)***

2 (i) It is possible to rescale the standard Cauchy distribution to give it a `location` x_0 and a `scale` γ. The rescaled density is then

$$f(x) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x-x_0}{\gamma}\right)^2\right]}$$

In a physics experiment, the energy of an unstable quantum state is recorded on n occasions and stored in a vector `x`. This data is known to have extremely heavy tails. The following R analysis is then applied

```
> theta.hat <- median(x)
> scale.hat <- IQR(x)/2
>
> n <- length(x)
> M <- 1000
> z.star <- rep(NA, M)
>
> for(i in 1:M) {
+   x.star <- rcauchy(n, location = theta.hat, scale = scale.hat)
+   z.star[i] <- median(x.star)
+ }
>
> quantile(z.star, c(0.005, 0.995))
    0.5%    99.5%
1.105624 3.638157
```

(a) Explain carefully the procedure that has been performed here, and state what the output in the last line represents ***(8 marks)***

(b) Why do you think we are using `median` and `IQR` as the parameter estimates instead of `mean` and `sd`? ***(2 marks)***

(c) Comment on the accuracy of the procedure in relation to the size of n and M ***(4 marks)***

(d) How would you find a 90% confidence interval for the location using the non-parametric bootstrap? ***(6 marks)***

2 (continued)

(ii) Suppose we have 30 observations (x_i, y_i) thought to arise from a model

$$y_i = \alpha x_i + \epsilon_i$$

where $\log \epsilon_i \sim N(0, \sigma^2)$. Let $\hat{\alpha}$ denote the least squares estimator of α. We wish to test a null hypothesis $H_0 : \alpha = 0$ against a two-sided alternative.

(a) Explain why a conventional parametric test comparing the test statistic

$$T = \frac{|\hat{\alpha}|}{\sqrt{Var(\hat{\alpha})}}$$

against a t-distribution is not suitable in this instance. ***(2 marks)***

(b) Explain the following alternative analysis including the conclusion of the investigation. Make sure you explain why the test statistic $T^\star = \left| \sum_{i=1}^{30} x_i y_i \right|$ is being used instead of the T statistic in part a).

```
> t.obs <- abs(sum(x*y))
>
> M <- 1000
> t.star <- rep(NA, M)
>
> for(i in 1:M) {
+  y.resamp <- sample(y, replace = FALSE)
+  t.star[i] <- abs(sum(x*y.resamp))
+ }
>
> mean(t.star > t.obs)
[1] 0.031
```

(8 marks)

3 Suppose X_1 and X_2 are independent random variables with a $N(\mu_1, \sigma^2)$ distribution while Y_1 and Y_2 are independent random variables with a $N(\mu_2, \sigma^2)$ distribution. The common variance σ^2 is presumed known. Let us define $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$.

(i) Use the factorisation theorem to show that $(X_1 + X_2)$ is a sufficient statistic for μ_1. What is a sufficient statistic for μ_2? ***(6 marks)***

(ii) Show that the complete data log-likelihood

$$\log\{f(X, Y|\mu_1, \mu_2)\} = M - \frac{1}{2\sigma^2}\left[x_1^2 + x_2^2 + y_1^2 + y_2^2 \right. \\ \left. -2\mu_1(x_1 + x_2) - 2\mu_2(y_1 + y_2) + 2\mu_1^2 + 2\mu_2^2\right],$$

where M is some constant, can be written as

$$l(X, Y|\theta) = \sum_{i=1}^{2} A_i(\theta)B_i(X, Y) + C(X, Y) + D(\theta)$$

where $\theta = (\mu_1, \mu_2)$. Give expressions for $A_1(\theta)$, $A_2(\theta)$, $B_1(X, Y)$, $B_2(X, Y)$ and $D(\theta)$ ***(6 marks)***

(iii) Suppose that only three observations are available: X_1, Y_1 and $Z = X_2 + Y_2$.

(a) Show that

$$X_2|Z \sim N\left(\frac{\mu_1 + z - \mu_2}{2}, \frac{\sigma^2}{2}\right)$$

Hint: Use the fact that

$$f(x_2|Z = z) \propto f_{(X_2,Z)}(x_2, z) = f_{(X_2,Y_2)}(x_2, z - x_2)$$

since $Z = z|X_2 = x_2$ *if and only if* $Y_2 = z - x_2$

(6 marks)

(b) The maximum likelihood estimates of μ_1 and μ_2 given X_1, Y_1 and Z are to be obtained using the EM algorithm. Let μ_1^{old} and μ_2^{old} denote the current estimates of $\hat{\mu}_1$ and $\hat{\mu}_2$. By maximising

$$Q(\mu_1, \mu_2|\mu_1^{old}, \mu_2^{old}) = E\left[\log\{f(X, Y|\mu_1, \mu_2)\}|X_1, Y_1, Z, \mu_1^{old}, \mu_2^{old}\right]$$

with respect to μ_1 and μ_2, derive improved estimates of $\hat{\mu}_1$ and $\hat{\mu}_2$. This can be done as follows:

- Use your result from part (a) to find

$$E\left[\log\{f(X, Y|\mu_1, \mu_2)\}|X_1, Y_1, Z, \mu_1^{old}, \mu_2^{old}\right].$$

- Solve

$$\frac{\partial Q}{\partial \mu_i} = 0$$

for $i = 1, 2$.

(12 marks)

4 (i) A financial model has been developed to predict the cost of an insurance policy. The model has two uncertain inputs:

θ : The cost of reinsurance

ϕ : The time for the first claim to be made

The model is a deterministic function $p(\theta, \phi)$ while θ and ϕ have densities

$$\theta : f(\theta) = \frac{5^{20}\theta^{19}\exp(-5\theta)}{\Gamma(20)} \qquad x > 0 \quad (\text{i.e.} \sim Gamma(20, 5)$$

$$\phi : g(\phi) = \frac{1}{30}e^{-\phi/30} \qquad x > 0 \quad (\text{i.e.} \sim Exp(Rate = 1/30))$$

The user wishes to determine the expected price C taking into account the uncertainty in the input parameters.

(a) Write down a formula for C in terms of $p(\theta, \phi)$, $f(\theta)$ and $g(\phi)$.
(2 marks)

(b) The model can be implemented in R via a user defined function `Price(theta, phi)`. Some output from an R analysis is shown below:

```
> theta <- rgamma(1000, 20, 5)
> phi <- rexp(1000, 1/10)
> C <- Price(theta, phi)
> mean(C)
[1] 923.2803
> var(C)
[1] 177026.3
```

Give the estimated expected price, and a 95% confidence interval.
(3 marks)

(c) How many sampled pairs would be needed such that the 95% confidence interval for C is no wider than 10? ***(2 marks)***

(d) An alternative distribution is proposed for θ: the $Exp(Rate = 0.25)$ distribution. Suppose the original analysis generated output values $C_1, \ldots, C_{1000}$ from input values $\theta_1, \ldots, \theta_{1000}$ and $\phi_1, \ldots, \phi_{1000}$. Give a formula for the Monte Carlo estimate of C corresponding to the new distribution of θ in terms of $C_1, \ldots, C_{1000}$ and $\theta_1, \ldots, \theta_{1000}$ which could be calculated without doing any further evaluations of the function `Price`. ***(4 marks)***

(e) How would you find a 95% confidence interval for C in this case?
(2 marks)

4 (continued)

(ii) The load on 10 web servers is measured and observed to be $x_1, \ldots, x_{10}$. A $Gamma(a, b)$ with density:

$$f(x) = \frac{b^a x^{a-1} \exp(-bx)}{\Gamma(a)} \quad x > 0$$

is believed to fit the data.

(a) Derive the profile log-likelihood for a ***(9 marks)***

(b) For our observed dataset of the load on 10 web servers, we find that $\sum_{i=1}^{10} x_i = 6.09$ and $\sum_{i=1}^{10} \log x_i = -6.212$. The profile likelihood function for a is maximised at $a = 4.159$. Give the maximum likelihood estimate for b. ***(2 marks)***

(c) Use the profile deviance function to test the null hypothesis that $a = 4$
Note: $\Gamma(4) = 6$ *and* $\Gamma(4.159) = 7.352$. ***(6 marks)***

End of Question Paper