2017 Examination Paper for STAT3004

Answer ALL questions. Section A carries 40% of the total marks and Section B carries 60%. The relative weights attached to each question are as follows: A1 (20), A2 (15), A3 (5), B1 (30), B2 (30). The numbers in square brackets indicate the relative weights attached to each part question.

Formula Sheet

- If X has a Beta(α, β) distribution then:

$$p(X|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} X^{\alpha-1}(1-X)^{\beta-1}, \quad 0 \le X \le 1$$

  where $B(\alpha,\beta)$ is the Beta function.

  The relationship between the parameters α, β and the mean/variance of X are:

$$E[X] = \frac{\alpha}{\alpha+\beta}, \quad Var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$\alpha = \left(\frac{1-E[X]}{Var[X]} - \frac{1}{E[X]}\right)[E[X]]^2, \quad \beta = \alpha\left(\frac{1}{E[X]} - 1\right)$$

- If X has a Binomial(N, θ) distribution then:

$$p(X = k|N,\theta) = \binom{N}{k}\theta^k(1-\theta)^{N-k}, \quad k = 0, 1, 2, \ldots, N$$

- If X has a Geometric(θ) distribution then:

$$p(X = k|\theta) = (1-\theta)^{k-1}\theta, \quad k = 0, 1, 2, \ldots$$

- If X has a Poisson(θ) distribution then:

$$p(X = k|\theta) = \frac{\theta^k e^{-\theta}}{k!}, \quad k = 0, 1, 2, \ldots$$

- If X has a Normal Distribution $N(\mu, \sigma^2)$ then:

$$p(X|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X-\mu)^2/2\sigma^2}$$

- If X has an Inverse-Gamma distribution $IG(\alpha, \beta)$ then:

$$p(X|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} X^{-\alpha-1} e^{-\beta/X}, \quad X > 0$$

- if X has a Generalized Pareto Distribution with threshold u and parameters k and σ, then the corresponding cumulative distribution function is:

$$p(X \leq D|u,k,\sigma) = 1 - \left(1 - k\frac{D-u}{\sigma}\right)^{(1/k)}, \quad X > u$$

The method of moments estimates of k and σ based on a sample $X_1, \ldots, X_n$ where all X_i are greater than u are:

$$\hat{\sigma} = \bar{X}'(\bar{X}'^2/s^2 + 1)/2$$
$$\hat{k} = (\bar{X}'^2/s^2 - 1)/2$$

where $\bar{X}'$ and s^2 denote the sample mean and variance:

$$\bar{X}' = \frac{1}{n}\sum_{i=1}^{n} X_i', \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i' - \bar{X}')^2$$

for the transformed observations $X_i' = X_i - u$.

- Given a distribution $p(Y|\theta)$ and observations $Y = (Y_1, \ldots, Y_n)$, the BIC approximation to the marginal likelihood used in this module is given by

$$\log p(Y|\hat{\theta}) - \frac{k}{2}\log(n)$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ and k is the number of parameters in θ

Section A

A1 A large community is being surveyed to determine whether they support building a flood defence system. Let θ by the unknown proportion of people who want to build the system. N people are surveyed, of whom Y answer that they support the system. The responses are assumed to be independent, and can hence be modelled by a Binomial distribution.

a) Give the definition of a conjugate prior, and explain why it is important to Bayesian statistics . [3]

b) Explain why the Beta distribution is the conjugate prior in this case. [3]

c) Before conducting the survey, you become aware of similar surveys that have been carried out in communities with similar demographic characteristics. In 4 such communities, the proportions of people supporting the system were $0.83, 0.71, 0.62, 0.54$. The sample size in these previous surveys is large enough that these proportions can be considered to be known exactly. Explain why this information might justify the use of a Beta(8.88, 4.27) prior for the current survey. What assumptions do you need to make? [6]

d) You find that 65% of the 1,000 people surveyed support building the system. Find the resulting posterior distribution, using the prior from part c). [3]

e) Suppose another M people are surveyed from the same community. Derive an expression for the predictive distribution $p(\tilde{Y}|Y)$ for predicting the number of people $\tilde{Y}$ who will say they support the system, given the above data and prior. Leave your answer as an expression involving Beta functions. [5]

A2 A village is to be classified as either low, medium, or high risk of suffering hurricane damage. Let θ denote the true risk status of the village, where $\theta = 0$ is low risk, $\theta = 1$ is medium risk, and $\theta = 2$ is high risk. You observe a random variable Y that represents the time between the last 2 major hurricanes to hit the village. The distribution of Y is:

$$p(Y|\theta = 0) = Exponential(0.5)$$
$$p(Y|\theta = 1) = Exponential(1)$$
$$p(Y|\theta = 2) = Exponential(2)$$

Your task is to estimate the value of θ, given you have observed the value $Y = 1.5$. Before observing this data, your prior beliefs are $p(\theta = 0) = 0.3$, $p(\theta = 1) = 0.5$, $p(\theta = 2) = 0.2$. Let actions a_0, a_1, a_2 correspond to classifying the village's risk status as $\theta = 0, 1, 2$ respectively. The loss function is:

	$\theta = 0$	$\theta = 1$	$\theta = 2$
a_0	0	3	7
a_1	3	0	4
a_2	7	4	0

a) Show that the posterior for θ is $p(\theta = 0|Y) = 0.350$, $p(\theta = 1|Y) = 0.551$, and $p(\theta = 2|Y) = 0.098$. [6]

b) Compute the risk of all 3 actions, and hence find the optimal decision. [4]

c) Let θ be an unknown parameter for a continuous probability distribution $p(Y|\theta)$. Based on data Y, the posterior distribution $p(\theta|Y)$ is formed. Explain how decision theory can be used to justify summarising the posterior by a particular single value $\hat{\theta}$, and discuss why different loss functions might be appropriate depending on the context. [5]

A3 a) Write down the form of the GARCH(1,1) model. [2]

b) When performing Bayesian estimation of the GARCH model, it is not generally possible to find a conjugate prior. Name some other techniques which can be used in this case to allow for Bayesian inference. [3]

Section B

B1 The government of an island community wants to model the number of hurricanes that occur in an annual season. One possible approach is to assume that the number of hurricanes comes from a Geometric distribution with unknown parameter θ. For the last few years, the number of hurricanes has been $Y = (Y_1, Y_2, Y_3)$ where:

$$Y_1 = 4, Y_2 = 12, Y_3 = 11$$

a) Using a conjugate improper Beta(0,0) prior, show that the predictive distribution for the number of hurricanes $\tilde{Y}$ in a future year is given by

$$p(\tilde{Y}|Y) = \frac{B(4, 23 + \tilde{Y})}{B(3, 24)}$$

[6]

b) An alternative model is to assume the number of hurricanes are observations from a Poisson distribution. Suppose you wish to compare the fit of the Geometric and Poisson distribution on the above data. Compute the BIC (see formula sheet) of both models and hence state which one gives a better fit.

You may use the fact that the maximum likelihood estimate of the parameter of a Geometric distribution is $n/\sum_{i=1}^{n} Y_i$, and the maximum likelihood estimate of the parameter of a Poisson distribution is $\frac{1}{n}\sum_{i=1}^{3} Y_i$ [6]

c) Comment on the limitations of using the BIC in part b). [2]

d) Since the available data are limited, there are not strong grounds for preferring one distribution to another. Instead, we could use the Generalised Pareto Distribution to model the tails. Give a precise statement of the Pickands-Balkema-de Haan theorem which gives a justification for the GPD approximation. [3]

e) Taking a threshold of $u = 10$, for the Generalised Pareto Distribution with the above data, use the method of moments to estimate the GPD parameters, and hence give a frequentist point prediction for $p(\tilde{Y} > 11.5)$ [10]

f) Discuss why your answer to e) might not be accurate, and suggest how it could be improved [3]

B2 The log-returns of a financial asset are given by:

$$Y_1 = 1.00, Y_2 = -1.00, Y_3 = 2.00, Y_4 = -1.00$$

Assume that the log returns have a $N(0, \sigma^2)$ distribution where σ^2 is unknown and is assigned an Inverse-Gamma$(1, 1)$ prior. A possible change point τ exists in this data.

a) Recalling that $\Gamma(1) = 1$, show that the marginal likelihood of the observations up to and including the change point is given by [6]

$$p(Y_1, \ldots, Y_\tau|\tau) = (2\pi)^{-\tau/2} \frac{\Gamma(1 + \tau/2)}{(1 + \sum_{i=1}^{\tau} Y_i^2/2)^{1+\tau/2}}$$

b) Deduce or derive an equivalent expression for the likelihood of the observations to the right of the change point [4]

c) Consider the following two possible models:

M_0: There is a change point immediately after the second observation Y_2

M_1 : There is a change point immediately after the third observation Y_3

Assuming that both models are equally likely a priori, compute the ratio $p(M_0|Y)/p(M_1|Y)$ as an expression involving Gamma functions (which you do not need to evaluate). Suppose that this expression evaluates to 1.02. Which model would give the better fit to the data, and why? [9]

d) Give the definition of the Value-at-Risk of a financial portfolio, and explain why it is a useful measurement of risk. [3]

e) Suppose a company wants to calculate the p=0.05 Value-at-Risk of a portfolio, given its last 2 years of daily returns. Explain how they could go about computing this quantity. Mention any practical challenges which might arise when computing this quantity using real financial data [8]

END OF PAPER