ST3189

Machine Learning

Suitable for all candidates

Instructions to candidates

This paper contains four questions. Answer **ALL FOUR**. All questions will be given equal weight (25%).

The marks in brackets reflect marks for each question.

Time allowed -	**Reading Time:**	*None*
	Writing Time:	*2 hours*
You are supplied with:		*Graph paper*
You may also use:		*No additional materials*
Calculators:		*Calculators are allowed in this examination*

1. (a) Suppose that $y_i \sim N(\mu, 1)$ for $i = 1, \ldots, n$ and that the y_i's are independent.

   i. Show that the sample mean estimator $\hat{\mu}_1 = \frac{1}{n}\sum_i^n y_i$ is obtained from minimising the least squares criterion [7 marks]

$$\hat{\mu}_1 = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mu)^2,$$

   and that $\hat{\mu}_1$ an unbiased estimator of μ. Also find the variance of $\hat{\mu}_1$.
   Answer: Show that the derivative of $\sum_{i=1}^n (y_i - \mu)^2$ wrt μ is equal to $2\sum_i y_i - 2n\mu$. Setting it equal to 0 and solving then yields $\hat{\mu}_1 = \frac{1}{n}\sum_i^n y_i$. We then get

$$E(\frac{1}{n}\sum_{i}^{n} y_i) = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \mu,$$

   which implies that the estimator is unbiased. For the variance not that

$$\text{var}(\frac{1}{n}\sum_{i}^{n} y_i) = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(y_i) = \frac{1}{n}$$

   ii. Consider adding a penalty term to the least squares criterion, and therefore using the estimator that minimises

$$\hat{\mu}_2 = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mu)^2 + \lambda\mu^2$$

   for the mean, where λ is a non-negative tuning parameter. Derive $\hat{\mu}_2$, find it bias and show that its variance is lower than that of $\hat{\mu}_1$ [7 marks]
   Answer: The derivative w.r.t. μ is $2\sum(y_i - \mu) + 2\lambda\mu$. Setting it to 0 gives

$$\hat{\mu}_2 = \frac{\sum_{i=1}^n y_i}{n + \lambda}.$$

   Then

$$E(\hat{\mu}_2) = \frac{n}{n+\lambda}\mu,$$

$$\text{Bias}(\hat{\mu}_2) = E(\hat{\mu}_2) - \mu = \frac{n}{n+\lambda}\mu - \mu = -\frac{\lambda}{n+\lambda}\mu$$

$$\text{var}\,(\hat{\mu}_2) = \text{var}\left(\frac{\sum_i^n y_i}{n+\lambda}\right) = \frac{1}{(n+\lambda)^2}\sum_{i=1}^{n} \text{var}(y_i) = \frac{n}{(n+\lambda)^2}.$$

   Note that $\text{var}\,(\hat{\mu}_2) < \text{var}(\hat{\mu}_1)$ since $\frac{n}{(n+\lambda)^2} < \frac{1}{n}$ as $\lambda > 0$.

   (b) Consider the multiple linear regression model

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \quad i = 1, \ldots, n, \quad j = 1, dots, p,$$

   where $\beta = (\beta_1, ..., \beta_p)^T$ and $\epsilon = (\epsilon_1, ..., \epsilon_n)^T \sim N(0, \sigma^2 I_n)$.

i. When p is comparable to n, the multicollinearity becomes an issue. Describe the effects of multicollinearity on the estimated coefficients, the associated standard errors and the significance of the coefficients using the ordinary maximum likelihood method. [3 marks]
Answer: The estimated coefficients and the associated standard errors can both become very large, making the coefficients non-significant in the end.

ii. The ridge regression estimate of β can be obtained by minimising a particular expression with respect to β. Write down this expression as well as an alternative formulation of it. [4 marks]
Answer: The expression is

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2,$$

where $\lambda>0$ is a tuning parameter.
It can be shown however, that minimise the above expression is equivalent to minimising

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2, \text{ subject to } \sum_{j=1}^{p}\beta_j^2\le s,$$

where $s>0$ is a tuning parameter.

iii. Explain why ridge regression can potentially correct the problems of multicollinearity. [2 marks]
Answer: It is because the magnitude of the estimated coefficients are restricted by setting $\sum_{j=1}^{p}\beta_j^2\le s$.

iv. Provide an advantage and a disadvantage of ridge regression over the standard linear regression. [2 marks]
Answer: Ridge regression estimates are biased but have lower variance.

2. Let $x = (x_1, \ldots, x_{100})$, with $\sum_i x_i = 20$, be a random sample from the Exponential(λ) distribution with probability density function given by

$$f(x_i|\lambda) = \frac{1}{\lambda} \exp\left(-\frac{x_i}{\lambda}\right), \quad x_i > 0, \quad \lambda > 0.$$

Note that $E(x_i) = \lambda$.

(a) Assign the IGamma$(0.1, 0.1)$ prior to λ and find the corresponding posterior distribution. [5 marks]

Answer: The likelihood can be written as

$$f(x|\theta) = \prod_{i=1}^{100} \frac{1}{\lambda} \exp\left(-\frac{x_i}{\lambda}\right) \lambda^{-n} \exp\left(-\frac{\sum_{i=1}^{n} x_i}{\lambda}\right) = \lambda^{-100} \exp\left(-\frac{20}{\lambda}\right),$$

and the prior is $\pi(\theta) \propto \lambda^{-0.1-1} \exp\left(-\frac{0.1}{\lambda}\right)$. Hence the posterior is

$$\pi(\theta|x) \propto \lambda^{-100} \exp\left(-\frac{20}{\lambda}\right) \lambda^{-0.1-1} \exp\left(-\frac{0.1}{\lambda}\right) = \lambda^{-(100.1)-1} \exp\left(-\frac{20.1}{\lambda}\right)$$

which can be recognised as the IGamma$(100.1, 20.1)$ distribution.

(b) Find the Jeffreys' prior for λ. Which is the corresponding posterior distribution? [6 marks]

Answer: We can write $l(x|\lambda) = \log f(x|\lambda) = -100 \log \lambda - \frac{20}{\lambda}$

$$\frac{\partial}{\partial \lambda} l(x|\lambda) = -\frac{100}{\lambda} + \frac{20}{\lambda^2}, \quad \frac{\partial^2}{\partial \lambda^2} l(x|\lambda) = \frac{100}{\lambda^2} - 2\frac{20}{\lambda^3}$$

$$\begin{aligned} I(\lambda) &= -E\left[\frac{\partial^2}{\partial \lambda^2} l(x|\lambda)\right] = -E\left[\frac{100}{\lambda^2} - 2\frac{20}{\lambda^3}\right] \\ &= -\frac{100}{\lambda^2} + 2\frac{\sum_i E(x_i)}{\lambda^3} = -\frac{100}{\lambda^2} + \frac{200}{\lambda^2} = \frac{100}{\lambda^2} \end{aligned}$$

Hence Jeffreys' prior is $\pi(\lambda) \propto I(\lambda)^{1/2} \propto (\lambda^{-2})^{1/2} = \lambda^{-1}$. The posterior becomes

$$\pi(\theta|x) \propto \lambda^{-100} \exp\left(-\frac{20}{\lambda}\right) \lambda^{-1} = \lambda^{-100-1} \exp\left(-\frac{20}{\lambda}\right)$$

which can be recognised as the IGamma$(100, 20)$

(c) Find a Bayes estimator for λ based on the priors of parts (a) and (b). [3 marks]

Answer: A standard Bayes estimator is the posterior mean which is equal to (see appendix)

$$\frac{20.1}{100.1 - 1} = 0.203$$

or

$$\frac{20}{100 - 1} = 0.202$$

depending on the chosen prior.

(d) Let y represent a future observation from the same model. Find the predictive distribution of y based either on the prior of part (a) or (b). [6 marks]

Answer:

$$\begin{aligned} f(y|x) &= \int_\Lambda f(y|\lambda)\pi(\lambda|x)d\lambda \\ &= \int_0^\infty \frac{1}{\lambda}\exp\left(-\frac{y}{\lambda}\right)\frac{(20)^{100}}{\Gamma(100)}\lambda^{-(100)-1}\exp\left(-\frac{20}{\lambda}\right)d\lambda \\ &= \frac{(20)^{100}}{\Gamma(100)}\int_0^\infty \lambda^{-(101)-1}\exp\left(-\frac{20+y}{\lambda}\right)d\lambda \\ &= \frac{(20)^{100}}{\Gamma(20)}\frac{\Gamma(101)}{(20+y)^{101}} \end{aligned}$$

for $y > 0$.

(e) Describe how you can calculate the mean the of the predictive distribution in software such as R. [5 marks].

Answer: Note that we can write the mean of the predictive distribution as

$$E(y|x) = \int_0^\infty y f(y|\lambda)\pi(\lambda|x)d\lambda.$$

Hence a Monte Carlo scheme would draw samples $y^{\{k\}}$, $k = 1, \ldots, N$ from $f(y|\lambda)\pi(\lambda|x)$ for some large N and then just take

$$\widehat{E(y|x)} = \frac{\sum_{k=1}^N y^{\{k\}}}{N},$$

To that in R once can

i. Draw -say- 10,000 samples from the Gamma(100,20) using `nu=rgamma(100,20)` .

ii. Invert those samples to make them samples from the IGamma(100,20) using `lambda=1/nu.`

iii. Using each of the samples in lambda, sample y by the model $y \sim$Exponential(lambda) using `y=rexp(lambda).`

iv. Calculate the sample mean of the values in y using `mean(y).`

3. (a) i. Suppose a non-linear model that can be written as

$$Y = f(X) + \epsilon,$$

where ϵ has zero mean and variance σ^2, and is independent of X. Show that the expected test error, conditional on X can be decomposed into the following three parts:

$$E\left[\left(Y - \hat{f}(X)\right)^2\right] = \sigma^2 + \text{Bias}\left[f(x)\right]^2 + \text{Var}\left[f(x)\right],$$

where $f(\cdot)$ is estimated from the training data. [7 marks]

Answer: Since $Y = f(x) + \epsilon$, we can write

$$\begin{aligned} E\left[\left(Y - \hat{f}(X)\right)^2 | X = x\right] &= E\left[\left(f(x) - \hat{f}(X) + \epsilon\right)^2\right] \\ &= E\left[\left(\hat{f}(X) - E\left(\hat{f}(X)\right) + E\left(\hat{f}(X)\right) - f(x) - \epsilon\right)^2\right] \\ = E\left[\left(\hat{f}(X) - E\left(\hat{f}(X)\right)\right)^2\right] &+ E\left[\left(E\left(\hat{f}(X)\right) - f(x)\right)^2\right] \\ +E\left[\left(\hat{f}(X) - E\left(\hat{f}(X)\right)\right)\left(E\left(\hat{f}(X)\right) - f(x)\right)\right] &+ E\left(\epsilon^2\right) \\ &= \text{Var}\left[f(x)\right] + \text{Bias}\left[f(x)\right]^2 + 0 + \sigma^2 \end{aligned}$$

The third equality comes from the fact that ϵ is independent of $\hat{f}(X)$. The fourth equality uses the definitions of variance and bias, $\text{var}(\epsilon) = E(\epsilon^2) = \sigma^2$ and the fact that the cross-product is equal to 0.

ii. To estimate the test error rate, one can use the 10-fold Cross Validation (CV) approach or the information criterion approach, e.g. AIC, BIC. What are the main advantage and disadvantage of using the 5-fold CV approach in comparison with AIC or BIC? [3 marks]

Answer: For the 10-fold CV, it is computational extensive because one need to fit the model 10 times, but only 1 time is needed for AIC or BIC. CV approaches provide direct estimates of the test error and make fewer assumption about the true model. For AIC or BIC, it is also hard to specify the model degrees of freedom.

iii. State which one of AIC and BIC tends to select smaller size model and explain the reason. [3 marks]

Answer: BIC places a heavier penalty on models with many variables and hence results in the selection of smaller models than AIC.

(b) i. The tree in Figure 1 provides a regression tree based on a dataset of patient visits for upper respiratory infection. The aim is to identify factors associated with a physicians rate of prescribing, which is a continuous variable. The variables appearing in the regression tree are *private:* percent of privately insured patients a physician has, *black:* the percent of black patients a physician has, and *fam* whether or not the physician specialises in family medicine. Provide an interpretation of this tree. [5 marks]

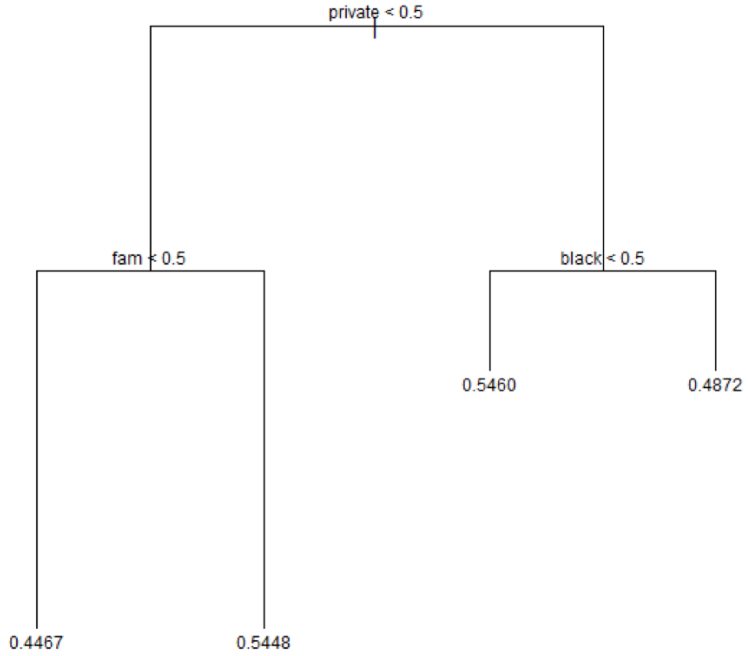

Figure 1: Regression tree for Question 3 (b) i.

Answer: Among those privately insured, black patient populations had a 48.72% average physician rate of prescribing, while physician's prescription rate for non-black populations was 54.60%. Among those without private insurance, the presence of a family medicine doctor raises the average provider prescribing rate by approximately 10%, to reach 54.48% (vs 44.67%), indicating that family medicine doctors systematically prescribe most antibiotics than non-family medicine doctors.

ii. Consider the regression tree of Figure 2 where the response variable is the log salary of a baseball player, based on the number of years that he has played in the major leagues (*Years*) and the number of hits that he made in the previous year (*Hits*). Create a diagram that represent the partition of the predictors spaces according to this tree.
Answer: The requested diagram showing the partition of the predictors spaces according to this tree is provided by Figure 3:

Figure 2: Regression tree for Question 3 (b) ii.

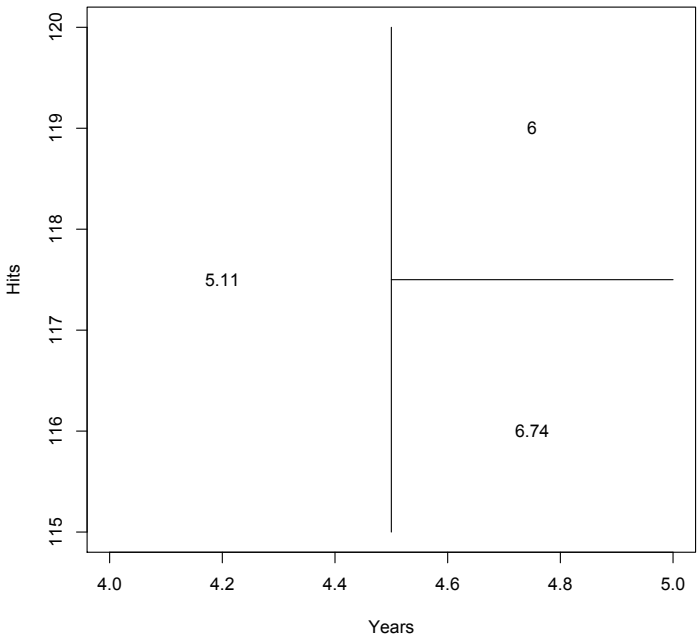

Figure 3: Partition of the predictor's according to the tree in Figure 2.

4 (a) i. Consider the following data:

10 20 40 80 85 121 160 168 195

Use the k-means algorithm with $k = 3$ to cluster the data set. Use the Euclidean distance to measure the distance between the data points. Suppose that the points 160, 168, and 195 were selected as the initial cluster means. Work from these initial values to determine the final clustering for the data. Provide results from each iteration. [9 marks]

Answer: The k-means clustering can be performed via the following steps:

- If we work through this you will see that initially observations (10, 20, 40, 80, 85, 121, 160) are closest to cluster centre 1, the observation 168 is closest to cluster centre 2, whereas the observation 195 is closer to cluster 3.
- The new cluster centres will then be the averages of the observations belonging to each cluster that are 73.71, 168 and 195 respectively.
- Now observations (10, 20, 40, 80, 85) are closest to cluster centre 1, the observations (121, 160, 168) are closest to cluster centre 2, whereas cluster centre 3 has the observation 195.
- The new cluster centre for cluster 1 will then be the average of the observations (10, 20, 40, 80, 85), which is 47, the new cluster centre for cluster 2 will be the average of observations (121, 160, 168), which is 149.67. Finally the centre of cluster 3 will remain unchanged to 195.
- As with the previous step, observations (10, 20, 40, 80, 85) are closest to cluster centre 1, the observations (121, 160, 168) are closest to cluster centre 2, whereas cluster centre 3 has the observation 195.
- Since there has been no change in the clusters k-means stops at this point with final cluster assignments of (10, 20, 40, 80, 85), (121, 160, 168), 195 and centres of 47, 149.67, 195.

ii. What are the main disadvantages of k-means clustering? Why one may want to consider hierarchical clustering as an alternative? [4 marks]

Answer: Regarding the k-means algorithm one drawback is that we can only find a local optimum rather than a global optimum, so the results obtained will depend on initial cluster assignment of each observation. Second, k-means clustering requires us to pre-specify the number of clusters k.

In hierarchical clustering there is no need to set k beforehand. Also, hierarchical clustering may also be appealing over K-means clustering in that it also offers a tree-based depiction of the data, called dendrogram.

(b) i. Data are available for students taking BSc degree in Data Science and in particular the variables X_1: average mark on project coursework, X_2: average hours studied per course, and Y: get a degree with distinction. The estimated coefficients of a logistic regression model were $\beta_0 =?5$, $\beta_1 = 0.02$, $\beta_2 = 0.1$. Estimate the probability that a student who takes on average 50% on project coursework and studies 30 hours on average for each course

gets a degree with distinction? How many hours would the student in part (a) need to study on average to have a 50 % chance of getting a degree with distinction ? [6 marks]

Answer: We have

$$p(X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)},$$

where X_1 = average coursework mark, X_2 = average hours studied per course, $\beta_0 = -5$, $\beta_1 = 0.02$ and $\beta_2 = 0.1$.

For $X_1 = 50$ and $X_2 = 30$ we get

$$p(X) = \frac{\exp(-5 + 0.02\ 50 + 0.1\ 30)}{1 + \exp(-5 + 0.02\ 50 + 0.1\ 30)} = 26.89\%.$$

For $X_1 = 50$ and $X_2 = x$ we get

$$p(X) = \frac{\exp(-5 + 0.02\ 50 + 0.1\ x)}{1 + \exp(-5 + 0.02\ 50 + 0.1\ x)} \text{ or else}$$
$$0.50 = \frac{\exp(-4 + 0.1\ x)}{1 + \exp(-4 + 0.1\ x)} \text{ or else } x = 40\text{hours}.$$

ii. Suppose that we wish to predict whether a high quality chip produced in a factory will pass the quality control ('Pass' or 'Fail') based on x, the measurement of its diameter. Diameter measurements are available for a large number of chips. After examining them it turns out that the mean value of x for chips that passed the quality control was 5mm, while the mean for those that didnt was 7mm. Moreover, the variance of x for these two sets of companies was $\sigma^2 = 1$. Finally, 70% of the produced chips passed the quality control. Assuming that x follows the normal distribution, predict the probability that a chip with $x = 5.8$ will pass the quality control. [6 marks]

Answer: For the probability of passing the quality control we get

$$\begin{aligned} p_{pass}(x) &= \frac{\pi_{pass} \exp(-\frac{1}{2\sigma^2}(x - \mu_{pass})^2)}{\pi_{pass} \exp(-\frac{1}{2\sigma^2}(x - \mu_{pass})^2) + \pi_{fail} \exp(-\frac{1}{2\sigma^2}(x - \mu_{fail})^2)} \\ &= \frac{0.70 \exp(-\frac{1}{2*1}(x - 5)^2)}{0.70 \exp(-\frac{1}{2*1}(x - 5)^2) + 0.30 \exp(-\frac{1}{2*1}(x - 7)^2)} \end{aligned}$$

Setting $x = 5.8$ we get

$$p_{pass}(5.8) = \frac{0.70 \exp(-\frac{1}{2*1}(5.8 - 5)^2)}{0.70 \exp(-\frac{1}{2*1}(5.8 - 5)^2) + 0.30 \exp(-\frac{1}{2*1}(5.8 - 7)^2)} = 77.68\%$$

Appendix: Table of Common Distributions

Binomial(n, θ)**:** number of successes in n independent Bernoulli trials with probability of success θ.

- $f(x|\theta) = P(x|\theta) = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}$ for $x = 0,\ 1, \ldots,\ n$.
- $E(X) = n\theta$, $\text{Var}(X) = n\theta(1-\theta)$.

NegBin(r, θ)**:** number of successes before r^{th} failures in repeated independent Bernoulli trials.

- $f(x|\theta) = P(x|\theta) = \binom{x+r-1}{x}\theta^x(1-\theta)^r$ for $x = 0,\ 1, \ldots$.
- $E(X) = \frac{r(1-\theta)}{\theta}$, $\text{Var}(X) = \frac{r(1-\theta)}{\theta^2}$.

Poisson(λ)**:** often used for the number of events which occur in an interval of time.

- $f(x|\lambda) = P(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$ for $x = 0,\ 1, \ldots$.
- $E(X) = \lambda$, $\text{Var}(X) = \lambda$.

Normal N(μ, σ^2)**:** characterized by first two moments.

- $f(x) = (2\pi\sigma^2)^{-1/2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for $-\infty < x < \infty$.
- $E(X) = \mu$, $\text{Var}(X) = \sigma^2$.

Beta(α, β)**:** characterized by parameters $\alpha > 0$ and $\beta > 0$.

- $f(x) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$ for $0 \le x \le 1$, $B(\alpha,\beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
- $E(X) = \frac{\alpha}{\alpha+\beta}$, $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$.

Gamma(α, β)**:** characterized by parameters $\alpha > 0$ and $\beta > 0$.

- $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x)$ for $0 \le x < \infty$, $\Gamma(t) = \int_0^\infty y^{t-1}e^{-y}dy$.
- $E(X) = \frac{\alpha}{\beta}$, $\text{Var}(X) = \frac{\alpha}{\beta^2}$.

IGamma(α, β)**:** characterized by parameters $\alpha > 0$ and β. If $X \sim \text{Gamma}(\alpha, \beta)$, then $1/X \sim \text{IGamma}(\alpha, \beta)$.

- $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}\exp\left(-\frac{\beta}{x}\right)$ for $0 \le x < \infty$.
- $E(X) = \frac{\beta}{\alpha-1}$, $\text{Var}(X) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$. for positive integer n.

END OF PAPER