Modelling and analysis of gene expression data

1. Introduction

The paper performs analysis and to fit a non-linear regression model to the gene expression dataset. The dataset comprises of 5 gene (x1, x2, x3, x4, x5) and Time (in mins). The input genes are x4 and x5 and the output gene is x5. The goal of the analysis is to find the best model with the least Mean Squared Error (MSE) from the input genes. We build the candidate input (x4, x5) set X using the given model structure and refer the output gene x3 as 'y'.

$$\mathbf{x}_3 = w_0 + a_1\mathbf{x}_4 + a_2\mathbf{x}_4^2 + a_3\mathbf{x}_4^3 + \cdots + b_1\mathbf{x}_5 + b_2\mathbf{x}_5^2 + b_3\mathbf{x}_5^3 + \cdots + \epsilon$$

To gain better understanding of the data provided in the problem, we perform exploratory data analysis and then proceed further with the model selection using the forward subset model selection algorithm. Once the best model is identified we validate the model. The uncertainty between the model parameters and predictions are identified. The posterior is estimated using the Approximate Bayesian computation.

1.1 Data

The data provided consist of 5 stimulated gene expression time series data. Protein is encoded within the gene which dictates the cell function. Thousands of such genes together determines the functionality of the cell. The flow of information from DNA to RNA to protein provide the cell with self-regulating functionalities by regulating the amount of and type of protein produced within them, this is also known as transcription factor. Gene regulation proves vital in determining a complex health issue. The main motive of this assignment is to fit a nonlinear model to the gene expression data set. The 'simulated' 5 gene expression time-series consist of 301 data points. The gene 'x3' is regulated by gene 'x4' and 'x5', although it is unknown if the regulation function is activation or suppression. The assignment fits a nonlinear polynomial regression model with 2 input genes 'x4' and 'x5' and output gene 'x3. The main objective of this task is to identify the (polynomial) model structure, estimate model parameters from the training data using forward model selection, and use the identified model to predict the response/output signal.

2. Exploratory data analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to understand the relationship between the input genes and output gene.
Initially, we begin by evaluating the time series graph of the two input genes with respect to the output gene. We plot the time expression of the gene on the y-axis and the time on the xaxis. We can observe that the scale of the input genes and output gene is very similar to each other. They fluctuate in similar patterns. If we observe the time series graph for output gene y, we can see that the graph, it seems that this time series could probably be described using an additive model, as the fluctuations in pattern scale out to be constant throughout and constant in size over time. Both the input gene follow a similar pattern.

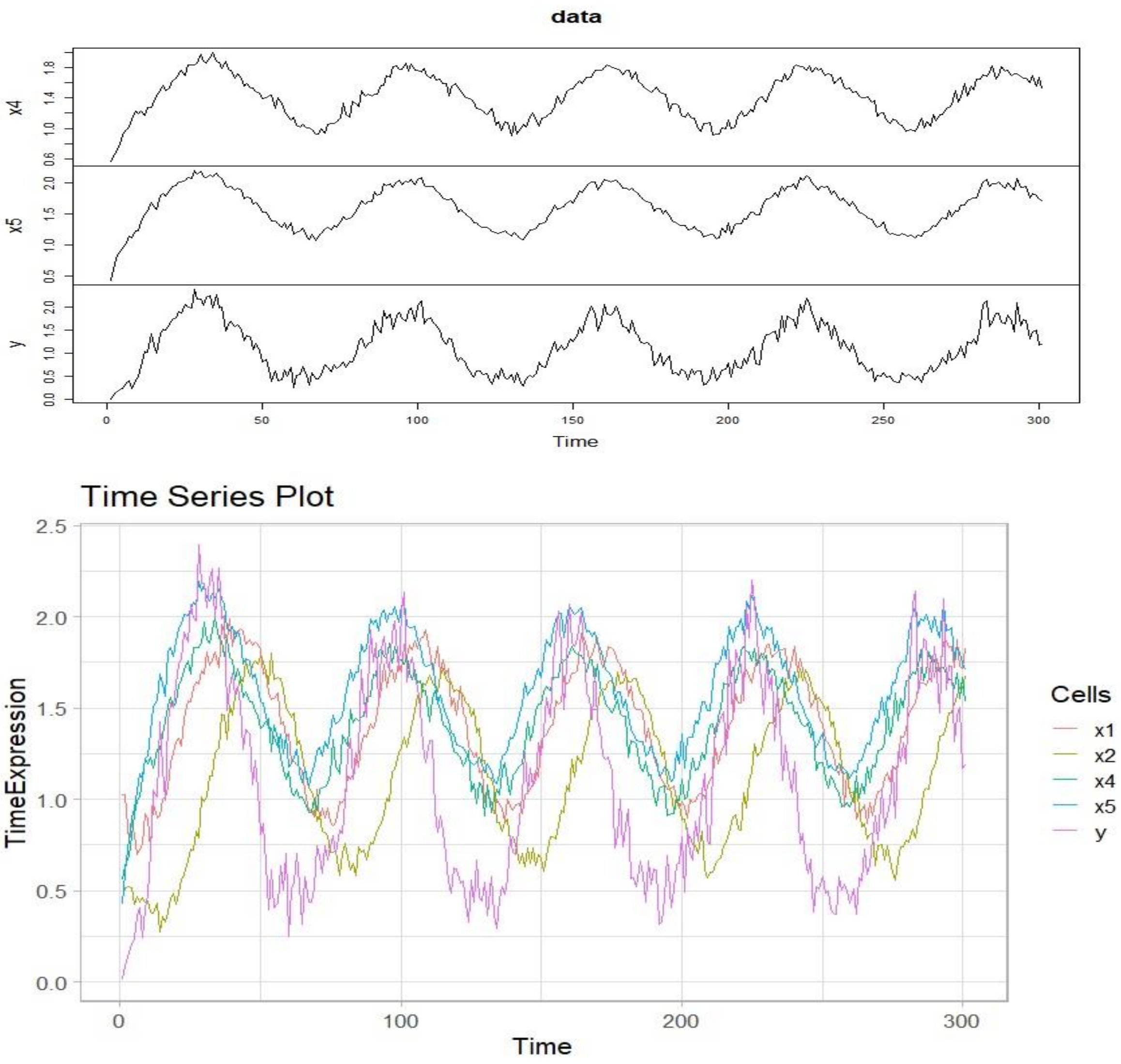
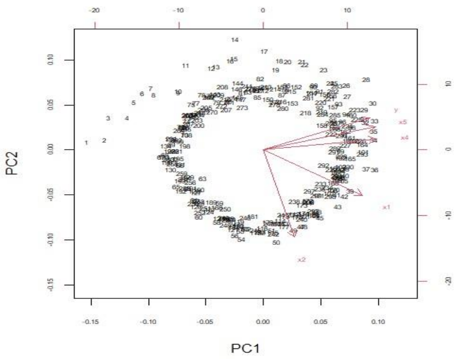


Figure 1 Top: Time series analysis for input and output gene
Bottom: Time series analysis for all genes in the data with respect to output gene y

Principal component analysis, was applied to the dataset to reduce the dimension of time. The results are attached below.



The relationship between the input gene x4 and x5 and output x3 is non-linear as we can see an increasing quadratic polynomial curve between them. (Figure 2)

The correlation significance between the input and output is calculated using Pearson's coefficient test. The results obtained proved that there is a positive correlation between the input gene and output gene. The observations are summarised in table 1 below.

Gene	Correlation	p-value	t
x4	0.8937071	< 2.2e-16	34.445
x5	0.9694559	< 2.2e-16	68.348

Table 1 Results for Pearson's coefficient test

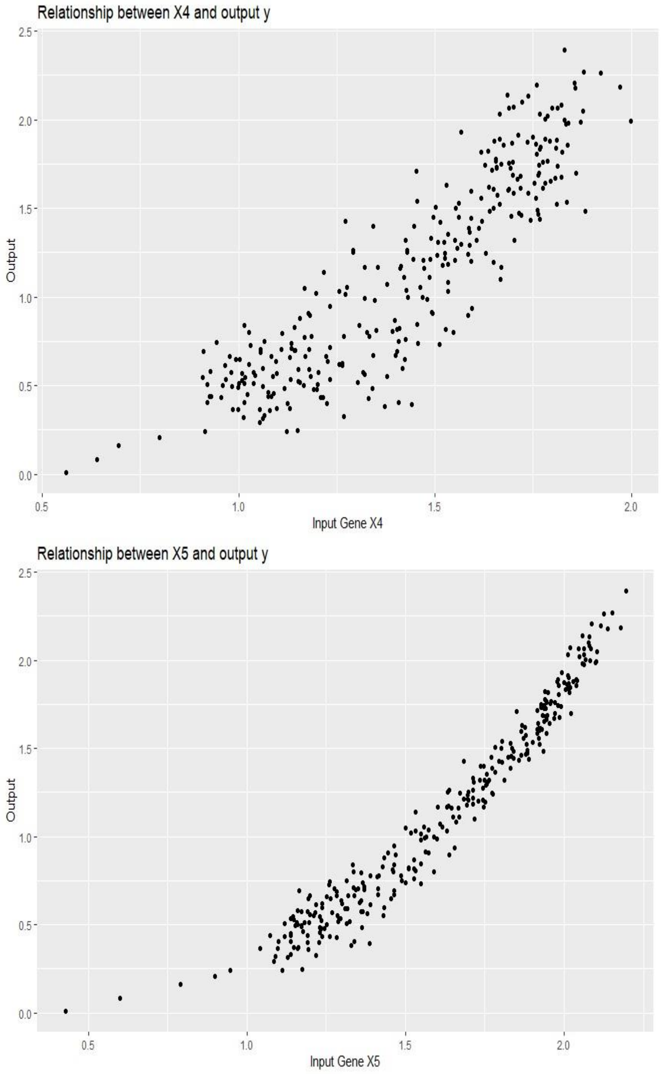


Figure 2 Top: Relationship between gene x4 and output gene x3
Bottom: Relationship between gene x5 and output gene x3

Q-Q plots are used to observe if the data is normally distributed, it plots the quantiles of the data to the quantiles of theoretical normal distribution. In this study, the Q-Q plots created do observe to deviate from the straight diagonal line, hence we conclude that the distribution is not normal. From figure 3, we can see that there is a deviation from the diagonal. We can also observe few outliers in both input and output.

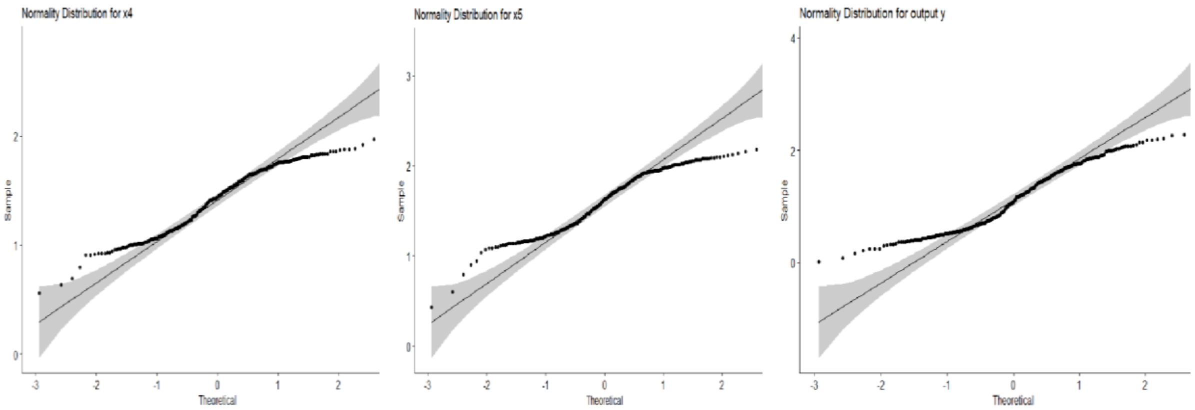


Figure 3 Left: Normal Distribution for input gene x4 | Centre: Normal Distribution for input gene x5 | Right: Normal Distribution for output gene x3

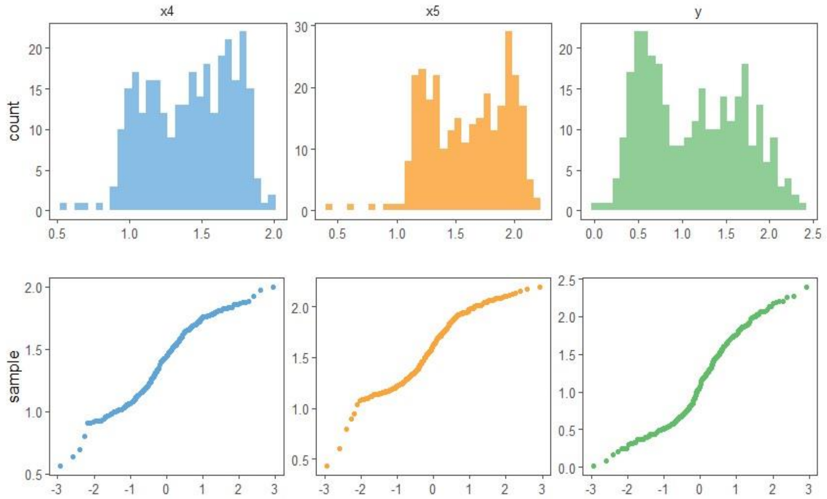


Figure 4 Top: Histogram for input genes and output y
Bottom: Q-Q plots

2.1 Linear Model Fitting

We have observed that the relationship between the input and output is not linear from the scatter plots above (Figure 2). In order to confirm the observation, we fit a linear model of the form $y = \text{ß}_0 + \text{ß}_1\text{x}$. The model is evaluated using mean squared value (MSE). The formula to calculate MSE is given below

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \qquad \textbf{(1)}$$

The MSE obtained is 0.09676259. The residual analysis shows that the graph distribution is not normal, as most of the residuals exist below the zero line. (Figure 5)

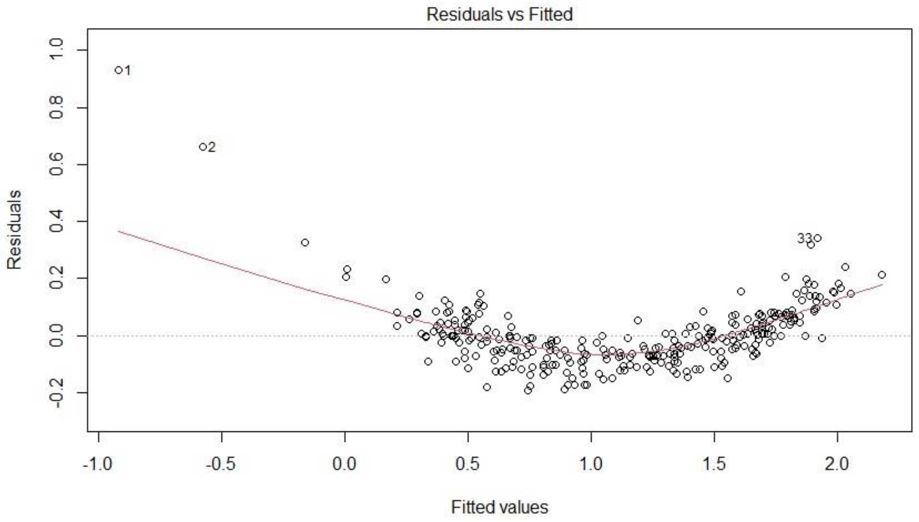


Figure 5 Residual Plot for Linear Model Fitting

To summarize we can say that linear model is a bad fit with high MSE value and poorly distributed residual plot.

3. Model selection (500)

There are various ways in which a model can be built from a candidate set, but it does not ensure how good the model can be. To build a good model, it is crucial to select the models accurately using various Model selection techniques such as Forward selection, Akaike information criterion (AIC), Bayesian information criterion (BIC), etc. In this paper we use Forward selection to select the best model parameters from the candidate set X. We build a generic non-linear model with two input genes (x4 and x5), and output x3. The exemplar model structure is mentioned below

$$x_3 = w_o + \alpha_1 x_4 + \alpha_2 x_{42} + \alpha_3 x_{43} + \alpha_4 x_{44} + \beta_1 x_5 + \beta_2 x_{52} + \beta_3 x_{53} + \beta_4 x_{54} + \varepsilon \qquad (3)$$

Here w_o denotes a bias term, ($\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3, \beta_4$) are the parameters of the model, ε denotes the denotes an additive, Gaussian, zero-mean noise. We also build the candidate set in the form of a matrix, X = (intercept, x_4, x_4^2, x_4^3, x_4^4, x_5, x_5^2, x_5^3, x_5^4). The intercept is a matrix which comprises of all ones. The X matrix is split into 80% training dataset and 20% testing dataset for the model selection procedure. Also, similarly we divide output y into training and testing. We perform Least squares method (LSE) on the training set of the input and output and obtain parameter estimates. Next step, is to calculate the predictions using only the test part of the dataset. The main motive of calculating the predictions on the testing set is to obtain “true” model prediction as only the training data was used to estimate the parameters and the testing is performed on the “unseen” testing dataset. In forward selection, we run iterations on the data and select the term which gives the minimum MSE and keep on adding it to the list of final models. MSE works as the performance metric in choosing the best parameter models out of the list. The model selected for the final model list is removed from the original list of possible model parameter list for the next iteration. This process is done iteratively, as per the condition. In this paper we have run the iteration thrice.

The results of the forward subset selection are recorded in the table below,

Model Terms	MSE
intercept	0.003982248
$\alpha_4 x_4^4$	0.004600914
$\beta_3 x_5^3$	0.01059285

Table2 Results for the Forward subset selection

Thus, the best model structure according to the results obtained is

$$y = w_0 + \alpha_4 x_4^4 + \beta_3 x_5^3 + \varepsilon \qquad (\text{MSE} = 0.004273755) \qquad \textbf{(4)}$$

Once the final model is evaluated, the covariance matric is obtained using the final model, recorded in the following table. Covariance defines the variability between the coefficients.

	One_Theta	Two _Theta	Three_Theta
One_Theta	0.033383151	0.004961579	-0.023556364
Two _Theta	0.004961579	0.003514412	-0.006698607
Three_Theta	-0.023556364	-0.006698607	0.020447659

Table 3 Results for the Covariance Matrix for the final model parameters

One_theta, Two_Theta, Three_Theta denotes intercept, x_4^4, x_5^3 respectively.

4. Model evaluation

To evaluate the model, we first look at the residual histogram and QQ plots.
The **residuals** define the deviation of the observed value from the predicted value in the model. As seen below in figure 6, we can observe that the mean and median for the distribution is very close to each other, the curve follows a roughly approximate normal curve. The QQ- plot falls nearly on the diagonal line with a few outliers.

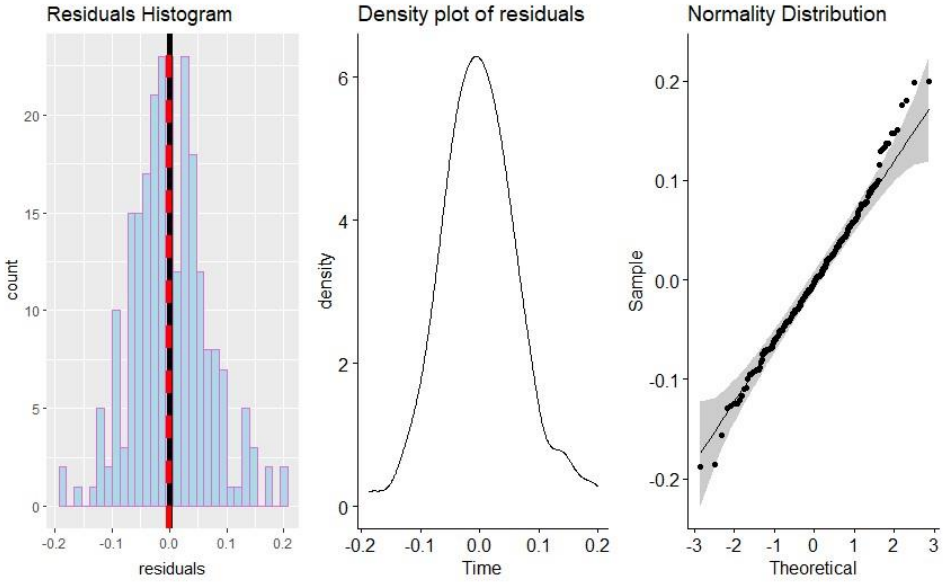


Figure 6 Residual Plot for Final Model after Forward subset selection.

The covariance matrix (table 3), helps to evaluate the relationship between the model parameters. The maximum covariance exists between one_theta and two_theta (~ 0.0049)

The parameter estimate uncertainty probability density function was calculated for each of the combination of two parameters and plotted in 3D.

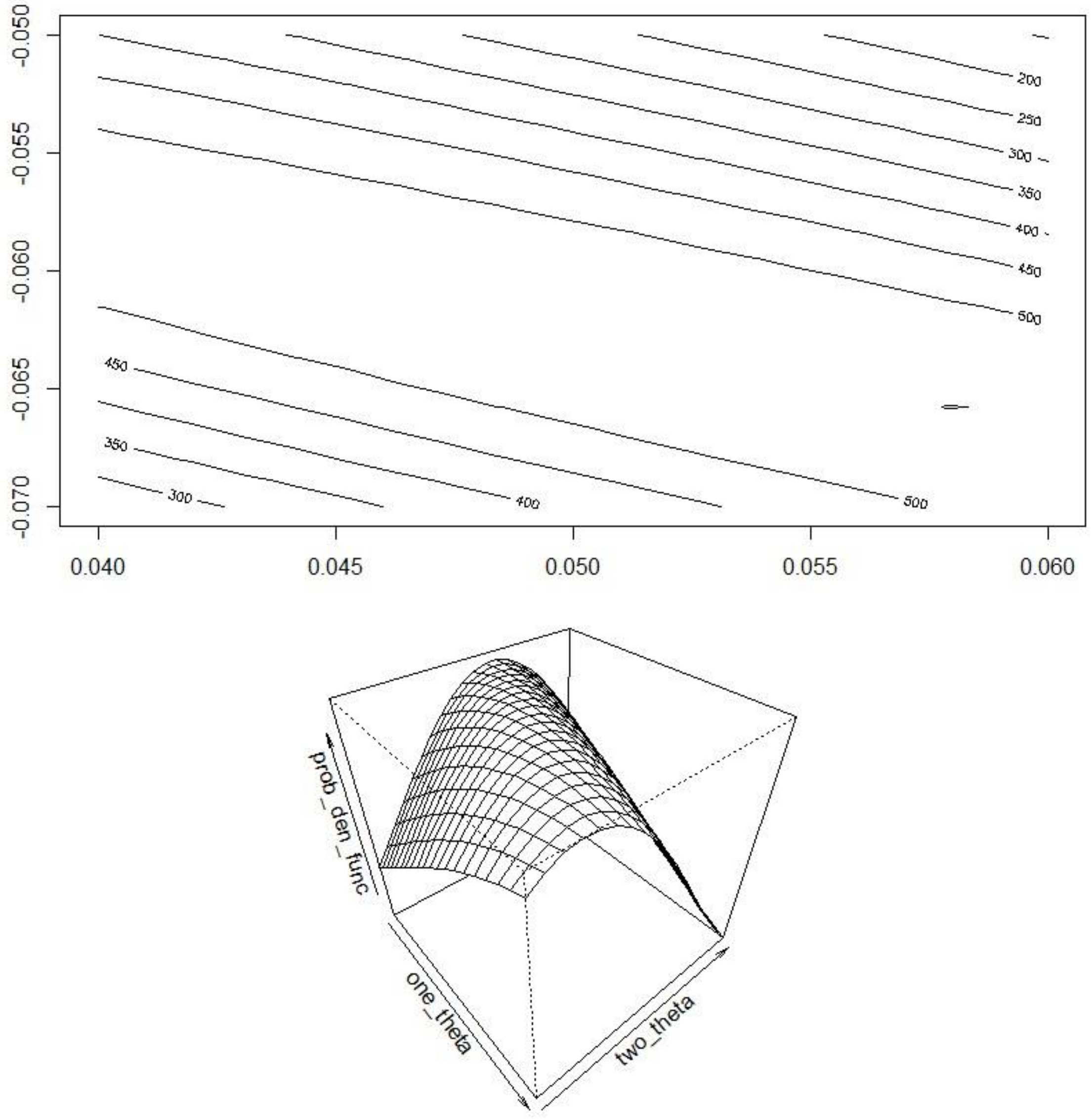


Figure 7 Top: Contour plots for one_theta and two_theta
Bottom: 3D perspective plot for one_theta and two_theta

The perspective plot for uncertainty density is smooth with a downward slope towards two_theta as it increases. This shows that they might have similar probability values

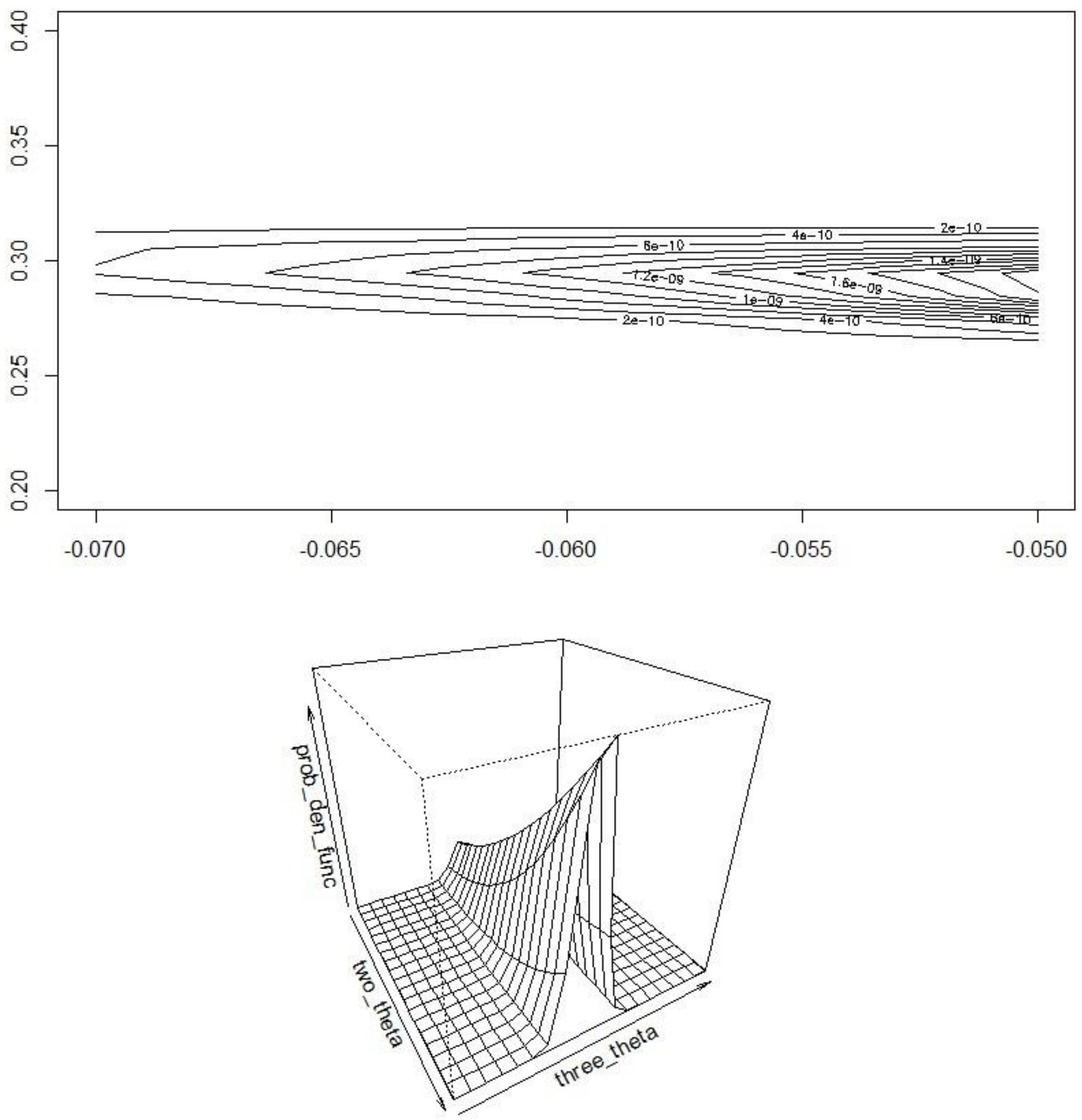


Figure 8 Top: Contour plots for two_theta and three_theta
Bottom: 3D perspective plot for two_theta and three_theta

The plots in figure 8 and 9, shows a sharp spike in probability density function, with small peak as well, this makes it bimodal in nature. The contours are rather narrow, showing less uncertainty of the parameters.

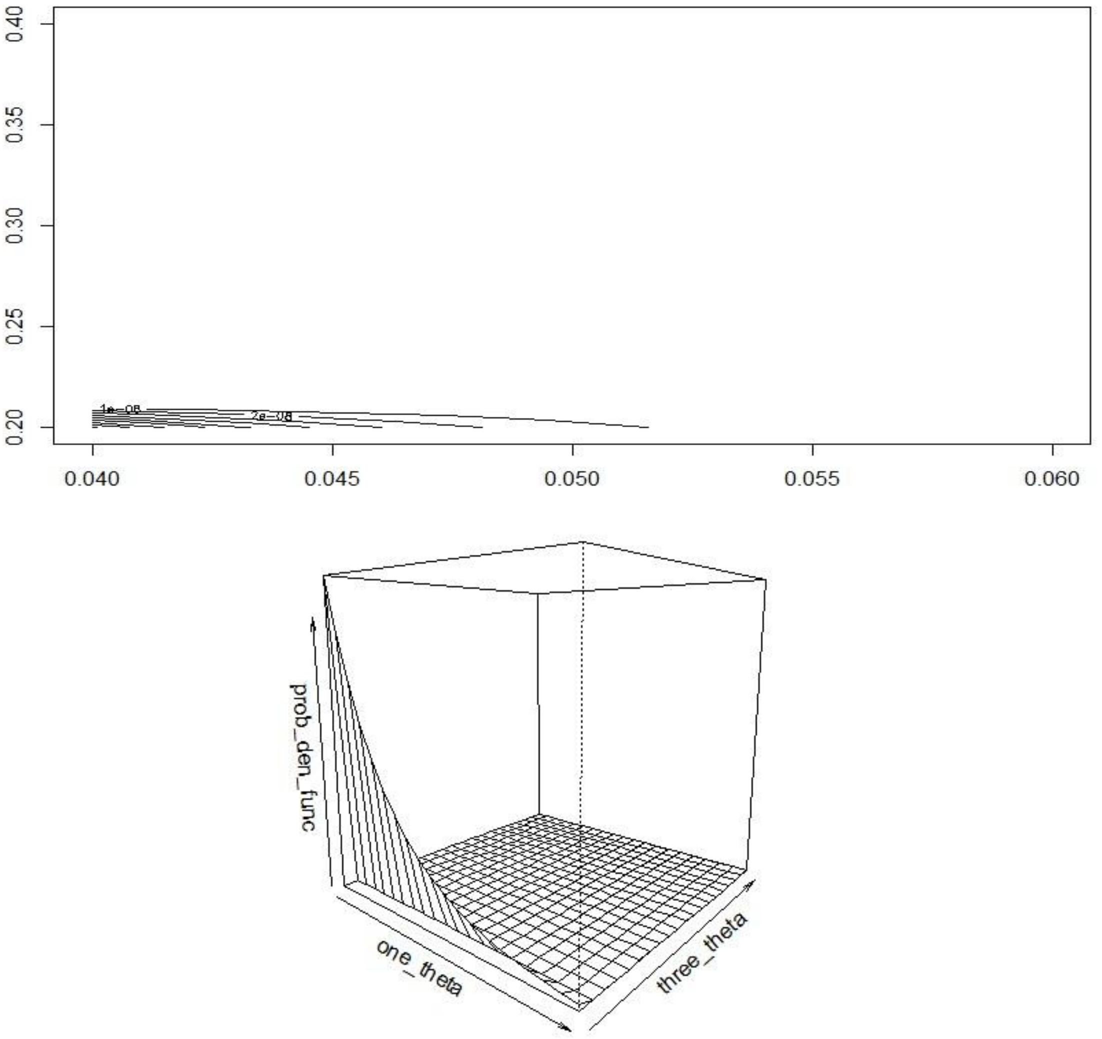


Figure 9 Top: Contour plots for one_theta and three_theta
Bottom: 3D perspective plot for one_theta and three_theta

4.1 Confidence Intervals

The model is also evaluated using 95% Confidence Intervals (Figure 7), Confidence Intervals define how well our sample is representing the total population.

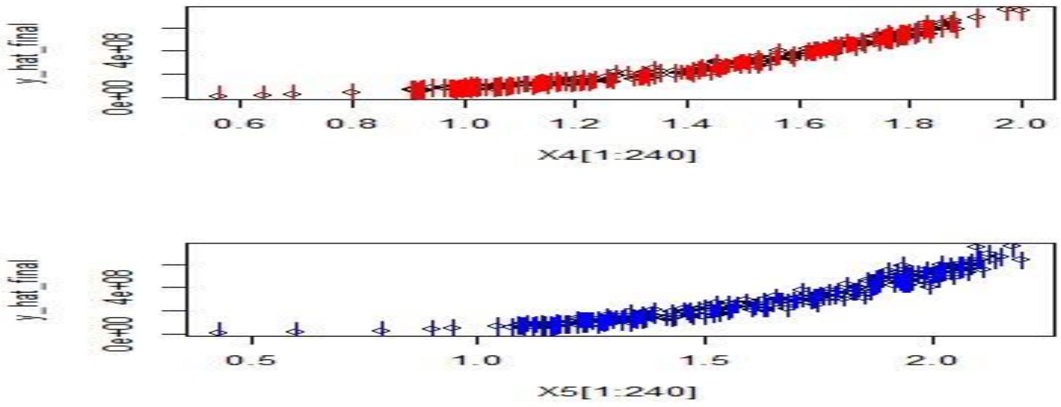


Figure 10 Confidence Intervals

As observed, we in the figure above, we can observe that the predicted values are very close to the true values of the dataset and also the intervals are narrow. This is due to the fact that the data on which the predictions generated and the data for fitting the model is the same.

4.2 Model Validation

Model validation was performed to check the performance of the model, we use the training dataset but with different ranges for the training and testing dataset, in our case we split the 70% of data into training set and 30% of it for the testing set. MSE is calculate and used as the performance metrics. The result for the training MSE obtained is 0.004399559 and testing MSE is 0.003810138. The difference between them is quite small, in an ideal model the MSE training and testing MSE would be almost identical.

5. Approximate Bayesian computation

Approximate Bayesian Computation is used to fit the Bayesian statistical methods without the calculating the likelihood function, it estimates the posterior distribution of the evaluated best selected model parameters. 30M samples are drawn from the prior generated using the runif() function which generates random deviates of the uniform distribution and plotted in the graph attached below. These priors are used to stimulate data for ABC, instead of the actual data. The priors are plotted below in Figure 11

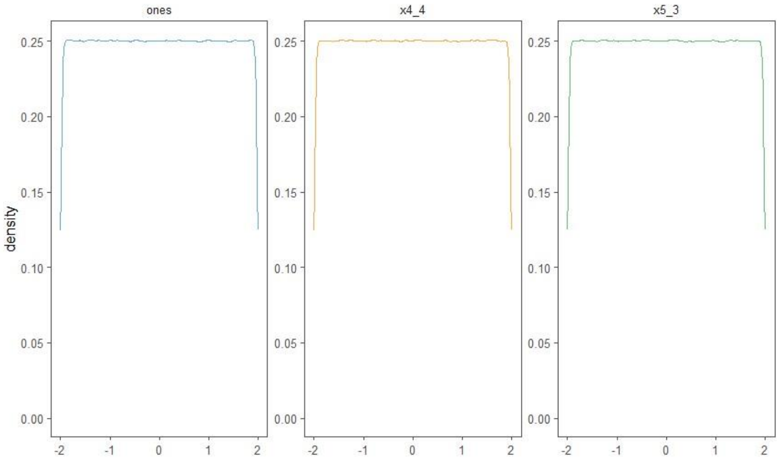


Figure 11 Priors distributions

The rejection threshold was set to 0.05, allowing only those values from the priors that gives the resultant MSE 0.05 or less. These selected values were accepted as the posterior samples for the experiment. 1379 samples were selected.

Marginal and joint posterior distribution combination are plotted below in Figure 12. We can see that all the combination follows almost a normal distribution throughout. There exists high probability given the width of the distribution. The yellow part in the joint probability distribution denotes the similarity of probability distribution between the parameters.

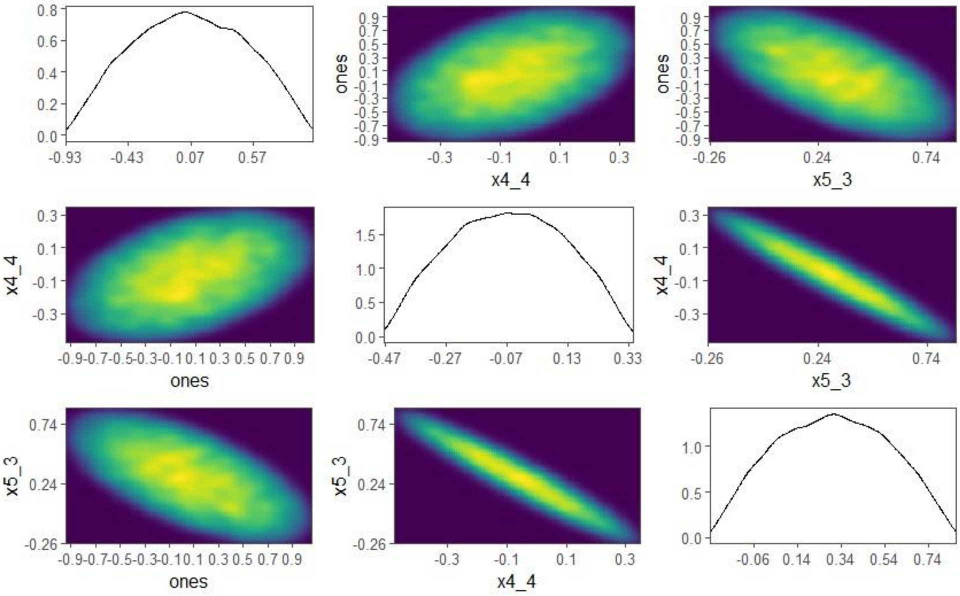


Figure 12 Marginal and Joint Posterior Distribution of Model parameters.

6. Conclusion

In conclusion, the model parameters were selected using forward subset selection, the dataset was divided into training and testing dataset for the experiment. Model validation was performed and MSE_ train (0.004399559) and MSE_test (0.003810138) was very close to each other, this implies that the model is one of the best fit solution. Approximate Bayesian Computation was utilised to calculate the posterior predictions. The marginal probability and joint posterior distribution was further evaluated by plots.