Faculty of Engineering, Environment and Computing

7089CEM: Introduction to Statistical Methods for Data Science

Assignment Brief

Module Title Introduction to Statistical Methods for Data Science	Individual	Cohort (Sept/Jan): Jan and May start	Module Code 7089CEM
Coursework Title **Modelling and analysis of gene expression data**			Hand out date: 22/05/2020
Lecturer Dr Fei He			Due date and time: 19/06/2020, 18:00
Estimated Time (hrs): 4 weeks Word Limit*: 3000 - 4000	Coursework type: Individual assignment		% of Module Mark: 100%

- Submission arrangement: online via CUMoodle, and Turnitin.
- File types and method of recording: Report (Word), Programme code (R, or Matlab script)
- Mark and Feedback date: 2 weeks after submission
- Mark and Feedback method (e.g. in lecture, written via Gradebook): provided in Moodle

Module Learning Outcomes Assessed:

- Demonstrate knowledge of underlying concepts in probability and statistics used in Data Science.
- Select and apply appropriate statistical methods or techniques to solve problems or analyse data sets.
- Use modern software to solve real world problems and analyse large data sets.
- Interpret the results of their analyses and communicate those results accurately.

Task and Mark distribution:

Coursework Description:

The aim of this assignment is to fit a non-linear time series model to the gene expression data set. Gene expression is one of the most important biological processes where information from a gene is used to synthesize a functional gene product, such as protein. The expression of a gene can be controlled (or regulated) by another gene or several other genes, through a gene product (protein) called transcription factor. Understanding how genes regulate each other, i.e. gene regulation, is important to investigate a complex diseases, and how cell respond to environmental stimuli.

Data:
The 'simulated' 5 gene expression time-series data, are given in the excel file (gene_data.csv). The first column contains the sampling time in minutes, the rest 5 columns are the time-course expression data

of 5 genes$\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$, respectively. All these 5 genes are subject to additive noise (assuming independent and identically distributed ("i.i.d") Gaussian with zero-mean) with unknown variance.

Task 1: Preliminary data analysis

You should first perform an initial exploratory data analysis, by investigating:

- Time series plots
- Distribution for each gene
- Correlation and scatter plots (between combination of two genes) to examine their dependencies

Task 2: Dimension reduction

- We would like to reduce the dimension of time (for all 5 genes) to two using PCA, you can choose to use either eigen-decomposition method or the singular value decomposition method.
- Plot these 5 genes in the reduced 2-dimensional space, with different notations or colours.

Task 3: Nonlinear regression – modelling gene regulation

We know one of the genes $\mathbf{x}_3$ is regulated by the other two genes $\mathbf{x}_4$ and $\mathbf{x}_5$, however, we do not know if such regulation is activation or repression, or if such a regulatory interaction is linear or nonlinear. Therefore, we will fit a generic nonlinear polynomial regression model (with 2 inputs) to the data with the following exemplar structure:

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

Here w_0 is a bias term (denotes the basal transcription rate); $\{a_1, a_2, a_3, \cdots, b_1, b_2, b_3, \cdots\}$ are the parameters of the regression model to be estimated, and ϵ denotes an additive, Gaussian, zero-mean noise.

The main objective of this task is to identify the (polynomial) model structure, estimate model parameters from the training data, and use the identified model to predict the response/output signal.

Then you need to identify the nonlinear regression model structure and estimate its parameters, by

- Identify the correct model structure (by using a model selection approach – e.g. subset selection, AIC/BIC, or explore all possible different model structures), so that the model provides you a good mean square error (MSE) and the model residual/error is close to Gaussian. You can either:

    i) Split the input and output dataset into two part: one part used to train the model, the other used for testing (e.g. 80% for training, 20% for testing). Apply the forward subset selection approach to select the best model structure iteratively (select the most significant term that reduce the MSE on testing data, in each iteration, and add it to the current model).
    ii) Or select the best model, using BIC or AIC goodness-of-fit criteria, by exploring all possible combinations (or out of the different possible model structures).

    The underlying nonlinear polynomial model may contain a bias term, a linear term, and one or few (input) nonlinear terms; the nonlinear terms can have a (maximum) nonlinearity up to 4th

order, the maximum model terms will be no more than 3 (including bias, linear and nonlinear terms).

- Estimate the model parameters using least squares method. This step will be embedded within the above model structure identification process (since for each candidate model structure, you will need to estimate its parameters, in order to evaluate the model's performance against observation data).
- Once the best model structure is selected and its parameters are estimated, estimate the parameter covariance matrix, plot corresponding parameter uncertainty p.d.f. in the 3D and/or contours (similar to the example given in the lecture/lab notes). Plot the pair-wise combinations of all parameters, if you have more than 2 parameters in the selected model.
- Compute the model's output/prediction (on the training data), and also compute the 95% confidence intervals and plot them (with error bars) together with the mean values of the model prediction.
- Validate the model using train-test split validation approach (may use different splitting portion as the subset model selection stage), to check whether the identified model provide good prediction on the testing dataset.
- Using "Approximate Bayesian Computation (ABC)" method to compute the posterior distribution of the regression model parameters (using rejection ABC and assuming a Uniform prior). Plot the marginal posterior distribution for each parameter, and the joint posterior probability distribution for all pair-wise combinations of parameters.

Marking Scheme

This coursework worth 15 credits (100%). This will be marked according to:

- 15% will be given for performing an initial data analysis (histogram plots, simple input-output correlation measures, time series plots, fitting linear model …). If you create any R code, you must include this in the report.
- 10% will be given for performing dimension reduction using PCA and plotting the result.
- 25% will be given for writing the R code that to select the correct model structure, estimate the model's parameters, use these estimates to calculate new predictions.
- 20% will be given for estimating the parameter estimation uncertainties (covariance matrix, plot the corresponding parameter estimates distribution) and the model's prediction confidence intervals (on the training input data). Again, if you create any R code, you must include this in the report.
- 5% will be given for performing model validation and analysing the performance of the identified nonlinear model.
- 5% will be given to perform the Approximate Bayesian computation to compute the (approximated) posterior distribution of the regression model.
- 10% will be given to appropriate discussion and interpretation of the results you obtained.
- 10% for writing the report (around 3000-4000 words) in a structured, readable form and submitting the executable R scripts. Report should be in sections with appropriate headings, an introduction and a conclusion.

Notes:

1. You are expected to use the Coventry University Harvard Referencing Style. For support and advice on this students can contact Centre for Academic Writing (CAW).
2. Please notify your registry course support team and module leader for disability support.
3. Any student requiring an extension or deferral should follow the university process as outlined here.
4. The University cannot take responsibility for any coursework lost or corrupted on disks, laptops or personal computer. Students should therefore regularly back-up any work and are advised to save it on the University system.
5. If there are technical or performance issues that prevent students submitting coursework through the online coursework submission system on the day of a coursework deadline, an appropriate extension to the coursework submission deadline will be agreed. This extension will normally be 24 hours or the next working day if the deadline falls on a Friday or over the weekend period. This will be communicated via your Module Leader.
6. *(ML's delete if not applying to this assessment) Assignments that are more than 10% over the word limit will result in a deduction of 10% of the mark i.e. a mark of 60% will lead to a reduction of 6% to 54%. The word limit includes quotations, but excludes the bibliography, reference list and tables.
7. You are encouraged to check the originality of your work by using the draft Turnitin links on your Moodle Web.
8. Collusion between students (where sections of your work are similar to the work submitted by other students in this or previous module cohorts) is taken extremely seriously and will be reported to the academic conduct panel. This applies to both courseworks and exam answers.
9. A marked difference between your writing style, knowledge and skill level demonstrated in class discussion, any test conditions and that demonstrated in a coursework assignment may result in you having to undertake a Viva Voce in order to prove the coursework assignment is entirely your own work.
10. If you make use of the services of a proof reader in your work you must keep your original version and make it available as a demonstration of your written efforts.
11. You must not submit work for assessment that you have already submitted (partially or in full), either for your current course or for another qualification of this university, unless this is specifically provided for in your assignment brief or specific course or module information. Where earlier work by you is citable, ie. it has already been published/submitted, you must reference it clearly. Identical pieces of work submitted concurrently will also be considered to be self-plagiarism.

Mark allocation guidelines to students (to be edited by staff per assessment)

0-39	40-49	50-59	60-69	70+	80+
Work mainly incomplete and /or weaknesses in most areas	Most elements completed; weaknesses outweigh strengths	Most elements are strong, minor weaknesses	Strengths in all elements	Most work exceeds the standard expected	All work substantially exceeds the standard expected

Marking Rubric (To be edited by staff per each assessment)

GRADE	ANSWER RELEVANCE		ARGUMENT & COHERENCE		EVIDENCE		SUMMARY	
First **≥70**	Innovative response, answers the question fully, addressing the learning objectives of the assessment task. Evidence of critical analysis, synthesis and evaluation.		A clear, consistent in-depth critical and evaluative argument, displaying the ability to develop original ideas from a range of sources. Engagement with theoretical and conceptual analysis.		Wide range of appropriately supporting evidence provided, going beyond the recommended texts. Correctly referenced.		An outstanding, well-structured and appropriately referenced answer, demonstrating a high degree of understanding and critical analytic skills.	
Upper Second **60-69**	A very good attempt to address the objectives of the assessment task with an emphasis on those elements requiring critical review.		A generally clear line of critical and evaluative argument is presented. Relationships between statements and sections are easy to follow, and there is a sound, coherent structure.		A very good range of relevant sources is used in a largely consistent way as supporting evidence. There is use of some sources beyond recommended texts. Correctly referenced in the main.		The answer demonstrates a very good understanding of theories, concepts and issues, with evidence of reading beyond the recommended minimum. Well organised and clearly written.	
Lower Second **50-59**	Competently addresses objectives, but may contain errors or omissions and critical discussion of issues may be superficial or limited in places.		Some critical discussion, but the argument is not always convincing, and the work is descriptive in places, with over-reliance on the work of others.		A range of relevant sources is used, but the critical evaluation aspect is not fully presented. There is limited use of sources beyond the standard recommended materials. Referencing is not always correctly presented.		The answer demonstrates a good understanding of some relevant theories, concepts and issues, but there are some errors and irrelevant material included. The structure lacks clarity.	
Third **40-49**	Addresses most objectives of the assessment task, with some notable omissions. The structure is unclear in parts, and there is limited analysis.		The work is descriptive with minimal critical discussion and limited theoretical engagement.		A limited range of relevant sources used without appropriate presentation as supporting or conflicting evidence coupled with very limited critical analysis. Referencing has some errors.		Some understanding is demonstrated but is incomplete, and there is evidence of limited research on the topic. Poor structure and presentation, with few and/or poorly presented references.	
Fail **<40**	Some deviation from the objectives of the assessment task. May not consistently address the assignment brief. At the lower end fails to answer the question set or address the learning outcomes. There is minimal evidence of analysis or evaluation.		Descriptive with no evidence of theoretical engagement, critical discussion or theoretical engagement. At the lower end displays a minimal level of understanding.		Very limited use and application of relevant sources as supporting evidence. At the lower end demonstrates a lack of real understanding. Poor presentation of references.		Whilst some relevant material is present, the level of understanding is poor with limited evidence of wider reading. Poor structure and poor presentation, including referencing. At the lower end there is evidence of a lack of comprehension, resulting in an assignment that is well below the required standard.	
Late submission	0		0		0		0	