Statistical Genetics and Bioinformatics
Coursework 2 (Weight: 50% of the module)

Marina Evangelou

General Instructions

Deadline: 05 May 2022 1pm (UK Time)

Write a report (up to 12 pages, including appendix; smallest font allowed 11pt).

The report should be in portable document format (PDF) and should be uploaded to blackboard (URL: bb.imperial.ac.uk) at the **Coursework 2** folder of the Statistical Genetics and Bioinformatics module page. Once the report is uploaded at blackboard there is no option for re-uploading so you should upload your final version only. Avoid last minute uploads, because the system can crash if it receives too many requests simultaneously.

In addition, the source files (.R, .Rmd, etc) of the conducted work should be uploaded in the **Coursework 2 Source Files** folder of the Statistical Genetics and Bioinformatics module page.

As this is assessed work you need to work on it individually. It must be your own and unaided work. You are not allowed to discuss the assessed coursework with your fellow students or anybody else. All rules regarding academic integrity and plagiarism apply. Violations of this will be treated as an examination offence. In particular, letting somebody else copy your work constitutes an examination offence.

All questions that you may have concerning the coursework must be addressed to the lecturer via e-mail. Any resulting clarifications will be communicated to the entire cohort via Blackboard announcements. The use of the Ed Discussion Forum for posting questions related to the coursework is not allowed.

Question 1: Gene expression analysis [20 marks]

The data for this question are available at:

`https://www.ma.ic.ac.uk/~me208/StatisticalGeneticsCoursework2_2022/data_CID.txt`

where you replace CID with your CID number (with the leading zeroes removed).

The dataset presents a gene expression dataset for rheumatoid arthritis patients where their C-reactive protein (CRP) levels have been measured (in mg/L). The first column of the file corresponds to the recorded CRP levels, and columns 2 to 1,501 correspond to the expressions of genes. The CRP levels have been scaled to have mean zero and standard deviation equal to 1. The column names correspond to real gene symbols. More information about gene symbols can be found at: `https://www.genecards.org`.

1. The researchers are interested in finding groups within the genes. Choosing your preferred clustering approach and measure for validating the number of groups, present the number of groups amongst the genes, and the number of genes in each group.

2. For the groups of the study, perform the following alternative analyses for testing the global null hypothesis: the group is not associated with CRP. The two approaches include:

   - Through a principal components regression. Clearly presenting the regression model applied and any choices made.
   - Through a self-contained method. Clearly presenting the chosen approach, models considered and any assumptions made.

   In your report present the models/ methods implemented. Present your findings and the significant groups. Are the results from the two different approaches in agreement?

Question 2: LASSO regression [30 marks]

The purpose of this question is to investigate the effect of correlation on the performance of LASSO for identifying the truly associated variables.

1. Describe the design of a simulation study for investigating the power of LASSO for identifying the truly associated variables when you analyse gene expression data. In your description include information on:

   (a) how you will generate the gene expression data with 500 samples and 3000 genes

   (b) how you will generate the response variable of the samples

   (c) how you will apply the LASSO method

   (d) how you will evaluate the performance of LASSO

2. Run the simulation study described in part 1. and illustrate the performance of LASSO as the correlation structure of the genes varies.

3. Modify appropriately the simulation study of part 2. to examine the effect of:

   (a) the number of samples in the study

   (b) the number of variables in the study

4. Prepare a Conclusions Section of few sentences that describes the findings of your simulation study and other parameters/quantities that you could have explored in addition to the ones considered in the conducted simulation study. Are your findings in agreement with the existing literature?

Question 3: Pathway analysis [10 marks]

Wang *et al.* (2007) presented gene set enrichment analysis (GSEA) for genome-wide association studies (GWAS). The following questions are related to the proposed approach and its steps.

1. Present the gene statistic proposed by the authors for representing the association of each gene with the response.

2. Discuss alternative gene statistics that the authors could have implemented. Discuss (1) how does the alternative statistic compare with the proposed one, and (2) if the researchers would need to amend any other steps of the proposed approach for implementing the alternative statistic.

3. As part of the proposed approach the authors have presented a permutation procedure. Describe the steps of the permutation procedure. Discuss its role and importance.

4. The authors discussed also an alternative permutation procedure for the proposed approach. Compare the two approaches, and discuss advantages and disadvantages of each approach [For this part you do not need to run any analysis].

References

[1] Wang, K., Li, M., and Bucan, M. (2007). *Pathway-Based Approaches for Analysis of Genomewide Association Studies*. The American Journal of Human Genetics, 81, 1278-1283, `https://www.sciencedirect.com/science/article/pii/S0002929707637756`