Assignment 1
COMP 565 ML in Genomics and Healthcare

This assignment is worth 8% of your total grade and due at **midnight on September 25, 2023**

Question 1 [2%] Implementing LD score regression

For a phenotype of interest, we have collected the marginal statistics $\tilde{\boldsymbol{\beta}}$ for $M = 4268$ SNPs and the $M \times M$ LD matrix $\mathbf{R}$ (i.e., pairwise SNP-SNP Pearson correlation). The marginal statistics are based on $N = 1000$ individuals. Download the marginal statistics and LD matrix from here:

`https://drive.google.com/drive/folders/1tq4bTdbsv1iwO4wHxq1smzoN9D5luapp?usp=sharing`

For this question, you may also assume there is no population stratification in this dataset. Both phenotype and genotype were standardized.

Implement the very basic LD score regression algorithm with a programming language of your choice (preferably Python or R) to estimate the heritability of the phenotype.

What's your estimate of the heritability?

Submit your answer to this question in iPython notebook with name `COMP565_A1_ldsr.ipynb` or R Markdown `COMP565_A1_ldsr.Rmd` on MyCourses. This way the TA can run your code to validate its output. Do not submit the data provided to you as long as you have the clear path to the data you run.

Question 2 [6%] Bayesian fine-mapping

For a phenotype of interest, we have identified a GWAS locus based on N=498 individuals, which harbour 100 SNPs. As shown in Figure 1, because of the extensive LD, identifying the

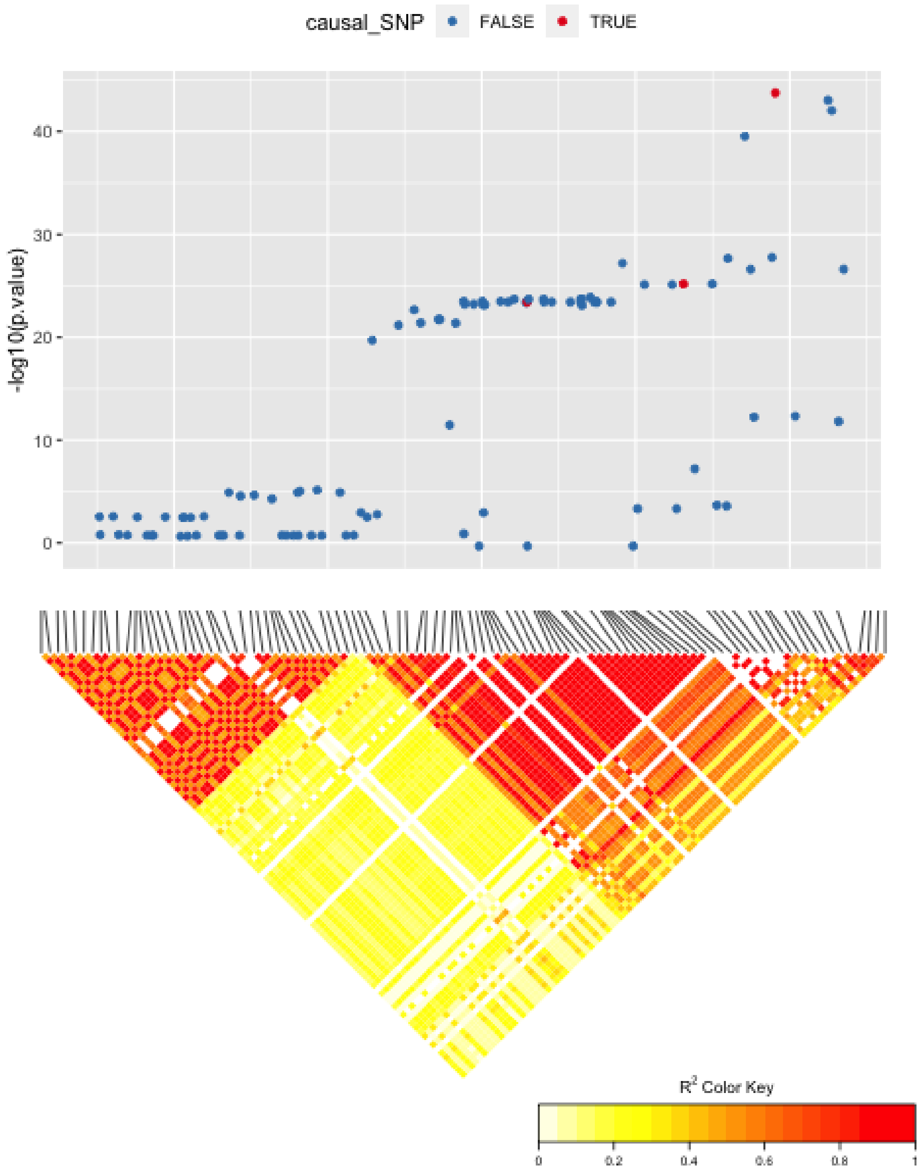

Figure 1: Manhattan plot for the GWAS locus to finemap. The causal SNPs are in fact coloured in red although in practice we will know which SNPs are causal.

causal SNPs based on the p-values of the z-scores alone is error prone. Because this is an assignment, I have highlighted the causal SNPs namely rs10104559, rs1365732, rs12676370 but of course in real world applications, we will not know them.

Download the marginal z-score and LD matrix from here:

`https://drive.google.com/drive/folders/1tr7BCceyIcKxiO_i6iCNjvk44HHpImgG?usp=sharing`

Your task is to implement a simplified version of the FINEMAP algorithm discussed in Lecture 5. To make the task easier, you may assume there are maximum 3 causal SNPs in the locus. You can divide the tasks into four small tasks:

1. (1%) Implement the efficient Bayes factor for each causal configurations:

   $$y = X\lambda + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I_n), \quad \lambda \sim \mathcal{N}(0, s_\lambda^2 \sigma^2 \Delta_\gamma)$$

   where s_λ^2 is user-defined prior variance in the unit of σ^2, Δ_γ is the diagonal matrix with diagonal equal to γ (causal configuration). You may assume that $s_\lambda^2 = 0.005$. Therefore, assuming there are k causal SNPs, then $\Sigma_{CC} = Ns^2 I_k = 2.49 I_k$

   For the multivariate Gaussian density function, we may find many existing libraries. In R, `mvtnorm`. In Python, it can be found in scipy `https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.multivariate_normal.html`

2. (1%) Implement the prior calculation for each configurations

3. (2%) Implement posterior inference over all possible configurations assuming at maximum 3 causal SNPs (i.e., $\binom{100}{1} + \binom{100}{2} + \binom{100}{3} = 100 + 4950 + 161700 = 166,750$ possible configurations). Therefore, *no* stochastic sampling is required here (as opposed to the original FINEMAP).

   To obtain all possible configurations, we may also use existing libraries. In R, this is done by `gtools:combinations`. In Python, check out `itertools.combinations` `https://docs.python.org/2/library/itertools.html`

   Some configurations may result in non-finite multivariate Gaussian density. Discard those configurations.

   Visualize the configuration posteriors by ranking them in increasing order as shown in Figure 2. As we can see, the vast majority of the configurations have very small posterior probabilities.

4. (2%) Implement posterior inclusion probabilities (PIP) to calculate SNP-level posterior probabilities.

   Visualize the normalized inferred PIP aligned with GWAS marginal -log10 p-values in Figure 3. It looks like we missed one of the 3 causal SNPs due to its nearly perfect LD with the other causal SNPs. But in general, we are able to pull down quite a few non-causal ones. That is, if we were going to experimentally validate the top SNPs, 2 out of 6 top SNPs based on PIP are true causal ones, whereas we would have got a lot more false positives if we were to follow the -log10 P-values instead.

Similar to Question 1, submit your code in `COMP565_A1_finemap.ipynb` or `COMP565_A1_finemap.Rmd` via myCourses. Your code should generate the plots illustrated in Figure 2 and 3. There may be some difference due to the numerical implementations of various MVN libraries but it should not differ too much from the provided PIP values in `SNP_pip.csv.gz`. This way the TA can run your notebook to validate its output. Do not submit the data provided to you as long as you have the clear path to the data you run in your notebook. You will be also evaluated based on the correctness of your code. Therefore, making your code readable is also very important.

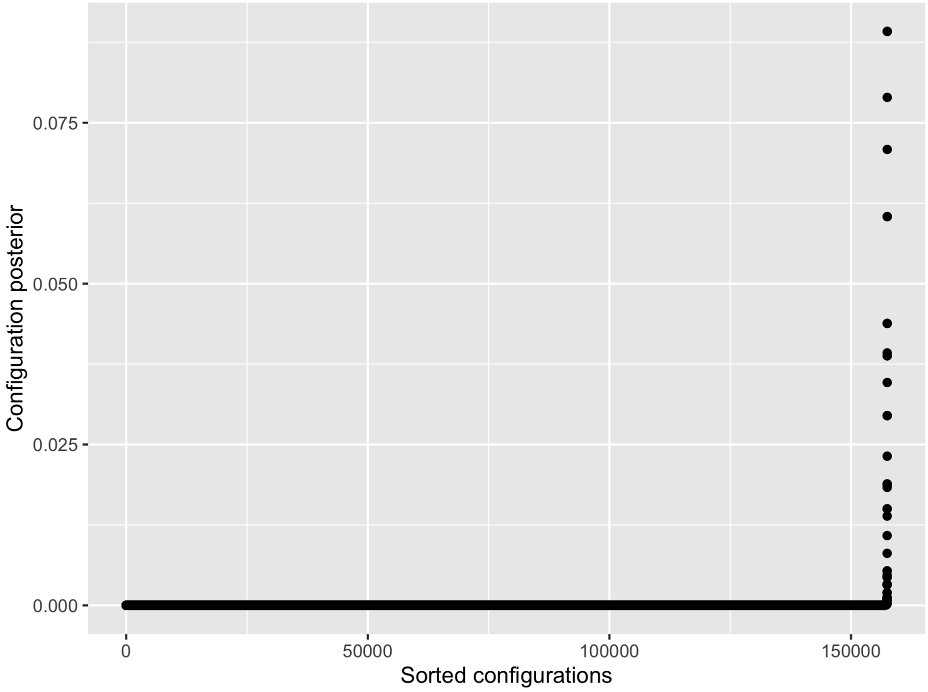

Figure 2: Posteriors of all of the valid configurations in increasing order.

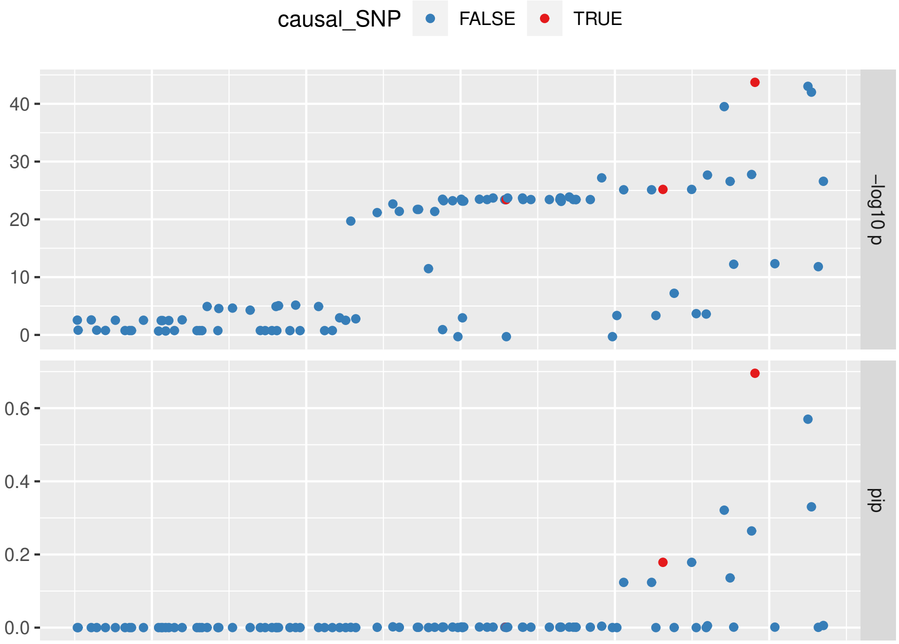

Figure 3: Inference results. Top panel. -log10 P-values of the marginal z-scores of the 100 SNPs. Bottom panel. The inferred posterior inclusion probabilities (PIP) of the 100 same SNPs.