STAT 447C Project Proposal: Comparing Methods for Linear Regression on High-Dimensional Molecular Data

Sarah Masri, 97415681

March 8, 2024

1 Abstract

Large-scale epigenetic analyses are an increasingly useful tool for prognostic prediction of disease and associated gene activation [1]. Problems involving high-dimensional gene expression data are often challenging by nature due the number of predictors being much larger than the number of observations. Regularized mechanisms with built-in variable selection are often employed to overcome this. The objective of this project is to compare performance and advantages of some regularized linear regression approaches to Bayesian hierarchical generalized linear models in high-dimensional inference. Namely, to compare elastic net regression to spike-and-slab regression in the context of prediction and gene detection for high-dimensional molecular data.

2 Project

This project explores the theme of a "careful and scientific comparison of a Bayesian estimator with another one, either Bayesian or non-Bayesian". This includes a motivation and background/review of literature of both elastic net regression and Bayesian regression using a spike-and-slab prior. If time permits, I would like to explore the possibility of proposing an extended technique from spike-and-slab regression.

3 Proposed Data

Proposed datasets are extracted from the NCBI Gene Expression Omnibus [`https://www.ncbi.nlm.nih.gov`]. If the reader wishes to avoid download the data, they can each be previewed on this projects public git repository [`https://github.com/sarahmasri/stat447C-Project`].

Breat cancer data: Single cell RNA sequencing of primary breast cancer in homo sapiens. Project investiage's gene expression from 515 single cell sequencing data from 11 breast cancer patients.
[`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75688`]

Melanoma data: Single cell RNA-seq analysis of melanoma in homo sapiens. The project investigate's the diversity of expression profiles in melanoma tumours.
[`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72056`]

References

[1] Zaixiang Tang, Yueping Shen, Xinyan Zhang, and Nengjun Yi. The spike-and-slab lasso generalized linear models for prediction and associated genes detection. *Genetics*, 205(1):77–88, January 2017.