Penalized Regression Methods for Interaction and Mixed-Effects Models with Applications to Genomic and Brain Imaging Data

Sahir Rai Bhatnagar

Doctor of Philosophy

Department of Epidemiology, Biostatistics and Occupational Health

McGill University
Montréal, Québec, Canada

July 2018

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Dedication

This thesis is dedicated to my family, Dadi, Papa, Maa, Sameer, Marie-Pierre, Louis, Mathieu, Chandni, Amir, Navdeep, Carlos, Gloria, and Karen.

Acknowledgements

I am most grateful to Celia Greenwood and Yi Yang for the supervision of this thesis, their advice and guidance not only in professional issues, but also in all other fundamental aspects.

Preface & Contribution of Authors

Manuscript 1: Bhatnagar SR, Yang Y, Khundrakpam B, Evans A, Blanchette M, Bouchard L, Greenwood CMT (2017). An analytic approach for interpretable predictive models in high dimensional data, in the presence of interactions with exposures. Genetic Epidemiology. Apr 1;42(3):233-49. DOI 10.1002/gepi.22112.
Software: `https://cran.r-project.org/package=eclust`
SRB, CMTG, YY, and MB contributed to the conceptualization of this research; SRB, LB, and BK contributed to the data curation; SRB contributed to the formal analysis, software, visualization; SRB and CMTG contributed to the methodology; SRB and CMTG contributed to writing the original draft; All authors contributed to editing the draft.

Manuscript 2: Bhatnagar SR, Yang Y, Lovato A, Greenwood CMT (2018+). Sparse Additive Interaction Models.
Software: `https://github.com/sahirbhatnagar/sail`

Manuscript 3: Bhatnagar SR, Oualkacha K, Yang Y, Forest M, Greenwood CMT (2018+). A General Framework for Variable Selection in Linear Mixed Models with Applications to Genetic Studies with Structured Populations.
Software: `https://github.com/sahirbhatnagar/ggmix`

Abstract

In high-dimensional (HD) data, where the number of covariates (p) greatly exceeds the number of observations (n), estimation can benefit from the bet-on-sparsity principle, i.e., only a small number of predictors are relevant in the response. This assumption can lead to more interpretable models, improved predictive accuracy, and algorithms that are computationally efficient. In genomic and brain imaging studies, where the sample sizes are particularly small due to high data collection costs, we must often assume a sparse model because there isn't enough information to estimate p parameters. For these reasons, penalized regression methods such as the lasso and group-lasso have generated substantial interest since they can set model coefficients exactly to zero. In the penalized regression framework, many approaches have been developed for main effects. However, there is a need for developing interaction and mixed-effects models. Indeed, accurate capture of interactions may hold the potential to better understand biological phenomena and improve prediction accuracy since they may reflect important modulation of a biological system by an external factor. Furthermore, penalized mixed-effects models that account for correlations due to groupings of observations can improve sensitivity and specificity. This thesis is composed primarily of three manuscripts. The first manuscript describes a novel strategy called `eclust` for dimension reduction that leverages the effects of an exposure variable with broad impact on HD measures. With `eclust`, we found improved prediction and variable selection performance compared to methods that do not consider the exposure in the clustering step, or to methods that use the original data as features. We further illustrate this modeling framework through the analysis of three data sets from very different fields, each with HD data, a binary exposure, and a phenotype of interest. In the second manuscript, we propose a method called `sail` for detecting non-linear interactions that automatically enforces the strong heredity property using both the ℓ_1 and ℓ_2 penalty functions. We describe a blockwise coordinate descent procedure for solving the objective function and provide performance metrics on both simulated and real data. The third manuscript develops a general penalized mixed model framework to account for correlations in genetic data due to relatedness called `ggmix`. Our method can accommodate several sparsity-inducing penalties such as the lasso, elastic net and group lasso and also readily handles prior annotation information in the form of weights. Our algorithm has theoretical guarantees of convergence and we again assess its performance in both simulated and real data. We provide efficient implementations of all our algorithms in open source software.

Abrégé

Il est aujourd'hui possible

Table of contents

List of Figures and Tables

Chapter 1

Literature Review

1.1 Introduction

Taken verbatim from Stephen Reid Tibs, Friedman (Reid, Tibshirani, & Friedman, 2016) Consider the linear model $Y = X\beta + \varepsilon$, where Y is an n-vector of independently distributed responses, X an $n \times p$ matrix with individual specific covariate vectors as its rows and ε an n-vector of i.i.d. random variables (usually assumed Gaussian) each with mean 0 and variance σ^2 . When $p > n$, one cannot estimate the unknown coefficient vector β uniquely via standard least squares methodology. In fact, it is probably ill-advised to use least squares to estimate the vector even when $p \leq n$ with p close to n, since standard errors are likely to be high and parameter estimates unstable. In this instance, if one can assume that β is reasonably sparse with many zero entries,

References

Reid, S., Tibshirani, R., & Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statistica Sinica*, 35–67.