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

SetDSpacefalse **DEDICATION**

This document is dedicated to the graduate students of the McGill University.

SetDSpacetrue

Acknowledgements

I am most grateful to Mathieu Blanchette and Rob Sladek for the supervision of this thesis, their advice and guidance not only in professional issues, but also in all other fundamental aspects. Many thanks to my PhD Committee: Jerome Waldispuhl, Doina Precup, Guillaume Bourque, and Derek Ruths for their helpful comments and suggestions.

I like to tank Douglas Ruden, Adrian Platts and Louis Letourneau for their insight and contributions to SnpEff and SnpSift projects.

I am grateful to Mark McCarthy, John Blangero and Mike Boehnke, and David Altshuler for their leadership in the T2D consortia. Special thanks to Pierre Fontanillas, Tanya Teslovich, Alisa Manning, Goo Jun, Anubha Mahajan, Jason Flannick, Andrew Morris, and Manuel Rivas for their helpful discussions that were instrumental in different aspects of this collaborative project.

I thank Fiona Cunningham, Will McLaren, and Kai Wang for their contributions to the VCF variant annotation standard as well as Sarah Hunt for her efforts on the GA4GH annotations specification.

Preface & Contribution of Authors

Manuscript 1: P. Cingolani, R. Sladek, and M. Blanchette. "BigDataScript: a scripting language for data pipelines." Bioinformatics 31.1 (2015): 10-16. For this paper, PC conceptualized the idea and performed the language design and implementation. RS & MB helped in designing robustness testing procedures. PC, RS & MB wrote the manuscript.

Manuscript 2: P. Cingolani, A. Platts, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D.M. Ruden, et al. "A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain $w^{1118}; iso-2; iso-3$". Fly, 6(2), 2012. In this paper, PC conceptualized the idea, implemented the program and performed testing. AP contributed several feature ideas, software testing and suggested improvements. XL, DR, SL, LW, TN, MC, LW performed mutagenesis and sequencing experiments. XL and DR performed the biological interpretation of the data. All authors contributed to the manuscript.

Manuscript 3: P. Cingolani, R. Sladek, and M. Blanchette. "A co-evolutionary approach for detecting epistatic interactions in genome-wide association studies". Ready for submission (data embargo restrictions). For this paper, PC designed the methodology under the supervision of MB and RS. PC implemented the algorithms. PC, RS & MB wrote the manuscript.

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 d'obtenir la séquence du génome de grandes cohortes d'individus, et cette information est permet de faciliter l'identification de variations génétiques liées à des maladies complexes. Dans ma thèse, j'étudie les défis informatiques et statistiques liés à l'analyse de grands ensembles de données génomiques. J'aborde trois aspects de l'analyse. Premièrement, afin d'analyser de grandes quantités de données provenant d'études génomiques nous concevons un langage de programmation, BigDataScript, qui simplifie la création de pipelines d'analyse de données robustes et évolutives. Deuxièmement, nous créons deux méthodes d'annotation et de classification de variantes génomiques (SnpEff et SnpSift) qui aident à prédire leur l'effet possible. Enfin, nous abordons le problème de l'identification de liens entre les maladies génétiques et les variantes qui les causent en proposant une méthodologie qui combine l'information génétique au niveau d'une la population avec informations évolutive afin d'augmenter la puissance statistique des études d'association considérant les interactions épistatiques.

Table of contents

List of Figures and Tables

Chapter 1

Epistatic GWAS analysis

1.1 Preface

In recent years over 80 genetic loci related to type II diabetes (T2D) have been identified