MASTER'S THESIS

BIVAS: a scalable Bayesian method for bi-level variable selection

Cai, Mingxuan

Date of Award:
2018

Abstract

In this thesis, we consider a Bayesian bi-level variable selection problem in high-dimensional regressions. In many practical situations, it is natural to assign group membership to each predictor. Examples include that genetic variants can be grouped at the gene level and a covariate from different tasks naturally forms a group. Thus, it is of interest to select important groups as well as important members from those groups. The existing methods based on Markov Chain Monte Carlo (MCMC) are often computationally intensive and not scalable to large data sets. To address this problem, we consider variational inference for bi-level variable selection (BIVAS). In contrast to the commonly used mean-field approximation, we propose a hierarchical factorization to approximate the posterior distribution, by utilizing the structure of bi-level variable selection. Moreover, we develop a computationally efficient and fully parallelizable algorithm based on this variational approximation. We further extend the developed method to model data sets from multi-task learning. The comprehensive numerical results from both simulation studies and real data analysis demonstrate the advantages of BIVAS for variable selection, parameter estimation and computational efficiency over existing methods. The BIVAS software with support of parallelization is implemented in R package 'bivas' available at https://github.com/mxcai/bivas.

Keywords: Bayesian variable selection; Variational inference; Group sparsity; Parallel computing.

Table of Contents

Declaration i

Abstract ii

Acknowledgements iii

Table of Contents iv

List of Figures vi

List of Tables vii

Chapter 1 Introduction 1

1.1 Overview 1

1.2 Bi-level variable selection 2

1.3 Outline of the thesis 3

Chapter 2 Statistical Models and Algorithms 4

2.1 Regression with BIVAS 4

2.1.1 Model setting 4

2.1.2 Algorithm 6

2.2 Multi-task learning with BIVAS 9

2.2.1 Model setting 9

2.2.2 Algorithm 10

2.3 Implementation details 11

2.4 Variable selection and prediction 13

2.5 Appendices 14

2.5.1 Variational EM Algorithm: Regression with BIVAS 14
2.5.2 Variational EM Algorithm: Multi-task Learning with BIVAS . 26

Chapter 3 Numerical Examples 35
3.1 Simulation study 35
3.2 Real data analysis 40
3.2.1 GWAS data 41
3.2.2 IMDB movie data 43

Chapter 4 Discussion 47

Curriculum Vitae 51