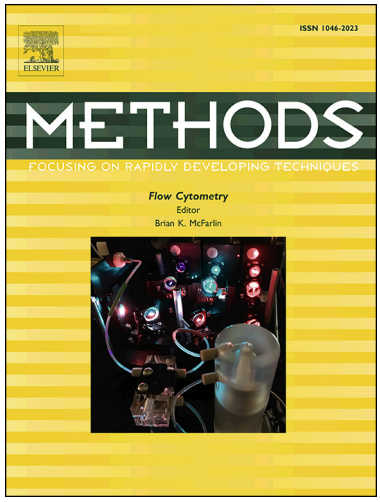
Accepted Manuscript



Variable Selection in Heterogeneous Datasets: A Truncated-rank Sparse Linear Mixed Model with Applications to Genome-wide Association Studies

Haohan Wang, Bryon Aragam, Eric P. Xing

PII: S1046-2023(17)30491-7
DOI: https://doi.org/10.1016/j.ymeth.2018.04.021
Reference: YMETH 4458

To appear in: *Methods*

Please cite this article as: H. Wang, B. Aragam, E.P. Xing, Variable Selection in Heterogeneous Datasets: A Truncated-rank Sparse Linear Mixed Model with Applications to Genome-wide Association Studies, *Methods* (2018), doi: https://doi.org/10.1016/j.ymeth.2018.04.021

Variable Selection in Heterogeneous Datasets: A Truncated-rank Sparse Linear Mixed Model with Applications to Genome-wide Association Studies

Haohan Wang[a], Bryon Aragam[b], Eric P. Xing[b,*]

[a] *Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA*
[b] *Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA*

Abstract

A fundamental and important challenge in modern datasets of ever increasing dimensionality is variable selection, which has taken on renewed interest recently due to the growth of biological and medical datasets with complex, non-i.i.d. structures. Naïvely applying classical variable selection methods such as the Lasso to such datasets may lead to a large number of false discoveries. Motivated by genome-wide association studies in genetics, we study the problem of variable selection for datasets arising from multiple subpopulations, when this underlying population structure is unknown to the researcher. We propose a unified framework for sparse variable selection that adaptively corrects for population structure via a low-rank linear mixed model. Most importantly, the proposed method does not require prior knowledge of sample structure in the data and adaptively selects a covariance structure of the correct complexity. Through extensive experiments, we illustrate the effectiveness of this framework over existing methods. Further, we test our method on three different genomic datasets from plants, mice, and human, and discuss the knowledge we discover with our method.

Keywords: Variable Selection, Genome-wide Association Study, Mixed Model, Heterogeneity, Confounding Correction

1. Introduction

Increasingly, modern datasets are derived from multiple sources such as different experiments, different databases, or different populations. In combining such heterogeneous datasets, one of the most fundamental assumptions in statistics and machine learning is violated: That observations are independent of one another. When a dataset arises from multiple sources, dependencies are introduced between observations from similar batches, regions, populations, etc. As a result, classical methods breakdown and novel procedures that can handle heterogeneous datasets and correlated observations are becoming more and more important.

In this paper, we focus on the important problem of variable selection in non-i.i.d. settings with possibly dependent observations. In addition to the aforementioned complications in analyzing

*Corresponding author
Email addresses: haohanw@cs.cmu .edu (Haohan Wang), epxing@cs.cmu.edu (Eric P. Xing)