DOCTORAL THESIS

Variable selection and dimension reduction in high-dimensional regression

Wang, Tao

Date of Award:
2013

Variable Selection and Dimension Reduction in High-dimensional Regression

WANG Tao

A thesis submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Principal Supervisor: Professor ZHU Lixing

Hong Kong Baptist University

August 2013

Abstract

In this thesis we develop some new methods for variable selection and dimension reduction in regression with high-dimensional data.

The first part of the thesis introduces a linear least squares framework for parameter estimation and variable selection in high-dimensional semi-parametric models. Firstly, we consider a general class of models with a single-index structure. We show that the coefficient vector from the least squares fit of any transformation of the response variable on the predictor vector can be used to identify the single-index parameter. Building upon the least squares formulation, we propose the non-convex penalized least squares estimation. We prove the oracle property of SCAD and the minimax concave penalty estimator when the number of predictors grows at some polynomial rate of the sample size. Secondly, we consider an additive multiple-index model in which each component function has a single-index structure, and of which the partially linear single-index model is a special case. Somewhat surprisingly, we show that all index vectors can be recovered through a single least squares coefficient vector, extending our previous result on the identifiability of single-index models. We propose the SCAD-penalized least squares estimation and establish the oracle property in sparse high-dimensional settings. As an application, for partially linear single-index models we develop a new two-stage estimation procedure that is iterative-free and easily implemented.

In the second part of the thesis, we establish a connection, from the viewpoint of dimension reduction, between classification and regression. Linear discriminant analysis (LDA) is a standard tool both for classification and dimension reduction. It is well-known that LDA can be re-formulated as a regression problem via optimal scoring. We show that sliced inverse regression (SIR), an innovative and effective method for dimension reduction in regression, can also be recast as an optimal scoring problem. Motivated by this regression interpretation of SIR, we propose a sparse and thus interpretable sufficient dimension reduction method that is applicable to high-dimensional settings. We also propose an alternating minimization algorithm for

efficiently solving the optimization problem.

In the third part, we consider the problem of semi-parametric variable selection for multiple-index regression models. We propose a knowledge-based method that takes into account predictor group information by combining a group-wise dimension reduction method and LASSO. Interestingly, the estimation via a LASSO penalty behaves like an adaptive LASSO estimation. We show that the new estimator is consistent in variable selection while retaining the root-n consistency. We use a Bayesian information criterion (BIC) type criterion to select the optimal tuning parameter. Moreover, we derive the consistency of the resulting BIC-type selector.

Keywords: Dimension reduction; High dimensionality; LASSO; Minimum average variance estimation; Multiple-index models; SCAD; Single-index models; Sliced inverse regression; Variable selection.

Table of Contents