Accepted Manuscript

Optimal estimation of slope vector in high-dimensional linear transformation models

Xin Lu Tan

PII: S0047-259X(17)30759-5
DOI: https://doi.org/10.1016/j.jmva.2018.09.001
Reference: YJMVA 4411

To appear in: *Journal of Multivariate Analysis*

Received date : 10 December 2017

Please cite this article as: X.L. Tan, Optimal estimation of slope vector in high-dimensional linear transformation models, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2018.09.001

Optimal estimation of slope vector in high-dimensional linear transformation models

Xin Lu Tan[a]

[a]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

Abstract

In a linear transformation model, there exists an unknown monotone, typically nonlinear, transformation function such that the transformed response variable is related to the predictor variables via a linear regression model. This paper presents CENet, a new method for estimating the slope vector and simultaneously performing variable selection in the high-dimensional sparse linear transformation model. CENet is the solution to a convex optimization problem which can be computed efficiently from an algorithm with guaranteed convergence to the global optimum. It is shown that when the joint distribution of the predictors and errors is elliptical, under some regularity conditions, CENet attains the same optimal rate of convergence as the best regression method in the high-dimensional sparse linear regression model. The empirical performance of CENet is shown on both simulated and real datasets. The connection of CENet with existing nonlinear regression/multivariate methods is also discussed.

Keywords: Canonical correlation analysis, Elastic net penalty, Elliptical distribution, Kendall's tau, Optimal rate of convergence, Variables transformation

1. Introduction

Recently, there has been significant interest in theoretically appealing and algorithmically efficient regression methods that are suitable for analyzing high-dimensional datasets whose dimension p is comparable or much larger than the sample size n. Under the linear regression model, researchers have proposed various computationally efficient regularization methods for simultaneously estimating the slope vector $\boldsymbol{\beta}^*$ and selecting predictors, and substantial progress has been made on understanding the theoretical properties of these regression methods; see, e.g., [4, 8, 10, 15, 25, 26, 59, 64, 65] and the references therein. In particular, when $\boldsymbol{\beta}^*$ is assumed to be sparse with a number s of non-zero coordinates such that $s = \|\boldsymbol{\beta}^*\|_0 \ll p$, the lasso estimator and the Dantzig selector have been shown to both attain an optimal rate of convergence $\sqrt{s\ln(p)/n}$ for estimating $\boldsymbol{\beta}^*$ [4].

When the relationship between the response variable and the predictors is nonlinear, however, the performance of regression methods based on linear models can be severely compromised. This paper is concerned with a family of sparse transformational regression models where a response variable $Y \in \mathbb{R}$ is related to p predictor variables $\mathbf{X} = (X_1, \ldots, X_p) \in \mathbb{R}^p$ through the equation

$$h^*(Y) = \mathbf{X}^\top \boldsymbol{\beta}^* + \epsilon, \quad \epsilon \mid \mathbf{X} \sim F(\epsilon). \tag{1}$$

Here h^* is an unknown, strictly increasing transformation function, $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)$ is a sparse slope vector of interest, and ϵ is a noise term independent of $\mathbf{X}$ with unknown distribution F.

Model (1) generalizes the linear regression model by allowing an unknown monotone transformation on the response variable Y. It has been widely studied in the classical $n \gg p$ setting [13, 16, 20, 22, 32], but the problem of estimating $\boldsymbol{\beta}^*$ in model (1) in the high-dimensional setting has received much less attention and a rate of convergence has not been derived. Han et al. [33] recently proposed a rank-based estimator for $\boldsymbol{\beta}^*$ and established consistency but

Email address: xinlutan92@gmail.com (Xin Lu Tan)