Journal of Multivariate Analysis 139 (2015) 266–282

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

High dimensional single index models

Peter Radchenko

University of Southern California, Bridge Hall 401, Los Angeles, CA, 90089, USA

ARTICLE INFO

Article history:
Received 29 April 2014
Available online 19 February 2015

AMS subject classification:
62J02
62J07
62F12

Keywords:
Nonlinear regression
High-dimensional data
Single index model
Sparsity
Regularization

ABSTRACT

This paper addresses the problem of fitting nonlinear regression models in high-dimensional situations, where the number of predictors, p, is large relative to the number of observations, n. Most of the research in this area has been conducted under the assumption that the regression function has a simple additive structure. This paper focuses instead on single index models, which are becoming increasingly popular in many scientific fields including biostatistics, economics and financial econometrics. Novel methodology is presented for estimating high-dimensional single index models and simultaneously performing variable selection. A computationally efficient algorithm is provided for constructing a solution path. Asymptotic theory is developed for the proposed estimates of the regression function and the index coefficients in the high-dimensional setting. An investigation of the empirical performance on both simulated and real data demonstrates strong performance of the proposed approach.

1. Introduction

Fields ranging from image processing and data compression to computational biology, climatology, economics and finance nowadays share the common feature of trying to extract information from vast noisy data sets. Such large scale problems may often be formulated under the framework of high-dimensional statistical regression, where the number of explanatory variables, p, is large relative to the number of observations, n. High-dimensional nonlinear regression is a research area that has recently generated a lot of interest with the emergence of powerful regularization methods. Due to the "curse of dimensionality" [1], most of the work on this subject has been performed under the assumption that the regression function has a simple additive structure. In other words, a typical regression model is

$$Y_i = \sum_{j=1}^{p} f_j^*(X_{ij}) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

Under the assumption that the number of "true" predictors is relatively small, this model has recently been extended into the high-dimensional setting. For example, the SpAM approach of Ravikumar et al. [28] fits a sparse additive model by imposing a penalty on the empirical L_2 norms of the functional components. Huang et al. [15] establish variable selection consistency for an adaptive variation on the SpAM approach using a B-spline implementation. Meier et al. [19] fit the same model as SpAM, but also incorporate a smoothness term in the penalty function, leading to interesting theoretical properties. Choi et al. [5] and Radchenko and James [26] extend this line of research to handle sparse high-dimensional models with interactions.

E-mail address: radchenk@usc.edu.

http://dx.doi.org/10.1016/j.jmva.2015.02.007
0047-259X/© 2015 Elsevier Inc. All rights reserved.

A different line of development uses models of the form

$$Y_i = f^*(X_i^T \boldsymbol{\alpha}^*) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2}$$

Here X_i is a realization of a p-dimensional predictor vector, and linear combination $X_i^T \boldsymbol{\alpha}^*$ is referred to as index. Eq. (2) defines the Single Index Model [11,16,10]. Single index models generalize linear regression by replacing the linear predictor with a semi-parametric component. Due to their flexibility and interpretability of the coefficients single index models are becoming increasingly popular in many scientific fields. Note that they are capable of modeling interactions among predictors, and thus serve as a useful alternative to additive models.

Single index models have been very popular in relatively low and moderate dimensional situations with a manageable number of predictors. The corresponding existing methods typically do not perform well in high-dimensional situations if applied directly. Limited research, however, has been conducted on extending these methods to the high-dimensional setting. This is mainly due to the non-convexity of the sum of squares with respect to the index coefficients; an issue that makes it very hard to either develop an efficient estimation algorithm or establish theoretical properties of the estimator. Wang and Yin [34] present an approach, SMAVE, which produces sparse index coefficient estimators by introducing L_1 regularization into the MAVE method of Xia et al. [36]. However, SMAVE cannot be implemented for $p > n$, and no high-dimensional asymptotic results have been established for it. Peng and Huang [24] estimate the single index model by minimizing a penalized least squares criterion, thus performing automatic variable selection. However, they focus on the case of n being larger than p and consider only fixed p asymptotics. Also, it is argued in Section 2 that such penalization may be problematic in high-dimensional situations due to the non-convexity of the sum of squares function.

Strong interest has been generated recently by the variable selection problem under the sufficient dimension reduction framework [6]. For example, Ni et al. [23] introduce L_1 regularization to sliced inverse regression; Zhou and He [39] use L_1 regularization together with thresholding for variable filtering; Li and Yin [18] implement a sliced inverse regression approach with L_2 and L_1 regularization; Bondell and Li [3] generalize the penalization idea to a family of inverse regression estimators; Zhu and Zhu [42] investigate variable selection for single-index models with a diverging number of predictors by using inverse regression; Wu and Li [35] establish asymptotic properties for a family of inverse regression estimators in the case where p is allowed to diverge; Wang et al. [32] analyze non-convex penalized estimation in high-dimensional models with single-index structure; Yu et al. [38] use the Dantzig selector approach, together with sliced inverse regression, to perform dimension reduction and predictor selection in semi-parametric models. The dimension reduction approaches are applicable to the estimation of index coefficients in single index models. However, they do not directly estimate the regression function, and are implemented under an additional linearity condition on the distribution of the predictors.

This paper considers a different approach and makes the following key contributions:

1. A new L_1 regularization method is introduced for efficiently estimating all components of the single index model and performing variable selection simultaneously. The method is designed for the high-dimensional setting, which includes situations where p is larger than n. The level of regularization is controlled by a tuning parameter, and an algorithm is presented for constructing a solution path with respect to this parameter in a computationally efficient manner.
2. Asymptotic theory is developed for the proposed estimates of the index coefficients and the regression function in the high-dimensional setting. In particular, under some assumptions, a polynomial rate of convergence for all estimators is established even in situations where p grows faster than n.

The organization of the paper is as follows. Section 2 introduces the new methodology, provides motivation for it and discusses the intuition behind it. A computationally efficient algorithm for constructing a solution path is presented in Section 3. Theoretical investigation is conducted in Section 4. Simulation and real data performance is discussed in Section 5, an extension to the Generalized Single Index Models is presented in Section 6 and concluding remarks are given in Section 7.

2. Methodology

This section presents a new approach for fitting the single index model, (2), in situations where the number of predictors, p, is large relative to the number of observations, n. As is very common in the high-dimensional statistical inference literature, it will be assumed that the true index incorporates only a small number of predictors, in other words $\boldsymbol{\alpha}^*$ is sparse. We will refer to the proposed method as HD-SIM, which stands for High Dimensional Single Index Models.

Let $\mathbf{Y}$ denote the response vector. For a given function f and vector $\boldsymbol{\alpha} \in \mathbb{R}^p$ we will define $\mathbf{f}_{\boldsymbol{\alpha}} = \big(f(X_1^T\boldsymbol{\alpha}), \ldots, f(X_n^T\boldsymbol{\alpha})\big)^T$. For a given $\boldsymbol{\alpha}$, candidate functions f will be chosen from a functional class $\mathcal{F}_n(\boldsymbol{\alpha})$. We will focus on cubic B-spline functions, but the proposed methodology works for other choices of $\mathcal{F}_n(\boldsymbol{\alpha})$. When $\boldsymbol{\alpha}$ is the jth coordinate vector, we will slightly abuse the notation and replace $\mathbf{f}_{\boldsymbol{\alpha}}$ and $\mathcal{F}_n(\boldsymbol{\alpha})$ with $\mathbf{f}_j$ and $\mathcal{F}_n(j)$, respectively. This is done for notational simplicity. Thus, we define $\mathbf{f}_j = \big(f(X_{1j}), \ldots, f(X_{nj})\big)^T$, where index j refers to the jth predictor. Formal definitions of functional classes $\mathcal{F}_n(\boldsymbol{\alpha})$ and $\mathcal{F}_n(j)$ will be given in Section 3.1. For the remainder of the paper $\|\cdot\|$ will refer to the usual Euclidean vector norm, while $\|\cdot\|_1$ will correspond to the L_1 vector norm.