Journal of Multivariate Analysis 109 (2012) 221–235

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

Non-convex penalized estimation in high-dimensional models with single-index structure

Tao Wang, Pei-Rong Xu, Li-Xing Zhu *

Hong Kong Baptist University, Hong Kong, China
East China Normal University, China

ARTICLE INFO

Article history:
Received 27 October 2011
Available online 29 March 2012

AMS 2000 subject classifications:
62H12
62G20

Keywords:
High-dimensional variable selection
Minimax concave penalty
Oracle property
Penalized least squares
SCAD
Single-index model

ABSTRACT

As promising alternatives to the LASSO, non-convex penalized methods, such as the SCAD and the minimax concave penalty method, produce asymptotically unbiased shrinkage estimates. By adopting non-convex penalties, in this paper we investigate uniformly variable selection and shrinkage estimation for several parametric and semi-parametric models with single-index structure. The new method does not need to estimate the involved nonparametric transformation or link function. The resulting estimators enjoy the oracle property even in the "large p, small n" scenario. The theoretical results for linear models are in parallel extended to general single-index models with no distribution constraint for the error at the cost of mild conditions on the predictors. Simulation studies are carried out to examine the performance of the proposed method and a real data analysis is also presented for illustration.

1. Introduction

Variable selection is a fundamental task for statistical modeling in high-dimensional settings, where the number of predictors is often comparable to, or even much larger than the total sample size. Traditional variable selection procedures follow either best subset selection or its stepwise variants. However, subset selection is computationally prohibitive when the number of predictors is large. Moreover, as analyzed by Breiman [6], subset selection may suffer from instability because of its inherent discreteness. To deal with these drawbacks, various penalized methods have been proposed during the past years to perform variable selection and shrinkage estimation simultaneously. In particular, the LASSO [36] and the SCAD [16] are two very popular methods with promising computational and statistical properties.

There is a huge literature devoted to studying the theoretical properties of the LASSO, particularly in the linear regression context. See, for instance, [25,15,42,41,3], among many others. Despite its popularity, the LASSO does suffer from several drawbacks, the most severe of which is its estimation bias. To this end, Fan and Li [16] proposed the SCAD in a general parametric framework. When the number of predictors is finite, they studied the oracle properties of general non-concave penalized likelihood estimators. Here, the oracle property means that the estimator is asymptotically as efficient as the ideal one assisted by an oracle who knows which coefficients are nonzero and which are zero. Their results were later extended by Fan and Peng [18] to the setting with a diverging number of predictors. Recently, Kim et al. [24] proved that for linear models, the oracle property of the SCAD continues to hold while the number of predictors can grow at a polynomial rate, up to exponentially fast, of the sample size. Other works on the advantages of penalized methods with non-convex penalties over the LASSO include [30,40]. In particular, Zhang [40] investigated in detail the properties of the minimax concave penalty

* Correspondence to: Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China.
E-mail address: lzhu@hkbu.edu.hk (L.-X. Zhu).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.03.009

approach. Fan and Lv [17] gave a selective overview on the theoretical properties as well as algorithmic implementations of penalized likelihood methods in high-dimensional settings. We mention that alternative sparsity promoting approaches, such as the PAC-Bayesian approach using sparsity favoring priors, have also been proposed and well-studied [12,13].

In most empirical applications of regression analysis, however, the working model such as the linear regression model is at best an approximation. Because of the so-called "curse of dimensionality", it is very difficult or even infeasible to formulate and then validate a parametric model with a large number of predictors. To mitigate the risk of model misspecification and to overcome the curse of dimensionality, semiparametric models have attracted much attention in the literature. Popular models include the response transformation model $g_1(Y) = \boldsymbol{\beta}^T X + \epsilon$ and the classical single-index model $Y = g_2(\boldsymbol{\beta}^T X) + \epsilon$. Here $g_1(\cdot)$ is an (unknown) monotone function, $g_2(\cdot)$ is an unknown link function, and ϵ is assumed to be independent of $X = (X_1, \ldots, X_p)^T$. See [21,22] for more details. Interestingly, these two classes of models are of a common feature in model structure: other than an unknown nonparametric model transformation or link function, the information of the response can be captured through a single linear combination of the predictors. We call this the single-index structure. In this paper, we consider the following class of models with the single-index structure

$$Y = g(\boldsymbol{\beta}^T X, \epsilon), \tag{1.1}$$

or equivalently

$$Y \perp\!\!\!\perp X | \boldsymbol{\beta}^T X, \tag{1.2}$$

where g is an unspecified bivariate function and $\perp\!\!\!\perp$ indicates independence. The statement is thus that, given $\boldsymbol{\beta}^T X$, the response variable Y and the predictor vector X are independent of each other. Many important regression models, including linear models and generalized linear models, naturally satisfy (1.2). Other examples are the transformation linear model and the classical single-index model mentioned above.

The family of general single-index models (1.2) have been well-studied in the literature. On one hand, promising methods for estimating the index $\boldsymbol{\beta}$ include least squares method [29], structural adaptation method [23,11], and those in the sufficient dimension reduction context, such as sliced inverse regression method [28], sliced average variance estimation method [10], directional regression method [27], discretization–expectation estimation method [43], and many others. On the other hand, some attempts have also been made to address the variable selection problem. Kong and Xia [26] and Naik and Tsai [31] proposed new selection criteria for variable selection in the classical single-index model. *See also* [2]. Other alternatives are model free, which typically integrate sufficient dimension reduction techniques with the regularization paradigm, see [4,39] and the references therein. In particular, Wu and Li [39] investigated the asymptotic properties of sufficient dimension reduction estimators equipped with a SCAD-type penalty, when the number of predictors diverges to infinity with the sample size. The approaches and results, however, cannot be directly extended to the "$p > n$" setting. Therefore, it is of great interest to see whether the model-based selection methods, such as those in [24,40], have their justifiable counterparts in the general setting of (1.2) where no parametric model is imposed.

In this article, we investigate index estimation and variable selection, with an emphasis on the latter, for the class of models (1.2) with high-dimensional predictors. First, we study the asymptotic properties of index estimation. For any bounded transformation of the response, we propose an index estimator and establish the consistency and asymptotic normality in the presence of a diverging number of predictors. Second, we briefly discuss the choice of response transformation and adopt a response–distribution transformation considered in [38]. Third, by introducing a non-convex penalty function, we consider the penalized least squares optimization. We prove the oracle property of the SCAD and the minimax concave penalty estimator, while we allow the number of predictors to grow at some polynomial rate of the sample size. Finally, we evaluate the finite sample performance of the proposed method through simulation studies as well as a real data analysis. All technical proofs are given in the Appendix.

2. Methodology and main results

2.1. Index estimation and asymptotics

We adopt the least squares approach to estimate the index $\boldsymbol{\beta}$. It is very simple to use. In addition, the proposed least squares estimation allows us to directly introduce the penalty function, as given in Section 2.3. Of course, before developing any justifiable variable selection procedure, it is important to establish the asymptotic properties for the unpenalized estimation. We shall address this problem in this subsection.

Let $\boldsymbol{\Sigma} = \mathrm{Cov}(X)$ and $\boldsymbol{\sigma} = \mathrm{Cov}\{X, h(Y)\}$ for a given function $h(\cdot)$ of the response. We assume that $\boldsymbol{\Sigma}$ is positive definite. Define $\boldsymbol{\beta}_h = \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}$ as the coefficient vector of the least squares type. The following proposition follows immediately from Theorem 2.1 in [29].

Proposition 1. *Assume that $E(X|\boldsymbol{\beta}^T X)$ is a linear function of $\boldsymbol{\beta}^T X$. Then $\boldsymbol{\beta}_h$ is proportional to $\boldsymbol{\beta}$, that is, $\boldsymbol{\beta}_h = \kappa_h \times \boldsymbol{\beta}$ for some constant κ_h.*

The design condition of Proposition 1, known as the linearity condition, is satisfied when X has an elliptical distribution. It is widely assumed in the sufficient dimension reduction literature, see [28,8,9], among others. Hall and Li [20] proved that, as p tends to infinity, such a linearity can hold to a reasonable approximation in many problems. Proposition 1 indicates that,