Journal of Multivariate Analysis 143 (2016) 383–393

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Nonconvex penalized reduced rank regression and its oracle properties in high dimensions

Heng Lian, Yongdai Kim *

University of New South Wales, Australia
Seoul National University, Republic of Korea

ARTICLE INFO

Article history:
Received 28 January 2015
Available online 19 October 2015

AMS subject classifications:
62J05

Keywords:
Model selection
Nonconvex penalty
Oracle property
Reduced rank regression

ABSTRACT

Sparse reduced rank regression achieves dimension reduction and variable selection simultaneously. In this paper, for a class of nonconvex penalties, we give sufficient conditions that guarantee the oracle estimator is a local minimizer and stronger conditions that guarantee it is a global minimizer, with probability tending to one in an ultra-high dimensional setting. We carry out simulations to investigate the performance of the estimator. A real data set is analyzed for illustration.

1. Introduction

Let $Y_i = (Y_{i1}, \ldots, Y_{iK})^\top$ and $X_i = (X_{i1}, \ldots, X_{ip})^\top$, $i = 1, \ldots, n$, be the response and the covariate vectors, respectively, where $K \leq p$. We allow K and p to depend on the sample size n to deal with high dimensional cases. The generative model is

$$Y_{ik} = X_i^\top \beta_k^* + \epsilon_{ik},$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, K$, where ϵ_{ik} are independent random variables with mean 0 and variance σ_k^2. Let $B^* = (\beta_1^*, \ldots, \beta_K^*)$ be the $p \times K$ matrix of the true regression coefficients. Without further constraint, this multivariate regression model can be estimated by the least squares procedure which reduces to apply least squares regression for each response separately. Reduced rank regression [1,16] assumes that $\text{rank}(B) \leq r$ for some positive integer r. Since such a low-rank matrix B can be written as $DA^\top$ where D and A are a $p \times r$ and a $K \times r$ matrix respectively, the rank constraint implies only r linear combinations of the p-dimensional predictors (the r columns of XD) suffice to predict the responses. Thus reduced rank regression is an effective dimension reduction method in multivariate regression. In practice, selection of r can be done by cross-validation or some information criteria. In this paper we assume r is known in our theoretical investigations and we give discussions about the choice of r in practice. Due to the reduced number of parameters, the estimated reduced rank regression is a more efficient estimator than the unrestricted estimator [2].

When p is large, it is desirable to remove those predictors that do not "explain" any of the responses from the regression model. Sparse penalized variable selection methods such as that uses the lasso penalty (least absolute shrinkage and selection operator, Tibshirani [22]), the SCAD penalty (smoothly clipped absolute deviation, Fan and Li [14]), or the MCP

* Correspondence to: Department of Statistics, Seoul National University, Seoul, Republic of Korea.
E-mail address: ydkim0903@gmail.com (Y. Kim).

http://dx.doi.org/10.1016/j.jmva.2015.09.023
0047-259X/

(minimax concave penalty, Zhang [23]) have received much attention recently. A penalized estimator for reduced rank regression is defined as

$$\widehat{B} = \text{argmin}_{B:\text{rank}(B)\le r}\|Y - XB\|^2 + 2n\sum_{j=1}^{p} J_\lambda(\|B_j\|), \tag{1}$$

with penalty J_λ, where λ is a tuning parameter and B_j is the jth row of B. Here Y is the $n \times K$ matrix whose ith row vector is Y_i and X is the $n \times p$ matrix whose ith row vector is X_i. $\|\cdot\|$ is the square root of the sum of squares of all entries, also called the Frobenius norm for matrices. Bunea et al. [7] and Chen and Huang [11] considered properties of the sparse reduced rank estimator with the lasso penalty and the adaptive lasso penalty, respectively. Other related works include Chen et al. [8,10]; She [21]. However, none of these demonstrated the oracle property of the sparse reduced-rank estimator. Here, by oracle property we mean the estimator (local minimizer or global minimizer of the objective function) is asymptotically the same as the oracle estimator $\widehat{B}^o$ which is defined as

$$\widehat{B}^o = \text{argmin}_B\|Y - XB\|^2,$$

subject to $\|B_j\| = 0$ for $j \in S(B^*)^c$ and $\text{rank}(B) \le r$, where $S(B) = \{j : \|B_j\| > 0\}$ for a given $p \times K$ matrix B. That is, the oracle estimator is the least square estimator subject to rank constraint when the zero rows of B^* are known in advance.

The aim of this paper is to prove that $\Pr(\widehat{B} = \widehat{B}^o) \to 1$, where $\widehat{B}$ is the global minimizer of $Q(B) = \|Y - XB\|^2 + 2n\sum_{j=1}^{p} J_\lambda(\|B_j\|)$ with $\text{rank}(B) \le r$. An easier problem is to investigate whether $\widehat{B}^o$ is a local minimizer of $Q(B)$, which we also address.

However, for high-dimensional models, it is difficult to consider $\widehat{B}$. Instead, we work with a nonconvex penalized estimator with sparsity constraint defined as follows. Let u be a given positive integer which serves as an upper bound of the number of nonzero coefficients. A nonconvex penalized estimator for reduced rank regression with sparsity constraint is defined as

$$\widehat{B}_u = \text{argmin}_{B:\text{rank}(B)\le r, |S(B)|\le u}\|Y - XB\|^2 + 2n\sum_{j=1}^{p_n} J_\lambda(\|B_j\|). \tag{2}$$

Sparsity constrained estimators are considered by most of literatures of the oracle property on high dimensions including Chen and Chen [9]; Kim and Kwon [18]; Kim et al. [19]; Zheng et al. [25]. In particular, the sparse constraint allows us to lower bound $\|X(B - B^*)\|^2/\|B - B^*\|^2$ by the smallest sparse eigenvalue of $X^\top X/n$ (see (B1) for the definition) if $S(B) \le u$. When $p > n$, the smallest eigenvalue of $X^\top X/n$ is zero but the sparse eigenvalue can be positive under mild assumptions, and thus is popularly used high-dimensional regression [4,3]. Of course we have $\widehat{B}_u = \widehat{B}$ if u is larger than $S(\widehat{B})$, but $S(\widehat{B})$ is not available before we obtain the estimator and thus it is more natural to directly impose this sparsity constraint. We are to prove $\Pr(\widehat{B}_u = \widehat{B}^o) \to 1$ under regularity conditions.

A main difficulty in proving the oracle property is that the parameter space is not convex due to the rank constraint. That is, even though $\text{rank}(B_1) \le r$ and $\text{rank}(B_2) \le r$, it is not necessarily true that $\text{rank}(B_1 + B_2) \le r$. Hence, standard techniques to prove the oracle property cannot be applied directly.

The paper is organized as follows. In Section 2, a class of penalties and computational algorithm are explained. In Section 3, the proofs of local and global oracle properties are given. In addition, the selection of r is discussed. Numerical results are given in Section 4 and conclusion follows in Section 5.

2. Nonconvex penalties and computational algorithm

We consider the class of nonconvex penalties which satisfy the following condition:

(A1) Let $\nabla_\lambda(\cdot)$ be the derivative of the penalty $J_\lambda(\cdot)$. Then, $\nabla_\lambda(\cdot)$ is nonnegative, nonincreasing and continuous on $(0, \infty)$. $\nabla_\lambda(0+) = \lambda$ and there exists $a > 0$ such that $\nabla_\lambda(t) = 0$ when $|t| \ge a\lambda$. Finally, there is a constant ν such that $J_\lambda(t) \ge \lambda t/2$ when $|t| \le \nu\lambda$.

This class includes two important penalties, the SCAD penalty [14], where

$$\begin{aligned} J_\lambda(t) = {} & \lambda t I\{0 \le t \le \lambda\} + [\{a\lambda(t - \lambda) - (t^2 - \lambda^2)/2\}/(a - 1) + \lambda^2]I\{\lambda \le t \le a\lambda\} \\ & + \{(a + 1)\lambda^2/2\}I\{t \ge a\lambda\} \quad (a > 2), \end{aligned}$$

and the MCP [23] where

$$J_\lambda(t) = \{\lambda t - t^2/(2a)\}I\{0 \le t \le a\lambda\} + (a\lambda^2/2)I\{t \ge a\lambda\} \quad (a > 1).$$

As in [5], we set $a = 4$ in the SCAD penalty and $a = 3$ in the MCP, which are also close to the suggested values in the original works that proposed these penalties. Previously proposed sparse reduced rank regression that used the lasso and the adaptive lasso penalty cannot satisfy $\widehat{B}^o = \widehat{B}$ exactly, due to the bias induced by the penalty. In addition, it is not obvious to choose a good initial solution for the adaptive lasso penalty for high dimensional models, and thus we do not consider those penalties in our theoretical investigations.