Journal of Computational and Applied Mathematics 260 (2014) 91–98

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

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

A note on the one-step estimator for ultrahigh dimensionality

Mingqiu Wang [a], Xiuli Wang [a], Xiaoguang Wang [b,*]

[a] School of Mathematical Sciences, Qufu Normal University, Shandong, 273165, PR China
[b] School of Mathematical Sciences, Dalian University of Technology, Dalian, 116023, PR China

ARTICLE INFO

Article history:
Received 2 March 2013
Received in revised form 23 July 2013

MSC:
33B15
26D15
60E15

Keywords:
Cross validation
High-dimensional data
Model selection consistency
One-step estimator
Partial orthogonality

ABSTRACT

The one-step estimator, covering various penalty functions, enjoys the oracle property with a good initial estimator. The initial estimator can be chosen as the least squares estimator or maximum likelihood estimator in low-dimensional settings. However, it is not available in ultrahigh dimensionality. In this paper, we study the one-step estimator with the initial estimator being marginal ordinary least squares estimates in the ultrahigh linear model. Under some appropriate conditions, we show that the one-step estimator is selection consistent. Finite sample performance of the proposed procedure is assessed by Monte Carlo simulation studies.

1. Introduction

Consider the linear regression model

$$Y_i = \beta_0 + \sum_{j=1}^{p_n} X_{ij}\beta_j + \varepsilon_i, \quad i = 1, \ldots, n,$$

where Y_i is the response variable, X_{ij} is the covariate or design variable and ε_i is the error term. In many applications, such as studies involving microarray or mass spectrum data, the total number of covariates p_n can be large or even much larger than n, but the number of important covariates is typically smaller than n. Without loss of generality, we assume that the outcome is centered and the predictors are standardized, i.e. $\sum_{i=1}^n Y_i = 0$, $\sum_{i=1}^n X_{ij} = 0$ and $n^{-1}\sum_{i=1}^n X_{ij}^2 = 1, j = 1, \ldots, p_n$, so the intercept β_0 is not included in the regression function.

Zou and Li [1] proposed the one-step sparse estimator, which is defined by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p_n} X_{ij}\beta_j\right)^2 + \sum_{j=1}^{p_n} p'_{\lambda_n}(|\tilde{\beta}_j|)|\beta_j|, \tag{1.1}$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$ is the initial value, and $p'_{\lambda_n}(\cdot)$ is the derivative of penalty function. Several important penalty functions have been proposed, and include the bridge [2] with $p_{\lambda_n}(|t|) = \lambda_n|t|^q$, the *least absolute shrinkage and selection operator* (Lasso, [3]) with $p_{\lambda_n}(|t|) = \lambda_n|t|$, the *smoothly clipped absolute deviation* (SCAD) penalty [4] with $p_{\lambda_n}(|t|, a)$

* Corresponding author. Tel.: +86 411 84708351 8040; fax: +86 411 84708354.
E-mail addresses: wangxg@dlut.edu.cn, wmq0829@gmail.com (X. Wang).

0377-0427/$ – see front matter
http://dx.doi.org/10.1016/j.cam.2013.09.037

$= \lambda|t|I(|t| < \lambda) + \{(a^2 - 1)\lambda^2 - [(a\lambda - |t|)_+]^2\}I(\lambda \leqslant |t|)/[2(a - 1)]$, the *minimax concave penalty* (MCP, [5]) with $p_{\lambda_n}(|t|, \gamma) = [\lambda_n|t| - t^2/(2\gamma)]I(|t| < \gamma\lambda_n) + \gamma\lambda_n^2/2I(|t| \geqslant \gamma\lambda_n)$, and so on.

High-dimensional data analysis has become increasingly frequent and important in diverse fields of sciences, engineering, and humanities. Much progress has been made in the ultrahigh dimensional linear model. Meinshausen and Buhlmann [6] and Zhao and Yu [7] studied the variable selection consistency of the Lasso when the number of covariates is much larger than the sample size. Huang, Horowitz and Ma [8] studied the bridge estimator in the sparse high dimensional linear model. Huang, Ma and Zhang [9] studied the asymptotic properties of the high dimensional adaptive Lasso estimator. Fan and Lv [10] proposed sure independence screening for high-dimensional regression problems. However, all the forgoing results only are suitable for a specific penalty. There is no general frame that can be suitable for various penalties. As suggested by Zou and Li [1], the initial value $\tilde{\boldsymbol{\beta}}$ in the objective function (1.1) is often chosen as the ordinary least squares estimate or maximum likelihood estimate. However, we cannot obtain these estimates in ultrahigh cases. So it is a challenge to study the theoretical properties of the one-step estimator in ultrahigh dimensionality.

Huang, Ma and Zhang [9] suggested that the marginal ordinary least squares estimates can be chosen as the initial values although they are not $\sqrt{n}$-consistent estimator of the parameters. In this paper, we study the one-step estimator with the initial estimator being marginal ordinary least squares estimates in the ultrahigh linear model. Under certain appropriate conditions, we show that the one-step estimator is selection consistent. Finite sample performance of the proposed procedure is assessed by Monte Carlo simulation studies.

The rest of the article is organized as follows. Section 2 states the results on the model selection under the partial orthogonality and some other appropriate conditions. In Section 3, we give some simulation studies to assess the performance of the proposed method. Section 4 gives some conclusions. Technical proofs are relegated to the Appendix.

2. Model selection consistency

Let the true parameter be $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p_n})^T$. Denote $A = \{j : \beta_{0j} \neq 0, j = 1, \ldots, p_n\}$, which are the indices of nonzero coefficients in the underlying model. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T, \boldsymbol{X}_j = (X_{1j}, \ldots, X_{nj})^T, \mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{p_n})$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$. The cardinality of A is denoted by $|A|$ and $|A| = k_n$. Let $\mathbf{X}_A = (\boldsymbol{X}_j, j \in A)$, $\Sigma = n^{-1}\mathbf{X}^T\mathbf{X}$ and $\Sigma_A = n^{-1}\mathbf{X}_A^T\mathbf{X}_A$. Let $\beta_{n*} = \min\{|\beta_{0j}|, j \in A\}$ and $\max_{j\in A}|\beta_{0j}| = O(1)$.

Assumption 1. There exists a constant c_0 such that the covariates with zero coefficients and those with nonzero coefficients are weakly correlated, i.e. $\left|n^{-1/2}\sum_{i=1}^n X_{ij}X_{ik}\right| \leqslant c_0,\ j \in A, k \notin A$.

The estimated marginal regression coefficient is

$$\tilde{\beta}_{nj} = \frac{\sum_{i=1}^n X_{ij}Y_i}{\sum_{i=1}^n X_{ij}^2} = \frac{1}{n}\boldsymbol{X}_j^T\boldsymbol{Y}.$$

Take $\eta_{nj} = E(\tilde{\beta}_{nj}) = n^{-1}\sum_{l\in A}(\sum_{i=1}^n X_{ij}X_{il})\beta_{0l}$. According to Assumption 1, we have

$$|\eta_{nj}| \leqslant \frac{1}{\sqrt{n}}\max_{l\in A}|\beta_{0l}|\sum_{l\in A}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ij}X_{il}\right| = \begin{cases} O(1), & j \in A, \\ O\left(\dfrac{k_n}{\sqrt{n}}\right) = o(1), & j \notin A. \end{cases}$$

For simplicity, we take $\eta_{nj} = 0$ for $j \notin A$. It is easy to know

$$\max_{1\leqslant j\leqslant p_n}|\tilde{\beta}_{nj} - \eta_{nj}| = O_P(n^{-k}), \quad \text{for } k < \frac{1}{2}.$$

It will be used in the proof of Theorem 2.1.

Consider the penalized objective function

$$Q_n(\boldsymbol{\beta}_n) = \frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_n\|^2 + \sum_{i=1}^{p_n} p'_{\lambda n}(|\tilde{\beta}_{nj}|)\cdot|\beta_{nj}| \tag{2.1}$$

$\hat{\boldsymbol{\beta}}_n = \arg\min\{Q_n(\boldsymbol{\beta}_n)\}$ is the one-step estimator. For any vector $\boldsymbol{x} = (x_1, x_2, \ldots)^T$, denote its sign vector by $\text{sgn}(\boldsymbol{x}) = (\text{sgn}(x_1), \text{sgn}(x_2), \ldots)^T$, with the convention $\text{sgn}(0) = 0$. Following Zhao and Yu [7], we write $\hat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0$ if and only if $\text{sgn}(\hat{\boldsymbol{\beta}}_n) = \text{sgn}(\boldsymbol{\beta}_0)$.

The following conditions are needed for the selection consistency.

(A1) Suppose that ε_i's are i.i.d. random variables, $E(\varepsilon_i) = 0$, $\text{Var}(\varepsilon_i) = \sigma^2, i = 1, \ldots, n$. Furthermore, suppose their tail probabilities satisfy $P(|\varepsilon_i| > x) \leqslant K\exp(-Cx^d)$ for constants $1 \leqslant d \leqslant 2$, C and K.