Statistics and Probability Letters 127 (2017) 173–177

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Simultaneous selection of predictors and responses for high dimensional multivariate linear regression

Baiguo An [*], Beibei Zhang

School of Statistics, Capital University of Economics and Business, Beijing, 100070, China

A R T I C L E I N F O

Article history:
Received 22 October 2016
Received in revised form 4 April 2017
Accepted 8 April 2017
Available online 19 April 2017

Keywords:
Canonical correlation
Group lasso
High dimensional
Multivariate linear regression
Variable selection

A B S T R A C T

Most existing variable selection methods for multivariate linear models focus only on predictor selection. In this article, we propose a two-step (double group lasso step and sparse canonical correlation step) method to conduct variable selection for predictors and responses simultaneously.

1. Introduction

Ordinary least squares estimation (OLS) is the most popular method for estimating the parameters for multivariate linear regression, but it has many drawbacks. First, it ignores the relationships among the responses. Second, OLS is unbiased but its variance may not be the smallest. Third, OLS is generally not sparse. Many improved penalized methods have been proposed to overcome the drawbacks of OLS. Applying a penalty to each row of the coefficient matrix can conduct predictor selection, such as simultaneous variable selection methods (SVS) (e.g., L_∞-SVS (Turlach et al., 2005) and L_2-SVS (Simila and Tikka, 2007)), RemMap (Peng et al., 2010), and SPLS (Chun and Keles, 2010). Low rank estimation is also a very popular approach. Thus, Yuan et al. (2007) proposed a rank reduction estimation method and Chen and Huang (2012) proposed a sparse reduced-rank method, which can guarantee the sparseness and rank reduction for the estimates.

However, most aforementioned methods only considered predictor selection. In many real data analysis (e.g., eQTL Sun et al., 2010), predictors and responses are both high dimensional. It is necessary to select important responses before further analysis. Su et al. (2016) proposed a sparse envelope model for response selection only. Hence, it is meaningful to develop a method for the simultaneous selection of both predictors and responses. An et al. (2013) proposed a sparse CCA method for selecting predictors and responses for multivariate linear models in large sample scenario. We here propose a two-step method for simultaneously selecting predictors and responses in high-dimensional scenario.

The remainder of the article is organized as follows. In Section 2, we describe our methodology. Section 3 presents simulation studies.

* Corresponding author.
E-mail address: anbg200@163.com (B. An).

http://dx.doi.org/10.1016/j.spl.2017.04.008
0167-7152/

2. Methodology

2.1. Notations and model

Let $(X_i^\top, Y_i^\top)^\top$ be the ith observation $(1 \le i \le n)$, $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ and $Y_i = (Y_{i1}, \ldots, Y_{iq})^\top \in \mathbb{R}^q$. Assume that $(X_i^\top, Y_i^\top)^\top$ with $1 \le i \le n$ are mutually independent with zero mean. We assume that

$$Y_i = B^\top X_i + \varepsilon_i, \tag{1}$$

where $B = (b_{kj}) \in \mathbb{R}^{p\times q}$, and $\varepsilon_i \in \mathbb{R}^q$ is random noise, which is independent of X_i. Let $b_{k.} = (b_{k1}, \ldots, b_{kq})^\top$ be the kth row of B, and the jth column of B is denoted by $b_{.j} = (b_{1j}, \ldots, b_{pj})^\top$. The true values of $B, b_{k.}, b_{.j}$ are denoted by $B^0, b^0_{k.}, b^0_{.j}$.

Obviously, only the predictors with nonzero $\|b^0_{k.}\|$ are relevant to Y_i, where $\|\cdot\|$ denotes the L_2 norm for a vector. Hence we define the predictor true model (PTM) as $\mathcal{M}_T = \{1 \le k \le p : \|b^0_{k.}\| > 0\}$. For the response, Y_{ij} is marginally independent of the predictor X_i if and only if the corresponding $b^0_{.j}$ is zero. Hence we define the response related model (RRM) as $\mathcal{N}_R = \{1 \le j \le q : \|b^0_{.j}\| > 0\}$. If $\mathcal{M}_T, \mathcal{N}_R$ are known, model (1) can be simplified as

$$Y_{i(\mathcal{N}_R)} = B^\top_{(\mathcal{M}_T,\mathcal{N}_R)} X_{i(\mathcal{M}_T)} + \varepsilon_{i(\mathcal{N}_R)}, \tag{2}$$

where $Y_{i(\mathcal{N}_R)}$ is the subvector of Y_i corresponding to the indices $\mathcal{N}_R$, and $X_{i(\mathcal{M}_T)}, \varepsilon_{i(\mathcal{N}_R)}$ are both defined similarly. $B_{(\mathcal{M}_T,\mathcal{N}_R)}$ is the submatrix of B corresponding to the row indices $\mathcal{M}_T$ and the column indices $\mathcal{N}_R$.

If $Y_{i(\mathcal{N}_R)}$ is still high dimensional, it is necessary to further reduce the elements of the responses considered in (2). Let $\mathcal{N}$ be an arbitrary subset of $\mathcal{N}_R$, and denote $\mathcal{N}_R \setminus \mathcal{N}$ by $\mathcal{N}^c$. We call $\mathcal{N}$ a sufficient response model (SRM) if $Y_{i(\mathcal{N}^c)}$ and $X_{i(\mathcal{M}_T)}$ are mutually independent conditioned on $Y_{i(\mathcal{N})}$. Obviously $\mathcal{N}_R$ is an SRM. We define the response true model (RTM) $\mathcal{N}_T$ as the intersection of all SRMs. Under certain regularity conditions, RTM is also an SRM and the smallest. It is sufficient to only consider the regression relationship between $X_{i(\mathcal{M}_T)}$ and $Y_{i(\mathcal{N}_T)}$ because all the information about $Y_{i(\mathcal{N}_T^c)}$ contained in $X_{i(\mathcal{M}_T)}$ is contained in $Y_{i(\mathcal{N}_T)}$. It is worthy to mention that our definition of $\mathcal{N}_T$ is different from that given by An et al. (2013). $\mathcal{N}_T$ was defined on Y_i by An et al. (2013), whereas we first discard the elements of Y_i that are independent of X_i, and then define $\mathcal{N}_T$ on $Y_{i(\mathcal{N}_R)}$. This is because an element of Y_i is independent of X_i, whereas it may be dependent on X_i given other elements of Y_i. The following example illustrates this characteristic: $Y_{i1} = \varepsilon_{i1}$, $Y_{i2} = X_{i1} + Y_{i1} + \varepsilon_{i2}$. Obviously, it is sufficient to only consider the relationships between Y_{i2} and X_i in this example. However, because Y_{i1} and X_i are dependent conditioned on Y_{i2}, then the final RTM will include Y_{i1} if we do not first discard Y_{i1} which is independent of X_i.

Su et al. (2016) only tried to identify the active responses that contribute to the material part. Our objective is to identify $\mathcal{M}_T$, $\mathcal{N}_R$ and $\mathcal{N}_T$. After obtaining them, it suffices to only study the relationships between $X_{i(\mathcal{M}_T)}$ and $Y_{i(\mathcal{N}_T)}$ in order to study the relationship between X_i and Y_i.

2.2. Estimating $\mathcal{M}_T$ and $\mathcal{N}_R$

Estimating $\mathcal{M}_T$ and $\mathcal{N}_R$ is equivalent to identifying the sparse structure of the coefficient matrix B. Let $Y = (Y_1, \ldots, Y_n)^\top$, $X = (X_1, \ldots, X_n)^\top$. The traditional least squares estimation solves $\hat{B}^{LS} = \arg\min_B \|Y - XB\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. However $\hat{B}^{LS}$ is generally not sparse, so it is impossible to estimate $\mathcal{M}_T$ and $\mathcal{N}_R$ based on $\hat{B}^{LS}$. Thus, we propose a sparse estimation method called double group lasso to identify the sparse structure of B, which solves the optimization problem

$$\widehat{B}^{GL}_{(\lambda_1,\lambda_2)} = \arg\min_B \frac{1}{2n}\|Y - XB\|_F^2 + \lambda_1 \sum_{j=1}^{q} \|b_{.j}\| + \lambda_2 \sum_{k=1}^{p} \|b_{k.}\|, \tag{3}$$

where the penalty terms $\lambda_1 \sum_{j=1}^q \|b_{.j}\|$ and $\lambda_2 \sum_{k=1}^p \|b_{k.}\|$ shrink columns and rows of B toward zero respectively.

2.3. ADMM algorithm for double group lasso

An alternating directions method of multipliers (ADMM) algorithm is proposed to solve the problem (3). One can refer to Boyd et al. (2011) for more details about ADMM. The problem (3) can be rewritten as

$$\min_{B,A} \frac{1}{2n}\|Y - XB\|_F^2 + \lambda_1 \sum_{j=1}^{q} \|b_{.j}\| + \lambda_2 \sum_{k=1}^{p} \|a_{k.}\| \quad \text{subject to } A = B, \tag{4}$$

where $a_{k.}$ is the kth row of matrix A. The scaled augmented Lagrangian (Boyd et al., 2011) for (4) is $L(B, A, C) = 1/(2n)\|Y - XB\|_F^2 + \lambda_1 \sum_{j=1}^q \|b_{.j}\| + \lambda_2 \sum_{k=1}^p \|a_{k.}\| + \rho/2\|A - B + C\|_F^2 - \rho/2\|C\|_F^2$, where C is the scaled dual variable. The ADMM