Journal of Multivariate Analysis 132 (2014) 138–150

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Lasso penalized model selection criteria for high-dimensional multivariate linear regression analysis

Shota Katayama [a,*], Shinpei Imori [b]

[a] Graduate School of Decision Science and Technology, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan

[b] Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

ARTICLE INFO

Article history:
Received 1 February 2014
Available online 23 August 2014

AMS 2000 subject classifications:
primary 62J05
secondary 62H12

Keywords:
Multivariate linear regression
Model selection
High-dimensional data
Consistency

ABSTRACT

This paper proposes two model selection criteria for identifying relevant predictors in the high-dimensional multivariate linear regression analysis. The proposed criteria are based on a Lasso type penalized likelihood function to allow the high-dimensionality. Under the asymptotic framework that the dimension of multiple responses goes to infinity while the maximum size of candidate models has smaller order of the sample size, it is shown that the proposed criteria have the model selection consistency, that is, they can asymptotically pick out the true model. Simulation studies show that the proposed criteria outperform existing criteria when the dimension of multiple responses is large.

1. Introduction

Multivariate linear regression is fundamental in statistical analysis, which is applied to biometrics, econometrics, marketing research, engineering, chemometrics and many other related research fields to study relationships between multiple responses and a set of predictors. Model selection criteria which identify the relevant predictors play an important role in these applications. By advances of information technology and data base system, however, the high-dimensional data in which the sample size n is comparable with the number of predictors p_n and the dimension of multiple responses q_n or larger than them often appear in these applications. For instance, in biometrics, one may wish to clarify the relationship between the RNA levels and the DNA copy number (Peng et al. [16]) or identify the differentially expressed genes among some groups (Xu and Cui [25]). In the former case, n is less than both p_n and q_n while n may be less than only q_n in the latter case.

Classical model selection criteria such as AIC (Akaike [1]), BIC (Schwarz [21]), GIC (Nishii [14]) including AIC and BIC as a special case and Mallows C_p criterion (Mallows [13]) are not applicable to the high-dimensional data since these methods have developed in the asymptotic framework $n \to \infty$ with both p_n and q_n are fixed. To handle the high-dimensionality, we need to consider an asymptotic framework where both p_n and q_n become to be large as n increases. Recently, Yamamura et al. [26] have derived an AIC type criterion when $(n, q_n) \to \infty$ and $n < q_n$ based on a ridge estimator proposed by Srivastava and Kubokawa [22] and Kubokawa and Srivastava [10]. Similarly, Kubokawa and Srivastava [11] have also derived an AIC type criterion and a Mallows C_p type criterion when $q_n/n \to \gamma \in (0, 1)$. On the other hand, Yanagihara et al. [27] have proved the consistency property of the classical AIC when $q_n/n \to \gamma \in (0, 1)$. It has not been shown whether or not the criteria given by Yamamura et al. [26] and Kubokawa and Srivastava [11] have the consistency property.

* Corresponding author.
E-mail addresses: katayama.s.ad@m.titech.ac.jp (S. Katayama), imori.stat@gmail.com (S. Imori).

http://dx.doi.org/10.1016/j.jmva.2014.08.002
0047-259X/

In the above literature on high-dimensional information criteria, they restrict that the size of considerable candidate models is not larger than n while n may be less than p_n. It is natural in high-dimensional data to assume that the number of important predictors is not so large even though p_n is larger than n. This type of restriction has also been imposed by Chen and Chen [3] and Kim et al. [9] in the linear regression model. Foygel and Drton [6] and Chen and Chen [4] have imposed it in the generalized linear regression model and the Gaussian graphical model, respectively.

In this paper, we propose two GIC type criteria with consistency property which allow $n < p_n$, $n < q_n$ and $q_n/n \to \infty$. To the best of our knowledge, there are no model selection criteria in such a case. In order to allow $n < p_n$ we search the true model among restricted candidate models with smaller size than n, similarly as the previous work. To allow the rest we use a sparse precision matrix estimator based on the penalized likelihood with the Lasso penalty on off-diagonal elements of precision matrices. In the last decades, various authors have investigated the estimation of sparse precision matrix with the Lasso penalty. For instance, Yuan and Lin [28], Friedman et al. [7] and Scheinberg et al. [19] have provided algorithms to optimize the Lasso penalized likelihood. Rothman et al. [17] and Lam and Fan [12] have studied asymptotic properties such as the convergence rate and the support recovery.

It is worth pointing out that the model selection is closely related to the tuning parameter selection in the penalized likelihood estimation since a model is selected when a tuning parameter is so. In the linear regression, Wang et al. [24] investigate the tuning parameter selector and its selection consistency while Zhang et al. [30] and Fan and Tang [5] do so in the generalized linear regression. In the multivariate linear regression, Rothman et al. [18] penalize both the coefficient matrix and the precision matrix with the Lasso penalty, and the tuning parameters for them are selected by the cross-validation. Similar procedures can be seen in Peng et al. [16] and Obozinski et al. [15], but there are no theoretical guarantees about the tuning parameter selection.

The organization of the paper is as follows. In Section 2, we introduce the classical GIC and then propose two GIC type criteria for high-dimensional data which we denote by high-dimensional GIC (HGIC), based on the Lasso penalized likelihood. The consistency property of the two HGIC under a high-dimensional asymptotic framework is provided in Section 3. In Section 4, we compare the two HGIC with existing criteria numerically. All the proofs are given in Section 5.

Here, we summarize the notations used throughout the paper. For a real matrix $\boldsymbol{A} = (a_{ij})$, we define the element-wise infinity norm as $\|\boldsymbol{A}\|_\infty = \max|a_{ij}|$, the element-wise ℓ_1 norm as $\|\boldsymbol{A}\|_1 = \sum|a_{ij}|$, the operator norm as $\|\boldsymbol{A}\| = \lambda_{\max}^{1/2}(\boldsymbol{A}^T\boldsymbol{A})$ and the Frobenius norm as $\|\boldsymbol{A}\|_F = (\sum a_{ij}^2)^{1/2}$ where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue. We also define the smallest eigenvalue as $\lambda_{\min}(\cdot)$. To denote the (i, j) element of $\boldsymbol{A}$, we write $(\boldsymbol{A})_{ij}$. When $\boldsymbol{A}$ is symmetric, we define $\boldsymbol{A}^- = \boldsymbol{A} - \mathrm{diag}(\boldsymbol{A})$ where $\mathrm{diag}(\boldsymbol{A})$ denotes the diagonal matrix of $\boldsymbol{A}$. For a set α, we write $|\alpha|$ to denote the cardinality of α.

2. HGIC in multivariate regression model

A multivariate linear regression model is given by

$$\boldsymbol{y}_i = \boldsymbol{B}^{*T}\boldsymbol{x}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iq_n})^T$ is a response vector, $\boldsymbol{B}^*$ is an unknown $p_n \times q_n$ non-random coefficient matrix, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip_n})^T$ is a set of predictors and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iq_n})^T$ is a random vector drawn from a q_n-dimensional multivariate normal distribution with the mean vector $\boldsymbol{0}$, the covariance matrix $\boldsymbol{\Sigma}^*$ and the precision matrix $\boldsymbol{\Omega}^* = \boldsymbol{\Sigma}^{*-1}$, which we denote $N_{q_n}(\boldsymbol{0}, \boldsymbol{\Sigma}^*)$ hereafter. Assume that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are independent and $\boldsymbol{\Sigma}^*$ is positive definite. Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ and $\boldsymbol{\mathcal{E}} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)^T$, then the model (2.1) can be given by the following matrix form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{\mathcal{E}}.$$

The log-likelihood function for the coefficient and the precision matrix multiplying $(-2)/n$ and ignoring the constant terms is given by

$$L(\boldsymbol{B}, \boldsymbol{\Omega}) = \frac{1}{n}\mathrm{tr}\boldsymbol{\Omega}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}) - \log\det(\boldsymbol{\Omega}),$$

where $\boldsymbol{B}$ is a $p_n \times q_n$ coefficient matrix and $\boldsymbol{\Omega}$ is a $q_n \times q_n$ precision matrix. Assume that the true coefficient matrix $\boldsymbol{B}^* = (b_{ij}^*)$ has a row support

$$\alpha_n^* = \{1 \le i \le p_n | b_{ij}^* \ne 0 \text{ for some } 1 \le j \le q_n\}$$

with $|\alpha_n^*|$ elements. Thus, the ith predictor is irrelevant for all responses if $i \notin \alpha_n^*$. Note that α_n^* is a subset of $\{1, \ldots, p_n\}$ and usually unknown. We wish to find the true support α_n^*, rather than the values of $\boldsymbol{B}^*$. This problem can be understood as a model selection problem to find the true model α_n^* from a set of possible candidate models α_n's where α_n is a subset of $\{1, \ldots, p_n\}$. The candidate model α_n can also be regarded as the row support of a coefficient matrix. Now we define the parameter spaces of $\boldsymbol{B}$ over the given candidate model α_n as

$$\Theta(\alpha_n) = \{\boldsymbol{B} = (b_{ij}) \in \mathbb{R}^{p_n \times q_n} | \ b_{i1} = \cdots = b_{iq_n} = 0 \text{ for } i \notin \alpha_n\}.$$

The parameter space of $\boldsymbol{\Omega}$ is also defined as

$$\Xi = \{\boldsymbol{\Omega} \in \mathbb{R}^{q_n \times q_n} | \boldsymbol{\Omega} = \boldsymbol{\Omega}^T\}.$$