Scientific Research Publishing

Advances in Pure Mathematics, 2023, 13, 284-302
https://www.scirp.org/journal/apm
ISSN Online: 2160-0384
ISSN Print: 2160-0368

A New Extended BIC and Sequential Lasso Regression Analysis and Their Application in Classification

Jie Chen, Wanzhou Ye

Department of Mathematics, College of Science, Shanghai University, Shanghai, China
Email: listenercjj@shu.edu.cn, wzhy@shu.edu.cn

How to cite this paper: Chen, J. and Ye, W.Z. (2023) A New Extended BIC and Sequential Lasso Regression Analysis and Their Application in Classification. *Advances in Pure Mathematics*, **13**, 284-302.
https://doi.org/10.4236/apm.2023.135020

Received: April 19, 2023
Accepted: May 28, 2023
Published: May 31, 2023

Abstract

In this paper, firstly, we propose a new method for choosing regularization parameter λ for lasso regression, which differs from traditional method such as multifold cross-validation, our new method gives the maximum value of parameter λ directly. Secondly, by considering another prior form over model space in the Bayes approach, we propose a new extended Bayes information criterion family, and under some mild condition, our new EBIC (NEBIC) is shown to be consistent. Then we apply our new method to choose parameter for sequential lasso regression which selects features by sequentially solving partially penalized least squares problems where the features selected in earlier steps are not penalized in the subsequent steps. Then sequential lasso uses NEBIC as the stopping rule. Finally, we apply our algorithm to identify the nonzero entries of precision matrix for high-dimensional linear discrimination analysis. Simulation results demonstrate that our algorithm has a lower misclassification rate and less computation time than its competing methods under considerations.

Keywords

Regularization Parameter, Sequential Procedure, BIC, Linear Discrimination Analysis, Feature Selection

1. Introduction

Sparse high-dimensional regression (SHR) models arise in many important contemporary scientific fields. A SHR model is:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i,\ i = 1, 2, \cdots, n \tag{1}$$

where the number of features p is much larger than the sample size n, and only a relatively small number of the β_j's are nonzero. Feature selection is crucial in the analysis of SHR models. Desboulets [1] pointed out that there are three main categories of variable selection procedures, they are test-based procedures, penalty-based procedures and screening-based procedures. Regularized regression approaches to the analysis of SHR models have attracted considerable attention of the researchers. A regularized regression approach selects the features and estimates the coefficients simultaneously by minimizing a penalized sum of squares of the form:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\sum_{j=1}^{p}p_\lambda\left(\left|\beta_j\right|\right) \tag{2}$$

where λ is a regulating parameter and p_λ is a penalty function such that the number of fitted nonzero coefficients can be regulated by λ; that is, only a certain number of β_j's are estimated nonzero when λ is set at a certain value. Various penalty functions have been proposed and studied, including Lasso: $p_\lambda\left(\left|\beta_j\right|\right)=\lambda\left|\beta_j\right|$, SCAD [2], which smoothly clips a L_1 penalty (for small $\left|\beta_j\right|$) and a constant penalty (for large $\left|\beta_j\right|$), adaptive Lasso [3]: $p_\lambda\left(\left|\beta_j\right|\right)=\lambda\omega_j\left|\beta_j\right|$, where ω_j are given weights, and MCP [4], which smoothly approaches the L_1 penalty from a constant penalty (for large $\left|\beta_j\right|$'s) by an asymptote.

Sequential methods have also received attention in recent decades for feature selection in SHR models. The traditional sequential procedures such as forward stepwise regression (FSR) were criticized for its greedy nature. However, it was discovered recently that the greedy nature is indeed a good one if the goal is to identify relevant features, see [5], especially, in the presence of high spurious correlations due to extremely high dimensionality of the feature space. A sequential procedure of a different nature called least angle regression (LAR) was proposed in [6]. The LAR continuously updates the estimate of the expected responses along a direction having equal angle with the features already selected and selects new features having the largest absolute correlation with the updated current residuals. Recently, Luo and Chen [7] identified the nonzero entries of the precision matrix by a sequential method called JR-SLasso proposed in [8]. Chen and Jiang [9] proposed a two-stage sequential conditional selection (TSCS) approach to the identification and estimation of the nonzeros of the coefficient matrix and the precision matrix. Sequential approach has also been considered for models other than SHR models. Besides their desirable theoretical properties, sequential approaches are computationally more appealing. They are more stable and less affected by the dimensionality of the feature space.

Apart from variable seletion, model selection is also an important part for regression analysis. In statistical modeling an investigator often faces the problem of choosing a suitable model from among a collection of viable candidates. Such a determination may be facilitated using a selection criterion, which assigns a score to every model in a candidate set based on some underlying statistical principle. The Bayesian information criterion (BIC) is one of the most widely

known and pervasively used tools in statistical model selection. However, as it was shown by [10] the ordinary Bayesian information criterion is too liberal that is the criteria select far more features than the relevant ones when the model space is large. Chen and Chen [10] proposed a family of extended Bayes information criteria (EBIC) to better meet the needs of variable selection for large model spaces.

Bsides regression, classification problem also receives widespread attention in statistical modeling. Classification problems with high-dimensional data rise in many important contemporary scientific fields such as genetics and medicine. The most popular method for classification is Fisher's linear discrimination analysis (LDA). In a K-class classification problem, the LDA assumes that the predictor $x=\left(x_1,\cdots,x_p\right)^{\mathrm{T}}$ given class $G=k$ follows the multivariate normal distribution $N\left(u_k,\Sigma\right)$, $k\in\left\{1,\cdots,K\right\}$. Let $\pi_k=\Pr\left(G=k\right)$ and:

$$d_{kj}\left(x\right)=\left[x-\left(u_k+u_j\right)/2\right]^{\mathrm{T}}\Omega\left(u_k-u_j\right)+\ln\left(\pi_k/\pi_j\right), \tag{3}$$

where $\Omega=\Sigma^{-1}$ is the so-called precision matrix. The Bayes rule which is theoretically optimal classifies x into class k if and only if:

$$d_{kj}\left(x\right)>0, \text{ for all } j\neq k.$$

Since the Bayes rule cannot be realized in practice due to the unknown u_k's and Ω, the Bayes rule is estimated in the LDA by replacing u_k's and Ω with their estimates. If the dimension of predictor p ($<n$) is fixed, or diverges under certain conditions, it has been shown that the LDA is asymptotically optimal but that, if $p > n$, it is asymptotically as bad as a random guessing, where n is the sample size, see [11]. The failure of LDA in the case $p > n$ is due to accumulated errors in the estimation of the unknowns, as argued in [12]. Thus, it is necessary to bypass the difficulties in the estimation of unknowns.

In this paper, firstly, we propose a novel method to choose the regularization parameter λ for lasso regression, traditional method such as multifold cross-validation chooses regularization parameter gradually, in details: for lasso regression, the larger regularization parameter λ is, the greater number of regression coefficients will be estimated zero, once λ exceeds the maximum value then there will be no features in the candidate model. Thus, it is important to know the exact maximum value of regularization parameter. However, multifold cross-validation method doesn't tell us the largest value of λ. Differing from multifold cross-validation, our method gives the exact maximum value of regularization parameter such that at least one of the β_j's will be estimated nonzero. Thus, if you have the data, you can get the largest λ for lasso regression immediately by our method. Then we apply this method to choose parameter for a sequence of partially penalized least squares problems, which means that the features selected in an earlier step are not penalized in the subsequent steps.

Secondly, the re-examination of BIC and EBIC prompts us naturally to consider other forms of prior over the model space in the Bayes approach. We pro-

pose a new extended Bayes information criterion family, and under some mild conditions our new EBIC (NEBIC) is shown to be consistent. Then our NEBIC is used as the stopping rule for sequential lasso algorithm, we dub the proposed procedure as sequential lasso with NEBIC. After that, we apply our algorithm to classification problem, the numerical study demonstrates that our algorithm performs well than its competing methods under consideration.

The remainder of the paper is arranged as follows. In Section 2, we introduce the basic properties of sequential lasso and our new method for choosing the regularization parameter λ. In Section 3, the selection consistency of our NEBIC is established. In Section 4, the structure of our algorithm is given. In Section 5, we apply our algorithm to classification problems, then the main method and simulation results are introduced sequentially.

2. Sequential Lasso Regression for Feature Selection

Let $X=\left(x_1,x_2,\cdots,x_p\right)$ be the $n\times p$ design matrix, for $j=1,\cdots,p$, $x_j=\left(x_{1j},x_{2j},\cdots,x_{nj}\right)^{\mathrm{T}}$ be the observation vector of predictor j on n individuals, $y=\left(y_1,y_2,\cdots,y_n\right)^{\mathrm{T}}$ be the response vector, and x_j,y all have been standardized, such that $x_j^{\mathrm{T}}1=0$, $y^{\mathrm{T}}1=0$, and $y^{\mathrm{T}}y=n$, $x_j^{\mathrm{T}}x_j=n$. Thus, in model (1) the intercept β_0 can be omitted. Let $\beta=\left(\beta_1,\beta_2,\cdots,\beta_p\right)^{\mathrm{T}}$, $\epsilon=\left(\epsilon_1,\epsilon_2,\cdots,\epsilon_n\right)^{\mathrm{T}}$, in matrix notation, model (1) is expressed as:

$$y=X\beta+\epsilon \tag{4}$$

Let S denote the set of indices $\{1,2,\cdots,p\}$, let s be any subset of S. Denote by $X(s)$ the matrix consisting of the columns of X with indices in s. Similarly, let $\beta(s)$ denote the vector consisting of the corresponding components of β. Let $\mathcal{R}(s)$ be the linear space spanned by the columns of $X(s)$ and $H(s)$ its corresponding projection matrix:

$$H(s)=X(s)\left[X^{\mathrm{T}}(s)X(s)\right]^{-1}X^{\mathrm{T}}(s).$$

At the initial step, sequential lasso minimizes the following penalized sum of squares:

$$l_1=\left\|y-X\beta\right\|_2^2+\lambda_1\sum_{j=1}^{p}\left|\beta_j\right| \tag{5}$$

Let $\lambda_1=2\max_{j\in\{1,2,\cdots,p\}}\left|x_j^{\mathrm{T}}y\right|$, from Theorem 1 we can prove that λ_1 is the largest value of the penalty parameter such that at least one of the β_j's, $j\in\{1,2,\cdots,p\}$, will be estimated nonzero. The features with nonzero estimated coefficients are selected and the set of their indices is denoted by s_{*1}.

For $k\geq 1$, let s_{*k} be the index set of the features selected until step k. At step $k+1$, sequential lasso minimizes the following penalized sum of squares:

$$l_{k+1}=\left\|y-X\beta\right\|^2+\lambda_{k+1}\sum_{j\in s_{*k}^{\mathrm{C}}}\left|\beta_j\right|. \tag{6}$$

From the proposition 1 we can prove that minimization of l_{k+1} is equivalent

to the minimization of

$$\tilde{l}_{k+1} = \left\| \tilde{y} - \tilde{X}\tilde{\beta} \right\|_2^2 + \lambda_{k+1} \sum_{j \in s_{*k}^{\complement}} \left| \beta_j \right|,$$

where $\tilde{y} = \left[I - H\left(s_{*k} \right) \right] y$, $\tilde{X} = \left[I - H\left(s_{*k} \right) \right] X\left(s_{*k}^{\complement} \right)$, $\tilde{\beta} = \beta\left(s_{*k}^{\complement} \right)$, $\lambda_{k+1} = 2\max_{j \in s_{*k}^{\complement}} \left| \tilde{x}_j^{\mathrm{T}} \tilde{y} \right|$, $\tilde{x}_j = \left[I - H\left(s_{*k} \right) \right] x_j$, from Theorem 1 we can prove that λ_{k+1} is the largest value of the penalty parameter such that at least one of the β_j's, $j \in s_{*k}^{\complement}$, will be estimated nonzero.

Next, we give the statements and proofs of proposition 1 and theorem 1.

Proposition 1. For $k \geq 1$, the minimization of l_{k+1} is equivalent to the minimization of $\tilde{l}_{k+1}$.

Proof: Differentiating l_{k+1} with respect to $\beta\left(s_{*k} \right)$, we have

$$\frac{\partial l_{k+1}}{\partial \beta\left(s_{*k} \right)} = -2X^{\mathrm{T}}\left(s_{*k} \right) y + 2X^{\mathrm{T}}\left(s_{*k} \right) X\left(s_{*k} \right) \beta\left(s_{*k} \right) + 2X^{\mathrm{T}}\left(s_{*k} \right) X\left(s_{*k}^{\complement} \right) \beta\left(s_{*k}^{\complement} \right).$$

Setting the above derivative to zero, we obtain:

$$\hat{\beta}\left(s_{*k} \right) = \left[X^{\mathrm{T}}\left(s_{*k} \right) X\left(s_{*k} \right) \right]^{-1} X^{\mathrm{T}}\left(s_{*k} \right) \left[y - X\left(s_{*k}^{\complement} \right) \beta\left(s_{*k}^{\complement} \right) \right].$$

Substituting $\hat{\beta}\left(s_{*k} \right)$ into $\left\| y - X\beta \right\|^2$ we have:

$$\begin{aligned} l_{k+1} &= \left\| y - X\beta \right\|_2^2 + \lambda_{k+1} \sum_{j \in s_{*k}^{\complement}} \left| \beta_j \right| \\ &= \left\| y - \left(X\left(s_{*k} \right) \hat{\beta}\left(s_{*k} \right) + X\left(s_{*k}^{\complement} \right) \beta\left(s_{*k}^{\complement} \right) \right) \right\|_2^2 + \lambda_{k+1} \sum_{j \in s_{*k}^{\complement}} \left| \beta_j \right| \\ &= \left\| \left[I - H\left(s_{*k} \right) \right] \left[y - X\left(s_{*k}^{\complement} \right) \beta\left(s_{*k}^{\complement} \right) \right] \right\|_2^2 + \lambda_{k+1} \sum_{j \in s_{*k}^{\complement}} \left| \beta_j \right| \\ &= \left\| \tilde{y} - \tilde{X}\tilde{\beta} \right\|_2^2 + \lambda_{k+1} \sum_{j \in s_{*k}^{\complement}} \left| \beta_j \right| = \tilde{l}_{k+1}. \end{aligned}$$

□

Theorem 1. $\lambda_1 = 2\max_{j \in \{1,2,\cdots,p\}} \left| x_j^{\mathrm{T}} y \right|$ is the largest value of the regularization parameter such that at least one of the β_j's, $j \in \{1,2,\cdots,p\}$, will be estimated nonzero.

Proof: this statement is equivalent to that λ_1 is the largest value of the regularization parameter such that the minimization of l_1 will obtain nonzero solutions.

By the KKT condition, let $\frac{\partial l_1}{\partial \beta} = 0$ we have:

$$2X^{\mathrm{T}}\left(y - X\hat{\beta} \right) = \lambda_1 \partial \left\| \hat{\beta} \right\|_1 \tag{7}$$

In component form:

$$2x_j^{\mathrm{T}}\left(y - X\hat{\beta} \right) = \lambda_1 \partial \left| \hat{\beta}_j \right|, j \in \{1,2,\cdots,p\}. \tag{8}$$

where $\partial \left| \hat{\beta}_j \right|$ is a sub gradient of $\left| \beta_j \right|$ at $\hat{\beta}_j$, according to the value of $\hat{\beta}_j$, there are three situations:

$$\begin{aligned} &\partial \left| \hat{\beta}_j \right| \in (-1,1), \text{ when } \hat{\beta}_j = 0 \\ &\partial \left| \hat{\beta}_j \right| = 1, \text{ when } \hat{\beta}_j > 0; \\ &\partial \left| \hat{\beta}_j \right| = -1, \text{ when } \hat{\beta}_j < 0. \end{aligned}$$

Then (8) is equivalent to:

$$2x_j^{\mathrm{T}}\left(y - X\hat{\beta}\right) = \pm\lambda_1, \text{ when } \hat{\beta}_j \neq 0 \tag{9.1}$$

or

$$\left|2x_j^{\mathrm{T}}\left(y - X\hat{\beta}\right)\right| < \lambda_1, \text{ when } \hat{\beta}_j = 0. \tag{9.2}$$

The statement "λ_1 is the largest value of the penalty parameter such that the minimization of l_1 will obtain nonzero solutions." is established by showing:

(i) Let $\lambda_1 = 2\max_{j\in\{1,2,\cdots,p\}}\left|x_j^{\mathrm{T}}y\right|$, $\min_\beta \left\|y - \sum_{j=1}^p \beta_j x_j\right\|_2^2 + \lambda_1\sum_{j=1}^p\left|\beta_j\right|$ will obtain nonzero solutions.

(ii) Let $\lambda > \lambda_1$, $\min_\beta \left\|y - \sum_{j=1}^p \beta_j x_j\right\|_2^2 + \lambda\sum_{j=1}^p\left|\beta_j\right|$ will only obtain zero solutions.

According to the proof by contradiction, we suppose that there are only zero solutions when $\lambda_1 = 2\max_{j\in\{1,2,\cdots,p\}}\left|x_j^{\mathrm{T}}y\right|$.

Substituting $\hat{\beta} = 0$ into (9.2) we have: $\left|2X^{\mathrm{T}}\left(y - X*0\right)\right| < \lambda_1$, which means:

$$\left|2x_j^{\mathrm{T}}y\right| < \lambda_1, j \in \{1,2,\cdots,p\}.$$

This contradicts with $\lambda_1 = 2\max_{j\in\{1,2,\cdots,p\}}\left|x_j^{\mathrm{T}}y\right|$, thus the supposition is wrong, (i) is proved.

Now let us turn to (ii):

Let $l = \min_\beta \left\|y - \sum_{j=1}^p \beta_j x_j\right\|_2^2 + \lambda\sum_{j=1}^p\left|\beta_j\right|$, and let $\frac{\partial l}{\partial\beta} = 0$ we have:

$$2X^{\mathrm{T}}\left(y - X\hat{\beta}\right) = \lambda\partial\left\|\hat{\beta}\right\|_1 \tag{10}$$

Then, the proof of (ii) is equivalent to prove this:

$$\left|2x_j^{\mathrm{T}}\left(y - X*0\right)\right| < \lambda, j \in \{1,2,\cdots,p\}$$

Substituting $\hat{\beta} = 0$ into the left of (10) we have:

$$\left|2x_j^{\mathrm{T}}\left(y - X*0\right)\right| = \left|2x_j^{\mathrm{T}}y\right|, j \in \{1,2,\cdots,p\}$$

By $\lambda_1 = 2\max_{j\in\{1,2,\cdots,p\}}\left|x_j^{\mathrm{T}}y\right|$ we have: $\left|2x_j^{\mathrm{T}}y\right| \le \lambda_1$, since $\lambda > \lambda_1$, then for all $j \in \{1,2,\cdots,p\}$, we have: $\left|2x_j^{\mathrm{T}}\left(y - X*0\right)\right| \le \lambda_1 < \lambda$, thus $\hat{\beta} = 0$ solves equation (10), (ii) is proved.

□

3. EBIC and New EBIC

Suppose the dimension of model space S is P, denote S_j is the collection of all models with j covariates, so that the model space S can be partitioned into $S = \bigcup_{j=1}^P S_j$, such that models within each S_j have equal dimension. Let $\tau\left(S_j\right)$ be the size of S_j, we know that $\tau\left(S_j\right) = C_P^j$. For example, suppose the number of covariates under consideration is $P = 1000$, the class of models containing a single covariate is denoted by S_1, then $S_1 = \left\{\{1\},\{2\},\cdots,\{P\}\right\}$ has size $\tau\left(S_1\right) = C_P^1 = 1000$, while the class of models containing two covariate is denoted by S_2, $S_2 = \left\{\{1,2\},\{1,3\},\cdots,\{1,P\},\cdots\right\}$ has size $\tau\left(S_2\right) = C_P^2 = \left(1000\times 999\right)/2$. We can see that the size of S_2 is much bigger than the size of S_1. Now let us

consider the prior distribution over S as follows.

For $s \in S$, we have $pr(s) = pr(S_j) * pr(s|S_j)$, $j = 1,2,\cdots,P$, since models within each S_j have equal dimension, it is reasonable to assign an equal probability $pr(s|S_j) = \frac{1}{\tau(S_j)}$ for any $s \in S_j$. The ordinary Bayesian information criterion assigns probabilities $pr(S_j)$ proportional to $\tau(S_j)$, that is $pr(S_j) \propto c * \tau(S_j)$. However, this would cause unreasonable situation by large model space. For example, as we have discussed in the above paragraph, we can see that $\tau(S_2)$ is 999/2 times bigger than $\tau(S_1)$, according to the constant prior by BIC, the probability assigned to S_2 is also 999/2 times that assigned to S_1. According to the knowledge of combinatorial number, the size of S_j increases as j increases to $j = P/2 = 500$, so that the probability assigned to S_j by the prior increases almost exponentially. In other word, models with larger number of covariates, 50 or 100 say, receive much higher probabilities than models with fewer covariates. This is obviously unreasonable, being strongly against the principle of parsimony.

Instead of assigning probabilities $pr(S_j)$ proportional to $\tau(S_j)$, as in the ordinary BIC, Chen and Chen [10] assign $pr(S_j)$ proportional to $\tau^{\varepsilon}(S_j)$ for $\varepsilon \in [0,1]$. This results in the prior probability $pr(s) = pr(S_j) * pr(s|S_j) = \tau^{\varepsilon-1}(S_j) = \tau^{-\gamma}(S_j)$, $\gamma = 1-\varepsilon$. This prior distribution gives rise to an extended BIC family (EBIC).

Notice that extended Bayesian information criterion is established by introducing the function $x^{\varepsilon}(x > 1, 0 < \varepsilon < 1)$, which aims to select models with fewer covariates. From the perspective of function, we know that $x^{\varepsilon}(x > 1, 0 \le \varepsilon \le 1)$ is a monotone increasing convex function, and the parameter ε is confined within [0, 1] to ensure upper convex property satisfied. Inspired by this, we consider other upper convex function like $\frac{x}{x+a}$, and the parameter a can be any positive numbers. In other word, we assign $pr(S_j)$ proportional to $\frac{\tau(S_j)}{\tau(S_j)+a}$, so that

$pr(s) = pr(S_j) * pr(s|S_j) = \left(\frac{\tau(S_j)}{\tau(S_j)+a}\right) * \frac{1}{\tau(S_j)} = \frac{1}{\tau(S_j)+a}$, this type of prior distribution on the model space gives rise to a new EBIC family as follows:

$$BIC_a(s) = -2\ln L_n\left\{\hat{\theta}(s)\right\} + v(s)\ln n + 2\ln\left(\tau(S_j)+a\right), a > 0 \tag{11}$$

where $\hat{\theta}(s)$ is the maximum likelihood estimator of $\theta(s)$ given model s. Now let us investigate the properties of our new EBIC for feature selection in linear regression models. Under some mild conditions our new EBIC is shown to be consistent.

Let y_n be the vector of n observations on the response variable, let X_n be the corresponding design matrix with all the covariates of concerns, and let β be the vector of regression coefficients. Assume that:

$$y_n = X_n\beta + e_n, \tag{12}$$

where $e_n \sim N(0, \sigma^2 I_n)$ and I_n is the identity matrix of size n. Let s_0 be the smallest subset of $\{1, \cdots, p_n\}$ such that $u_n = Ey_n = X_n(s_0)\beta(s_0)$, where $X_n(s_0)$ and $\beta(s_0)$ are respectively the design matrix and the coefficients corresponding to s_0. Let $v(s_0)$ be the number of components in s_0. We call s_0 the true submodel and denote $v(s_0)$ by K_0, and K is an upper bound for K_0. Let the projection matrix of $X_n(s)$ be $H_n(s) = X_n(s)\left[X_n^{\mathrm{T}}(s)X_n(s)\right]^{-1}X_n^{\mathrm{T}}(s)$. Define $\Delta_n(s) = \left\|u_n - H_n(s)u_n\right\|^2$.

The family of our new EBIC under model (12) is defined as:

$$BIC_a(s) = n\ln\left(\frac{y_n^{\mathrm{T}}\left(I_n - H_n(s)\right)y_n}{n}\right) + v(s)\ln n + 2\ln\left(\tau\left(S_j\right) + a\right), s \in S_j, a > 0. \tag{13}$$

Under the asymptotic identifiability condition proposed by [10], the consistency of our new EBIC is established. The asymptotic identifiability condition is as follows:

Condition 1: asymptotic identifiability. Model (12) with true submodel s_0 is asymptotically identifiable if:

$$\lim_{n\to\infty}\min\left\{(\ln n)^{-1}\Delta_n(s) : s \neq s_0, v(s) \leq K_0\right\} = \infty.$$

And other two lemmas proposed by [13] are also useful to our proof.

Lemma 1: if $\frac{\ln j}{\ln p} \to \delta$ as $p \to +\infty$, we have:

$$\ln\left(\frac{p!}{j!(p-j)!}\right) = j\ln p(1-\delta)\left(1 + o(1)\right).$$

Lemma 2: let χ_k^2 denote a χ^2 random variable with degrees of freedom k. If $m \to +\infty$ and $\frac{K}{m} \to 0$, then:

$$P\left(\chi_k^2 \geq m\right) = \frac{1}{\Gamma(k/2)}(m/2)^{(k/2)-1}\mathrm{e}^{-m/2}\left(1 + o(1)\right),$$

uniformly for all $k \leq K$.

We now state the consistency result as follows.

Theorem 2. Assume that $p_n = O\left(n^k\right)$ for some constant k. If $a > 0$, then under the asymptotic identifiability condition we have:

$$pr\left[\min\left\{BIC_a(s) : v(s) = j, s \neq s_0\right\} > BIC_a(s_0)\right] \to 1,$$

for $j = 1, \cdots, K$, as $n \to +\infty$.

Proof:

$$\begin{aligned}
&BIC_a(s) - BIC_a(s_0)\\
&= n\ln\left(\frac{y_n^{\mathrm{T}}\left(I_n - H_n(s)\right)y_n}{y_n^{\mathrm{T}}\left(I_n - H_n(s_0)\right)y_n}\right) + \left(v(s) - v(s_0)\right)\ln n, \text{ say,}\\
&\quad + 2\left[\ln\left(\tau\left(S_j\right) + a\right) - \ln\left(\tau\left(S_{K_0}\right) + a\right)\right]\\
&\triangleq T_1 + T_2
\end{aligned}$$

where:

$$T_1 = n\ln\left(\frac{y_n^{\mathrm{T}}\left(I_n - H_n(s)\right)y_n}{y_n^{\mathrm{T}}\left(I_n - H_n(s_0)\right)y_n}\right)$$
$$= n\ln\left(1 + \frac{y_n^{\mathrm{T}}\left(I_n - H_n(s)\right)y_n - e_n^{\mathrm{T}}\left(I_n - H_n(s_0)\right)e_n}{e_n^{\mathrm{T}}\left(I_n - H_n(s_0)\right)e_n}\right),$$
$$T_2 = \left(v(s) - v(s_0)\right)\ln n + 2\left[\ln\left(\tau(S_j) + a\right) - \ln\left(\tau(S_{K_0}) + a\right)\right].$$

Without loss of generality, we assume that $\sigma^2 = 1$.

Case 1: $s_0 \not\subset s$,

First, we will show $T_1 \geq n\ln\left(1 + \frac{C\ln n}{n}\right)$. We can write:

$$y_n^{\mathrm{T}}\left(I_n - H_n(s_0)\right)y_n = e_n^{\mathrm{T}}\left(I_n - H_n(s_0)\right)e_n = \sum_{i=1}^{n-v(s_0)} Z_j^2 = n\left\{1 + o_p(1)\right\},$$

where Z_j's are i.i.d. standard normal variables, we have:

$$y_n^{\mathrm{T}}\left(I_n - H_n(s)\right)y_n - e_n^{\mathrm{T}}\left(I_n - H_n(s_0)\right)e_n$$
$$= u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\}u_n + 2u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\}e_n - e_n^{\mathrm{T}}H_n(s)e_n + e_n^{\mathrm{T}}H_n(s_0)e_n \qquad (14)$$

By asymptotic identifiability condition, uniformly over s such that $v(s) \leq K$, we have:

$$(\ln n)^{-1} u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\}u_n \to \infty. \qquad \text{(I)}$$

Write

$$u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\}e_n = \sqrt{\left[u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\}u_n\right]}Z(s),$$

where:

$$Z(s) = \frac{u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\}e_n}{\sqrt{u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\}u_n}} \sim N(0,1).$$

We hence arrive at

$$\begin{aligned}
&\max\left[u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\}e_n : s \in S_j\right] \\
&\leq \sqrt{\left[u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\}u_n\right]}\max\left\{Z(s) : s \in S_j\right\} \\
&\leq \sqrt{\left[u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\}u_n\right]}O_p\left\{\sqrt{2\ln p_n}\right\} \\
&= o_p\left[u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\}u_n\right]
\end{aligned} \qquad \text{(II)}$$

where the last inequality follows the Bonferroni inequality, in detail we have:

$$P\left(\max\left\{Z(s) : s \in S_j, j \leq K\right\} \geq \sqrt{m}\right)$$
$$\leq \sum_{j=1}^{K}\tau(S_j)P\left(Z(s) \geq \sqrt{m}\right) = \sum_{j=1}^{K}\tau(S_j)P\left(\chi_1^2 \geq m\right)$$
$$\leq \sum_{j=1}^{K}\tau(S_j)P\left(\chi_j^2 \geq m\right)$$

where $m = 2K \ln p_n$. The last inequality comes from Lemma 2 that $P\left(\chi_1^2 \geq m\right) < P\left(\chi_j^2 \geq m\right)$. And according to Lemma 2, we can prove that $\sum_{j=1}^{K} \tau\left(S_j\right) P\left(\chi_j^2 \geq m\right) \to 0$.

Thus, we have

$$\begin{aligned} &\max\left\{e_n^{\mathrm{T}} H_n(s) e_n : v(s) \leq K\right\} = m\left(1 + o_p(1)\right) \\ &= 2K \ln p_n \left(1 + o_p(1)\right) = O_p\left(\ln p_n\right) = O_p\left(\ln n\right) \end{aligned} \tag{III}$$

We know the term $e_n^{\mathrm{T}} H_n\left(s_0\right) e_n$ is a χ^2-distributed statistic with a fixed degrees of freedom K_0. Then take (I), (II), (III) into (14) we have:

$$\begin{aligned} &y_n^{\mathrm{T}}\left(I_n - H_n(s)\right) y_n - e_n^{\mathrm{T}}\left(I_n - H_n\left(s_0\right)\right) e_n \\ &= u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\} u_n \left[1 + \frac{2u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\} e_n}{u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\} u_n} - \frac{e_n^{\mathrm{T}} H_n(s) e_n}{u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\} u_n}\right. \\ &\quad \left. + \frac{e_n^{\mathrm{T}} H_n\left(s_0\right) e_n}{u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\} u_n}\right] \\ &= u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\} u_n \left(1 + o_p(1)\right). \end{aligned}$$

Under the asymptotic identifiability condition, we know that $u_n^{\mathrm{T}}\left\{I_n - H_n(s)\right\} u_n$, which goes to infinity faster than $\ln n$, is the dominating term in $y_n^{\mathrm{T}}\left(I_n - H_n(s)\right) y_n - e_n^{\mathrm{T}}\left(I_n - H_n\left(s_0\right)\right) e_n$. Thus,

$$\frac{y_n^{\mathrm{T}}\left(I_n - H_n(s)\right) y_n - e_n^{\mathrm{T}}\left(I_n - H_n\left(s_0\right)\right) e_n}{e_n^{\mathrm{T}}\left(I_n - H_n\left(s_0\right)\right) e_n} \geq \frac{C \ln n}{n},$$

for any large constant C in probability, thus we have:

$$T_1 = n \ln\left(\frac{y_n^{\mathrm{T}}\left(I_n - H_n(s)\right) y_n}{y_n^{\mathrm{T}}\left(I_n - H_n\left(s_0\right)\right) y_n}\right) \geq n \ln\left(1 + \frac{C \ln n}{n}\right).$$

Next, we will show $BIC_a(s) - BIC_a\left(s_0\right) \to +\infty$, as $n \to +\infty$.

We know that:

$$\begin{aligned} &BIC_a(s) - BIC_a\left(s_0\right) = T_1 + T_2 \\ &\geq n \ln\left(1 + \frac{C \ln n}{n}\right) + \left(v(s) - v\left(s_0\right)\right) \ln n \\ &\quad + 2\left[\ln\left(\tau\left(S_j\right) + a\right) - \ln\left(\tau\left(S_{K_0}\right) + a\right)\right] \end{aligned} \tag{15}$$

On the one hand,

$$\begin{aligned} &n \ln\left(1 + \frac{C \ln n}{n}\right) + \left(v(s) - v\left(s_0\right)\right) \ln n \\ &\geq n \ln\left(1 + \frac{C \ln n}{n}\right) - v\left(s_0\right) \ln n \\ &= \ln n \frac{n}{\ln n}\left[\ln\left(1 + \frac{C \ln n}{n}\right) - v\left(s_0\right) - v\left(s_0\right) \frac{\ln n}{n}\right] \end{aligned}$$

Let $\frac{n}{\ln n} = a_n$, we know that $\frac{1}{a_n} \to 0$ as $n \to +\infty$, so that $\ln\left(1 + \frac{1}{a_n}\right) \sim \frac{1}{a_n}$.

Thus, the above equation can be expressed as

$$\begin{aligned}&= a_n(\ln n)\left[\ln\left(1+C\frac{1}{a_n}\right)-K_0\frac{1}{a_n}\right]\\&\approx a_n(\ln n)\left[C\frac{1}{a_n}-K_0\frac{1}{a_n}\right]\\&= \ln n(C-K_0) > \ln n(C-K)\end{aligned}$$

Which means:

$$n\ln\left(1+\frac{C\ln n}{n}\right)+\left(v(s)-v(s_0)\right)\ln n \geq \ln n(C-K). \quad \text{(IV)}$$

On the other hand,

According to Lemma 1 we have $\ln\left(\tau(S_j)+a\right)\approx \ln\tau(S_j)\approx j\ln p_n$ as $n\to+\infty$. Thus, we obtain:

$$2\left[\ln\left(\tau(S_j)+a\right)-\ln\left(\tau(S_{K_0})+a\right)\right]\approx 2(j-K_0)\ln p_n > -2K\ln p_n. \quad \text{(V)}$$

Bring (IV), (V) into (15) we can see that choosing $C>K(2k+1)$, we obtain:

$$BIC_a(s)-BIC_a(s_0) > \ln n\left(C-K(2k+1)\right)\to+\infty.$$

Case 2: $s_0\subset s$,

First, we will show:

$$T_1 = n\ln\left(\frac{y_n^{\mathrm{T}}\left(I_n-H_n(s)\right)y_n}{y_n^{\mathrm{T}}\left(I_n-H_n(s_0)\right)y_n}\right)\geq -2j\ln p_n\left\{1+o_p(1)\right\}.$$

In this case we have $\left\{I_n-H_n(s)\right\}X_n(s_0)=0$, thus $y_n^{\mathrm{T}}\left(I_n-H_n(s)\right)y_n = e_n^{\mathrm{T}}\left(I_n-H_n(s)\right)e_n$.

And

$$\begin{aligned}&e_n^{\mathrm{T}}\left(I_n-H_n(s_0)\right)e_n - y_n^{\mathrm{T}}\left(I_n-H_n(s)\right)y_n\\&= e_n^{\mathrm{T}}\left(H_n(s)-H_n(s_0)\right)e_n = \sum_{i=1}^{j}Z_i^2(s).\end{aligned}$$

where $j=v(s)-v(s_0)$, $Z_i(s)$ are some independent standard normal random variables depending on s. Let $\hat{e}_n=\left\{I_n-H_n(s_0)\right\}e_n$, we obtain that:

$$\begin{aligned}&n\ln\left(\frac{y_n^{\mathrm{T}}\left(I_n-H_n(s_0)\right)y_n}{y_n^{\mathrm{T}}\left(I_n-H_n(s)\right)y_n}\right)\\&= n\ln\left(1+\frac{e_n^{\mathrm{T}}\left(I_n-H_n(s_0)\right)e_n - e_n^{\mathrm{T}}\left(I_n-H_n(s)\right)e_n}{e_n^{\mathrm{T}}\left(I_n-H_n(s)\right)e_n}\right)\\&= n\ln\left\{1+\frac{\sum_{i=1}^{j}Z_i^2(s)}{\hat{e}_n^{\mathrm{T}}\hat{e}_n-\sum_{i=1}^{j}Z_i^2(s)}\right\}\leq n\frac{\sum_{i=1}^{j}Z_i^2(s)}{\hat{e}_n^{\mathrm{T}}\hat{e}_n-\sum_{i=1}^{j}Z_i^2(s)}.\end{aligned}$$

As $n\to+\infty$, $n^{-1}\hat{e}_n^{\mathrm{T}}\hat{e}_n\to\sigma^2=1$. And in case 1 we have shown that:

$$\max\left\{\sum_{i=1}^{j}Z_i^2(s), s\in S_{v(s)}\right\} = 2j\ln p_n\left\{1+o_p(1)\right\}.$$

Thus,

$$\max\left\{\frac{n\sum_{i=1}^{j}Z_i^2(s)}{\hat{e}_n^{\mathrm{T}}\hat{e}_n-\sum_{i=1}^{j}Z_i^2(s)}:s\in S_{v(s)}\right\}$$
$$\le\frac{2nj\ln p_n\{1+o_p(1)\}}{(n-2j\ln p_n)\{1+o_p(1)\}}=2j\ln p_n\{1+o_p(1)\}.$$

So, we can see that:

$$T_1=-n\ln\left(\frac{y_n^{\mathrm{T}}(I_n-H_n(s_0))y_n}{y_n^{\mathrm{T}}(I_n-H_n(s))y_n}\right)\ge-2j\ln p_n\{1+o_p(1)\}.$$

Consequently, uniformly in s such that $v(s)=j+v(s_0)$, we have:

$$\begin{aligned}&BIC_a(s)-BIC_a(s_0)=T_1+T_2\\&\ge-2j\ln p_n\{1+o_p(1)\}+(v(s)-v(s_0))\ln n\\&\quad+2\left[\ln(\tau(S_j)+a)-\ln(\tau(S_{K_0})+a)\right]\\&\approx-2j\ln p_n\{1+o_p(1)\}+(v(s)-v(s_0))\ln n+2(v(s)-v(s_0))\ln p_n\\&=-2j\ln p_n\{1+o_p(1)\}+j\ln n+2j\ln p_n=j\ln n.\end{aligned}$$

Thus, we have:

$BIC_a(s)-BIC_a(s_0)\ge j\ln n\to+\infty$ as $n\to+\infty$. The conclusion hence follows.

□

4. The Structure of Our Algorithm

Initial step:

Standardize $y,x_j,j=1,\cdots,p$, such that $y^{\mathrm{T}}1=0$, $x_j^{\mathrm{T}}1=0$, $y^{\mathrm{T}}y=n$, $x_j^{\mathrm{T}}x_j=n$, let $\lambda_1=2\max_{j\in\{1,2,\cdots,p\}}\left|x_j^{\mathrm{T}}y\right|$, consider the following penalized sum of squares:

$$\min_{\beta}\left\{\|y-X\beta\|_2^2+\lambda_1\|\beta\|_1\right\},$$

the features with nonzero estimated coefficients are selected and the set of their indices is denoted by s_{TEMP}, let $s_{*1}=s_{TEMP}$. Compute $I-H(s_{*1})$ and $\mathrm{NEBIC}(s_{*1})$.

General step:

For $k\ge1$, compute $\tilde{x}_j\tilde{y}$ for $j\in s_{*k}^{\complement}$, where $\tilde{y}=\left[I-H(s_{*k})\right]y$, $\tilde{x}_j=\left[I-H(s_{*k})\right]x_j$. Let $\lambda_{k+1}=2\max_{j\in s_{*k}^{\complement}}\left|\tilde{x}_j\tilde{y}\right|$, consider the following penalized sum of squares:

$$\min_{\beta}\left\{\left\|\tilde{y}-\tilde{X}\tilde{\beta}\right\|_2^2+\lambda_{k+1}\left\|\tilde{\beta}\right\|_1\right\},$$

the features with nonzero estimated coefficients are selected and the set of their indices is denoted by s_{TEMP}, let $s_{*k+1}=s_{*k}\cup s_{TEMP}$. Compute $\mathrm{NEBIC}(s_{*k+1})$, if $\mathrm{NEBIC}(s_{*k+1})>\mathrm{NEBIC}(s_{*k})$, stop; otherwise, compute $I-H(s_{*k+1})$ and continue.

Suppose that the above procedure stops at k^*+1 step, then our algorithm outputs the index set of selected features which is denoted by s_{*k^*}. Then the parameters in the selected model are estimated by their least-square estimates:

$$\hat{\beta}\left(s_{*k^*}\right)=\left[X^{\mathrm{T}}\left(s_{*k^*}\right)X\left(s_{*k^*}\right)\right]^{-1}X^{\mathrm{T}}\left(s_{*k^*}\right)y,\hat{\beta}\left(s^{\complement}_{*k^*}\right)=0.$$

The NEBIC for s_{*k}, $k=1,2,\cdots,k^*$ in the above algorithm is given by:

$$\mathrm{NEBIC}\left(s_{*k}\right)=n\ln\left(\frac{y^{\mathrm{T}}\left(I_n-H_n\left(s_{*k}\right)\right)y}{n}\right)+v\left(s_{*k}\right)\ln n$$
$$+2\ln\left(\tau\left(S_{v(s*k)}\right)+a\right)\quad\text{where } a>0.$$

5. Application to High-Dimensional Classification

5.1. Method

As we have discussed before, it is necessary to give better methods to estimate the unknown u_k's and Ω for the small-n-large-p problem. In this paper, we apply our algorithm to identify the non-zero entries of precision matrix, and the estimate of Ω is then obtained by the constrained maximum likelihood estimation. And we adopt the method proposed by [7] to estimate class means. The estimated class means and precision matrix are finally used to construct the discrimination rule.

5.2. Procedure

1) Constrained estimation for the class means

The predictor components are first ordered according to the F-statistic for testing the significance of class effects, instead of being pairwise fused, the class means for each component are then clustered by a divisive hierarchical clustering procedure. The class means are estimated under the structure revealed by the clustering. For details, we refer the reader to [7].

2) Constrained estimation for the precision matrix

The identification of non-zero entries in a concentration matrix has attracted considerable attention of the researchers, concentration matrix is the inverse of the covariance matrix of a random vector. A concentration matrix is closely related to an undirected graphical model. An undirected graphical model is specified by a vertex set V and an edge set E, and is denoted by $G=(V,E)$. The vertex set V represents a collection of random variables $\{Y_1,\cdots,Y_p\}$. The edge set E describes the inter-relationship among the random variables: there is an edge connecting vertices Y_i and Y_j if they are dependent conditioning on all the remaining variables. Suppose that $Y=(Y_1,\cdots,Y_p)$ follows a multivariate normal distribution with concentration matrix $\Omega=(\Omega_{ij})$. Then, there is an edge between Y_i and Y_j if and only if $\Omega_{ij}=\Omega_{ji}\neq 0$. Thus, the detection of edges of G is equivalent to the identification of non-zero entries of Ω.

In this paper, we adopt the method, which was proposed by [14], is based on

the relationship between entries of Ω and the coefficients of p regression models where each component of Y is regressed on the remaining $p-1$ components. A non-zero entry of Ω corresponds to a non-zero regression coefficient in the regression models. In other word, the detection and estimation of non-zero entries of Ω are then boiled down to the selection and estimation of non-zero coefficients in p regression models. According to this, we first apply our algorithm to identify non-zero entries of Ω, and the estimate of Ω is then obtained by the constrained maximum likelihood estimation. The details as follow.

For an undirected graph $G=(V,E)$, let V be modeled as the set of the components of a random vector $Y=(Y_1,\cdots,Y_p)$. We assume that Y follows a multivariate normal distribution $N(u,\Sigma)$. Without loss of generality, assume that $u=0$. Let Y_{i-} be the vector obtained from Y by eliminating component Y_i. Denote by Σ_{i-i-} the variance-covariance matrix of Y_{i-}, by σ_i^2 the variance of Y_i, and by Σ_{ii-} the covariance vector between Y_i and Y_{i-}. We know that the conditional distribution of Y_i given Y_{i-} is still normal, with the following conditional mean and conditional variance:

$$E(Y_i \mid Y_{i-}) = \Sigma_{ii-}\Sigma_{i-i-}^{-1}Y_{i-}, Var(Y_i \mid Y_{i-}) = \sigma_i^2 - \Sigma_{ii-}\Sigma_{i-i-}^{-1}\Sigma_{i-i}.$$

Let β_{ij} be the jth component of $\Sigma_{ii-}\Sigma_{i-i-}^{-1}$. We can then express the conditional distributions in the form of the following regression models:

$$Y_i = \sum_{j\neq i}\beta_{ij}Y_j + \epsilon_i, \epsilon_i \sim N(0,D_i), i=1,\cdots,p, \tag{16}$$

where $D_i = \sigma_i^2 - \Sigma_{ii-}(\Sigma_{i-i-})^{-1}\Sigma_{i-i}$. Without loss of generality, suppose that Y_i is the first component of Y. Let the covariance matrix Σ be partitioned as:

$\Sigma = \begin{pmatrix} \sigma_i^2 & \Sigma_{ii-} \\ \Sigma_{i-i} & \Sigma_{i-i-} \end{pmatrix}$, let $\Omega = (\omega_{jl})_{j,l\in\{1,\cdots,p\}}$, we can obtain:

$$\Omega = \Sigma^{-1} = \begin{pmatrix} D_i^{-1} & -D_i^{-1}\Sigma_{ii-}\Sigma_{i-i-}^{-1} \\ -\Sigma_{i-i-}^{-1}\Sigma_{i-i}D_i^{-1} & \Sigma_{i-i-}^{-1} + \Sigma_{i-i-}^{-1}\Sigma_{i-i}D_i^{-1}\Sigma_{ii-}\Sigma_{i-i-}^{-1} \end{pmatrix}.$$

By comparing the left upper block of Ω with $\Sigma_{ii-}\Sigma_{i-i-}^{-1}$, we see that:

$$\beta_{ij} = -\frac{\omega_{ij}}{D_i^{-1}} = -\frac{\omega_{ij}}{\omega_{ii}},$$

noting that $D_i^{-1} = \Omega_{ii}$. These connections establish the equivalence:

$$\omega_{ij} = 0 \Leftrightarrow \beta_{ij} = 0.$$

As we can see that the identification of non-zero entries of Ω reduces to the identification of the non-zero β_{ij} in the above regression models. Thus, firstly, we apply our algorithm to model (16) to identify all the non-zero entries of Ω. Let $\mathcal{E} = \{(j,l): \omega_{jl} \neq 0\}$, according to our algorithm we have $\hat{\mathcal{E}}$.

Next, we summary our proposed approaches. First, we introduce some notation. Denote by y_i the standardized n-vector of observed Y_i, let X_i be the $n\times(p-1)$ matrix consisting of all y_j with $j\neq i$. Here is our algorithm based on above data sets:

Algorithm for identifying $\mathcal{E}$

Initial step:

Let $\lambda_i = 2\max\left|X_i^{\mathrm{T}} y_i\right|$, let β_i be $(p-1)$-vector, for $i = 1,2,\cdots,p$. Consider minimizing following p penalized sum of squares separately:

$$\min_{\beta_i} \frac{1}{2}\left\|y_i - X_i\beta_i\right\|_2^2 + \lambda_i\left\|\beta_i\right\|_1, i = 1,2,\cdots,p,$$

Let $s_{1*1},\cdots,s_{p*1}$ denote the indices of features with nonzero estimated coefficients for the above p regression models respectively. Let $s_{*1} = \left\{s_{1*1},\cdots,s_{p*1}\right\}$. Then compute $\mathrm{NEBIC}\left(s_{*1}\right)$ and $I - H\left(s_{1*1}\right)$ for $i = 1,2,\cdots,p$.

General step $k+1$ ($k \geq 1$):

Let $\lambda_i = 2\max\left|X_i^{\mathrm{T}}\left(I - H\left(s_{i*k}\right)\right)y_i\right|$ for $i = 1,2,\cdots,p$. Consider minimizing following p penalized sum of squares separately:

$$\min_{\tilde{\beta}_i} \frac{1}{2}\left\|\tilde{y}_i - \tilde{X}_i\tilde{\beta}_i\right\|_2^2 + \lambda_i\left\|\tilde{\beta}_i\right\|_1, i = 1,2,\cdots,p.$$

where $\tilde{y}_i = \left(I - H\left(s_{i*k}\right)\right)y_i$, $\tilde{X}_i = \left(I - H\left(s_{i*k}\right)\right)X_i$, $\tilde{\beta}_i = \beta_i\left(s_{i*k}^{\complement}\right)$. Let s_{iTEMP} be the indices of features with nonzero estimated coefficients for the above ith regression model, and $s_{i*k+1} = s_{iTEMP} \bigcup s_{i*k}$ for $i = 1,2,\cdots,p$. Then let $s_{*k+1} = \left\{s_{1*k+1},\cdots,s_{p*k+1}\right\}$. Compute $\mathrm{NEBIC}\left(s_{*k+1}\right)$. If $\mathrm{NEBIC}\left(s_{*k+1}\right) > \mathrm{NEBIC}\left(s_{*k}\right)$, stop; otherwise, compute $I - H\left(s_{i*k+1}\right)$ for $i = 1,2,\cdots,p$ and continue.

Suppose the above algorithm stops at step $k^* + 1$, then our algorithm output the index set $s_{*k^*} = \left\{s_{1*k^*},\cdots,s_{p*k^*}\right\}$.

The NEBIC for set s_{*k} is given by the following formula:

$$\mathrm{NEBIC}\left(s_{*k}\right) = \sum_{i=1}^{p}\left\{n\ln\left(\left\|\left[I - H\left(s_{i*k}\right)\right]y_i\right\|_2^2\right) + \left|s_{i*k}\right|\ln n + 2\ln\left(\tau\left(S_{\left|s_{i*k}\right|}\right) + a\right)\right\}.$$

Estimation of precision matrix:

From the above algorithm we identify all the non-zero entries of $\hat{\Omega}$, let $\hat{\mathcal{E}}$ be the set of all the indices. Then $\hat{\Omega}$ is obtained by the constrained maximum likelihood estimation as follows:

$$\hat{\Omega} = \min_{\omega_{jl}=0:(j,l)\in\mathcal{E}^{\complement}}\left\{\mathrm{trace}\left(\hat{S}\Omega\right) - \ln\det\left(\Omega\right) + \lambda\left\|\Omega\right\|_1\right\},$$

where $\hat{S}$ is the empirical covariance matrix.

3) Discrimination rule

The discrimination rule is constructed by replacing the unknowns in $d_{kj}(x)$ by their estimates obtained above. Here $d_{kj}(x)$ is:

$$d_{kj}(x) = \left[x - \left(u_k + u_j\right)/2\right]^{\mathrm{T}}\Omega\left(u_k - u_j\right) + \ln\left(\pi_k/\pi_j\right).$$

Then we have $\hat{d}_{kj}(x)$:

$$\hat{d}_{kj}(x) = \left[x - \left(\hat{u}^k + \hat{u}^j\right)/2\right]^{\mathrm{T}}\hat{\Omega}\left(\hat{u}^k - \hat{u}^j\right) + \ln\left(n_k/n_j\right).$$

Classify x into class k if and only if:

$$\hat{d}_{kj}(x) > 0, \text{ for all } j \neq k.$$

5.3. Simulation Studies

We compare our method with other methods available in the literature through numerical studies in this section. The methods considered for the comparison are: 1) linear discrimination with detected sparsity (LDwDS) in [7]. 2) The adaptive hierarchically penalized nearest shrunken centroid (ahp-NSC) and the adaptive L_∞ -norm penalized NSC (alp-NSC) proposed in [15]. 3) Pairwise SIS (PSIS) in [16].

The performance of the methods is evaluated by the misclassification rate (MCR) and computation time. We have three simulation settings, repeat each setting 100 times. At each replicate under a simulation setting, we simulate a training data set and a testing data set. The training data set is used to construct discrimination rules with the methods, and the testing set is used to evaluate the MCR and record computation time.

We design the simulation setting inspired by [7], consider $K=2$, $p=400$. The following single scheme for the class means is taken throughout:

$$u_1=0, u_2=\left(\underbrace{0.5,\cdots,0.5}_{100},\underbrace{0,\cdots,0}_{p-100}\right).$$

The covariance matrix is generated through the precision matrix. First, the nonzero positions of the precision matrix are decided in the following schemes:

ES1: the non-zero entries of Ω is randomly determined with:

$$\Pr\left(\omega_{ij}=0\right)=1-\Pr\left(\omega_{ij}\neq 0\right)=0.99.$$

ES2: Ω is a diagonal block matrix with 10 blocks of size 10.

ES3: Ω is a diagonal block matrix with 10 blocks of size 10, each block is a diagonal band matrix with $\omega_{ij}\neq 0$ if $|i-j|\leq 2$.

For the nonzero values of the precision matrix, we first generate $\tilde{\Omega}=\tilde{\omega}_{ij}$ as follows: $\tilde{\omega}_{ii}=1$, $\tilde{\omega}_{ij}=\tilde{\omega}_{ji}$ are generated as i.i.d. observations from the uniform distribution on $[-0.3,0.7]$. Then we take $\Omega=\tilde{\Omega}+\left(0.1-\lambda_{\min}\left(\tilde{\Omega}\right)\right)I$. Eventually, take the common covariance matrix as the correlation matrix corresponding to Ω^{-1}. The training sample size $n_1=200$ and the testing sample size $n_2=1000$. The covariance matrix is generated as a diagonal block matrix with four 100 × 100 identical blocks. The 100 × 100 block is generated by the same generating schemes above. The simulation results under these three settings are reported in **Tables 1-3**.

From all three tables we can see that as for the MCR, the performance of our method is much better than the two NSC methods and PSIS. Although our method has the same MCR with LDwDS from **Table 1** and **Table 2**, it takes shorter computation time than LDwDS. From **Table 3** we find that our method performs better than LDwDS both in misclassification and computation time. From [7] we notice that the misclassification rates of LDwDS are universally lower than all the other methods under consideration. Nevertheless, it is not without drawbacks. LDwDS takes longer computation time than other methods.

Table 1. Averaged misclassification rate (MCR) and computation time (in seconds) under simulation ES1 (numbers in the parentheses are standard deviations).

Method	MCR	Time
LDwDS	0.005 (0.005)	14.339 (1.502)
ahp-NSC	0.045 (0.026)	0.213 (0.049)
alp-NSC	0.193 (0.050)	0.202 (0.026)
PSIS	0.500 (0.006)	0.115 (0.009)
our method	**0.005 (0.005)**	**11.008 (0.962)**

Table 2. Averaged misclassification rate (MCR) and computation time (in seconds) under simulation ES2 (numbers in the parentheses are standard deviations).

Method	MCR	Time
LDwDS	0.001 (0.003)	28.532 (3.524)
ahp-NSC	0.147 (0.043)	0.205 (0.027)
alp-NSC	0.291 (0.043)	0.199 (0.024)
PSIS	0.500 (0.006)	0.114 (0.007)
our method	**0.001 (0.002)**	**10.684 (1.199)**

Table 3. Averaged misclassification rate (MCR) and computation time (in seconds) under simulation ES3 (numbers in the parentheses are standard deviations).

Method	MCR	Time
LDwDS	0.017 (0.025)	20.852 (2.384)
ahp-NSC	0.167 (0.049)	0.194 (0.007)
alp-NSC	0.310 (0.041)	0.190 (0.009)
PSIS	0.500 (0.006)	0.110 (0.004)
our method	**0.016 (0.023)**	**8.725 (0.709)**

Therefore, from the perspectives of computation time and misclassification, our algorithm performances better than its competing algorithms.

6. Summary and Discussion

In this paper, on the one hand, based on the characteristic of lasso regression we propose a novel method to choose the regularization parameter λ, which gives the exact maximum value of λ for lasso regression. On the other hand, since the ordinary Bayes information criterion is too liberal for model selection when the model space is large, inspired by [10] we propose a new EBIC (NEBIC). Compared with EBIC the parameter in our new criterion doesn't need to be restricted in a small range, what's more, under some mild conditions our new EBIC is also shown to be consistent. Then, based on these two findings, we have a new algorithm for feature selection. In detail, we apply our new method to choose regula-

rization parameter for sequential lasso, and our NEBIC is used as the stopping rule. After that, we apply our algorithm to classification problem, the numerical study demonstrates that our algorithm performs better than its competing methods under consideration.

Further research may consider these two questions:

Firstly, our method chooses the largest regularization parameter λ such that at least one of the $\beta_j's$, $j \in \{1,2,\cdots,p\}$, will be estimated nonzero. A nature question is that instead of solving the optimization problems can we obtain the indices of those features with nonzero estimated coefficients by easier methods? Luo and Chen [17] pointed out that these indices are related to the choice of regularization parameter λ under some conditions. However, these conditions are too strict. Further research may consider if there exist milder conditions to connect our new method for choosing regularization parameter with the indices of non-zero estimated coefficients.

Secondly, we choose the upper convex function $\frac{x}{x+a}, a>0$ as the prior function over the model space to extend EBIC, and our new EBIC is shown to be consistent. Subsequent research may consider if there exists a special function class for model space priors such that within this class the Bayes information criteria is consistent.

Conflicts of Interest

The authors declare no conflicts if interest regarding the publication of this paper.

References

[1] Desboulets, L.D.D. (2018) A Review on Variable Selection in Regression Analysis. *Econometrics*, **6**, Article 45. https://doi.org/10.3390/econometrics6040045

[2] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[3] Zou, H. (2012) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. https://doi.org/10.1198/016214506000000735

[4] Zhang, C. (2010) Nearly Unbiased Variable Selection under Minimax Concave Penalty. *The Annals of Statistics*, **38**, 894-942. https://doi.org/10.1214/09-AOS729

[5] Tropp, J.A. (2004) Greed Is Good: Algorithmic Results for Sparse Approximation. *IEEE Transactions on Information Theory*, **50**, 2231-2242. https://doi.org/10.1109/TIT.2004.834793

[6] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least Angle Regression. *The Annals of Statistics*, **32**, 407-499. https://doi.org/10.1214/009053604000000067

[7] Luo, S. and Chen, Z. (2020) A Procedure of Linear Discrimination Analysis with Detected Sparsity Structure for High-Dimensional Multi-Class Classification. *Journal of Multivariate Analysis*, **179**, Article ID: 104641. https://doi.org/10.1016/j.jmva.2020.104641

[8] Luo, S. and Chen, Z. (2014) Edge Detection in Sparse Gaussian Graphical Models.

Computational Statistics & Data Analysis, **70**, 138-152.
https://doi.org/10.1016/j.csda.2013.09.002

[9] Chen, Z. and Jiang, Y. (2020) A Two-Stage Sequential Conditional Selection Approach to Sparse High-Dimensional Multivariate Regression Models. *Annals of the Institute of Statistical Mathematics*, **72**, 65-90.
https://doi.org/10.1007/s10463-018-0686-5

[10] Chen, J. and Chen, Z. (2008) Extended Bayesian Information Criteria for Model Selection with Large Model Spaces. *Biometrika*, **95**, 759-771.
https://doi.org/10.1093/biomet/asn034

[11] Bickel, P.J. and Levina, E. (2004) Some Theory for Fisher's Linear Discriminant Function, 'Naive Bayes', and Some Alternatives When There Are Many More Variables than Observations. *Bernoulli*, **10**, 989-1010.
https://doi.org/10.3150/bj/1106314847

[12] Fan, J. and Fan, Y. (2008) High Dimensional Classification Using Features Annealed Independence Rules. *Annals of Statistics*, **36**, 2605-2637.
https://doi.org/10.1214/07-AOS504

[13] Luo, S. and Chen, Z. (2013) Extended BIC for Linear Regression Models with Diverging Number of Relevant Features and High or Ultra-High Feature Spaces. *Journal of Statistical Planning and Inference*, **143**, 494-504.
https://doi.org/10.1016/j.jspi.2012.08.015

[14] Meinshausen, N. and Bühlmann, P. (2006) High-Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, **34**, 1436-1462.
https://doi.org/10.1214/009053606000000281

[15] Wang, S. and Zhu, J. (2007) Improved Centroids Estimation for the Nearest Shrunken Centroid Classifier. *Bioinformatics*, **23**, 972-979.
https://doi.org/10.1093/bioinformatics/btm046

[16] Pan, R., Wang, H. and Li, R. (2016) Ultrahigh-Dimensional Multiclass Linear Discriminant Analysis by Pairwise Sure Independence Screening. *Journal of the American Statistical Association*, **111**, 169-179.
https://doi.org/10.1080/01621459.2014.998760

[17] Luo, S. and Chen, Z. (2014) Sequential Lasso Cum EBIC for Feature Selection with Ultra-High Dimensional Feature Space. *Journal of the American Statistical Association*, **109**, 1229-1240. https://doi.org/10.1080/01621459.2013.877275