Correlation search for screening and selecting nonlinear features

By

Yao Mingkang

Supervisor:

Professor Chen Zehua

ST5199 Honours Project in Statistics

Department of Statistics and Applied Probability

National University of Singapore

2018/2019

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor, professor Chen Zehua. He led me to a new field in this academic year and gave me a great number of valuable ideas during the accomplishment of the whole project. He was always patient to help me whenever I met difficulty in my project or researched in a wrong way. It is my honour and luck to be guided by him.

Besides, I would like to thank my friends in Singapore as well as China for their accompaniment and encouragement. I thank this department for providing this opportunity for me so that I can focus on a particular field and enhance my ability of programming.

Finally, I want to thank my whole family including my parents, grandparents. They support me, trust me and encourage me all the time.

Contents

Summary

Sure independence screening method and sequential sparse group search procedure provides us with a superior method to screen and select features among various variables. In this project, we will focus on a sparse high-dimensional nonlinear model and our purpose is to pick out all the important features and remain the irrelevant covariates as less as possible.

We aim to compare several measurements in the screening and selecting procedure. In the procedure, we introduce two major measurements: one is Distance Correlation that would be performed on single covariates and a covariates group and the other one is group Canonical Correlation.

It's our contribution to apply Distance Correlation and Canonical Correlation to screening and selecting stages and perform sparse group canonical correlation selection procedure in a nonlinear model. All above procedures will follow a stopping rule called Extended Bayesian Information Criterion in the selecting stage.

In the simulation step, we enlarge each covariate to a covariate group such as cubic form (x, x^2, x^3). We use R to generate the covariates along with responses, write three algorithms and show the results to compare the accuracy along with efficiency of the approaches under a variety of settings. Therefore we find the superior one or conclude the situations that the exact method fits.

Chapter 1

Introduction

Feature selection of high-dimensional models is quite popular nowadays in that it is widely applied in a large number of scientific fields such as DNA screening, medical research, Consumer behavior forecasting, financial analysis and so on. With the development of modern technology for data collection, researchers are able to collect ultrahigh-dimensional data at low cost and they find that in a great number of issues the data is nonlinear rather than ideal linear relationship. Hence the next problem that we concern about is how to analyze data.

Similarly to some real situations, the data we need to consider contains the characteristic that the number of covariates q is larger than the number of observations n but the number of important features that relate to the responses p is very small, which is called the small-n-large-q structure. The former regularization methods may not perform well for this type of data structure due to the simultaneous challenges of computational expediency, statistical accuracy, and algorithmic stability. Meanwhile some existing regularized regression approaches or sequential approaches can do with it but they all mainly focus on the linear condition. The method for nonlinear feature selection is not sufficient up to now.

So a new approach that combines dimensionality reduction method with sequential sparse group search approach is proposed. In this approach, there are two stages: (I) screening an acceptable number (<n) of features to reduce the dimension of the covariate matrix and (II) from the new model, selecting the p most important features relied on the n observations

with the help of Extended Bayesian Information Criterion (Jiahua Chen and Zehua Chen 2008). This algorithm allows the number of predictors to be greater than the sample size and we introduce two particular measurements, Distance Correlation (DC) or Canonical Correlation (CC), to describe the nonlinear relationship in the algorithm. Moreover, we generate the covariates and use a bunch of functions that are nonlinear to relate important covariates to response variable. We will demonstrate that the algorithm performs well for the high-dimensional data that we simulate in advance.

The remainder of the thesis is arranged as following. In chapter 2 we introduce some classic regularization methods, sequential approaches and dimensionality reduction method refers to current literatures. In chapter 3 we build the model and demonstrate our two-stage algorithm via DC and CC exactly. In chapter 4 we use R to do the simulation study and analyze the result. Eventually we will conclude our findings and discuss the aspects that can be improved in the future.

Chapter 2

Literature Review

There are several methods to select features from a high-dimensional model. The first category consists of methods using regularized regression approach with various penalty functions. It includes LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), Adaptive LASSO (Zou, 2006) so here we will introduce them in first section.

The second category consists of sequential approaches such as Forward regression (FR) (Hansheng Wang, 2009), L_2Boosting (Buehlmann et al., 2006), Orthogonal matching pursuit (OMP) (Cai and Wang, 2011), Sequential Lasso (Luo and Chen, 2014). We will discuss FR, L2Boosting, OMP, Sequential Lasso and Sequential canonical correlation search (SCCS) procedure (Luo and Chen, 2018) in second section.

Furthermore, in that we focus on the small-n-large-q data structure where q is huge, we need to screen some candidates from a large number of covariates before the selection procedure for accuracy along with efficiency, which is called dimensionality reduction method. So we will introduce the sure independence screening (SIS) (Fan and Lv, 2008) and DC-SIS (Li, Zhong, Zhu, 2012) in the last section.

2.1 Regularization Methods

2.1.1 LASSO

Robert Tibshirani proposed a new technique, called the lasso, for 'least absolute shrinkage and selection operator' in 1996 based on Leo Breiman's nonnegative garrote. It shrinks some coefficients and sets others to 0, hence it is able to keep the good features of both subset selection and ridge regression.

Consider the simplest univariate linear regression model. There is a sample of size N, each of which has p covariates and one dependent variable denoted by $X = \left(x_{i1}, x_{i2} \dots x_{ip}\right)$ and the response set $Y = (y_1 \dots y_i)$. The Lasso estimate, denoted by $\hat{\beta}_{lasso}(\lambda)$, is the solution to minimize

$$\sum_{i=1}^{n} \{Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\}^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Compare it with ridge regression, this penalty $\sum_{j=1}^{p} |\beta_j|$ can be called L_1-penalty, Thus, it is sometimes called L1-penalty method as well.

Another equivalent form is to solve

$$\sum_{i=1}^{n} \{Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\}^2, \text{ s.t. } \sum_{j=1}^{p} |\beta_j| < t, \text{ for a constant t}$$

Because of the nature of the constraint, letting t sufficiently small (or λ sufficiently large) will cause some of the coefficients to be exactly zero. There are zeros values in the estimated β, the actual assumption behind the method is the "sparsity". This will help us for feature (model) selection. The lasso does a kind of continuous subset selection. If t is chosen larger than $t_0 = \sum_{j=1}^{p} |\widehat{\beta_j}|$, where $\widehat{\beta_j}$ is the LSE, then the lasso estimates are the $\widehat{\beta_j}$. If t = 0, then no variable is selected. Let $s = \frac{t}{t_0}$, then when s = 0, no variable is selected, when s = 1 all are selected and the estimator is the LSE.

However, the drawbacks of Lasso should not be ignored. Lasso creates large covariance and does not always hold the oracle property (Fan and Li 2001). It only obtain it when the condition called irrepresentablility is satisfied, which is too strong to be satisfied in practical problems. So Fan and Li proposed a new penalty function called Smoothly Clipped Absolute Deviation (SCAD) penalty in 2001, which holds the oracle property.

2.1.2 Smoothly Clipped Absolute Deviation (SCAD)

The penalty function of SCAD is a continuous differentiable function defined by

$$p_\lambda(\beta) = \lambda\left\{I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta \le \lambda)\right\}$$

For some $a > 2$ and $\beta > 0$.

SCAD penalty corresponds to a quadratic spline function with knots at λ and $a\lambda$. Compared with Lasso, it keep the three properties of a good penalty. (I) Unbiasedness: The resulting estimator is nearly unbiased when the true unknown parameter is large. (II) Sparsity: The resulting estimator is singular at the origin, which automatically sets small estimated coefficients to zero to reduce model complexity. (III)Continuity: The final estimator is continuous to ensure the stability of the model selection.

The solution of it is given by Fan(1997):

$$\hat{\theta} = \begin{cases} sgn(z)(|z| - \lambda)_+ & when\ |z| \le 2\lambda \\ \dfrac{\{(a-1)z - sgn(z)a\lambda\}}{(a-2)} & when\ 2\lambda < |z| \le a\lambda \\ z & when\ |z| > a\lambda \end{cases}$$

In practice, we could search the best pair (λ, a) over the two-dimensional grids using some classic feature selection criteria, such as cross-validation and generalized cross-validation, which is expensive in computation. For the standard error of the estimates, Fan and Li developed a sandwich formula that is good enough for moderate sample sizes.

2.1.3 Adaptive LASSO

In order to make up the LASSO's drawback that it fails to hold the oracle property, a modified version of LASSO called adaptive LASSO was proposed by Hui Zhou (2006). It is also an l_1 penalization method. Suppose that $\hat{\beta}$ is a consistent estimator to β^*. The adaptive lasso estimates $\hat{\beta}^{*(n)}$ are defined as:

$$\hat{\beta}^{*(n)} = argmin||y - \sum_{j=1}^{p} x_j\beta_j||^2 + \lambda_n \sum_{j=1}^{p} \widehat{w_j}|\beta_j| \quad (*)$$

where $\gamma > 0$ is a constant and $\widehat{w_j}$ is the weight vector $\widehat{w} = 1/|\widehat{\beta_j}|^{\gamma}$. Every coefficient in lasso will be equally penalized but here adaptive lasso assign different weights to different coefficients. This paper shows that if the weights are data-dependent and cleverly chosen, then it can enjoy the oracle properties. Thus $(*)$ is a convex optimization problem and its global minimizer can be found by software.

The oracle methods including SCAD and adaptive lasso tend to be more accurate than lasso when the signal-to-noise ratio (SNR) is high. Adaptive lasso can combine the edge of lasso with that of SCAD. When it comes to a middle or low SNR, adaptive lasso performs better than SCAD and Garotte. Therefore the overall performance of the adaptive lasso is the best among lasso, SCAD and Garotte. However adaptive lasso still cannot be directly applied in a model where the interaction exists. In those models, the main-effect features are harder to be picked out in that the methods impose hierarchical structures through penalties and computation difficulty is also often faced up with. Therefore, we discuss some sequential methods in next section.

2.2 Sequential Approaches

2.2.1 Forward Regression (FR)

Forward regression starts with no covariates in the model. The algorithm tries out the covariates one by one and includes them if they can increase the predictability. At each step of the forward selection, feature selection criteria such as AIC and BIC are used to accept covariates that improve the model best to the new model. The steps repeat again and again and stop when all the covariates are checked or the new value of feature selection criteria is bigger than the former one. Hangsheng Wang (2009) shows that (both theoretically and numerically) FR can identify all relevant predictors consistently, even if the predictor dimension is considerably larger than the sample size.

Under the assumption that the true model ζ exists, the main objective is to discover all relevant predictors consistently. So we use the following FR algorithm.

- Set $S^{(0)} = \emptyset$
- Forward regression:
  - In the kth step $(k \geq 1)$, we are given $S^{(k-1)}$. Then for every $j \in \mathcal{F} \backslash S^{(k-1)}$, we construct a candidate model ${M_j}^{(k-1)} = S^{(k-1)} \cup \{j\}$. We then computer sum of residual square $RSS_j^{(k-1)} = Y^T\{I - H_j^{(k-1)}\}Y$, where $H_j^{(k-1)}$ is the Hessian Matrix $H_j^{(k-1)} = X_{{M_j}^{(k-1)}}\{{X_{{M_j}^{(k-1)}}}^T X_{{M_j}^{(k-1)}}\}^{-1}{X_{{M_j}^{(k-1)}}}^T$ and I is an identity matrix
  - Then we find

$$a_k = argmin\{RSS_j^{(k-1)}\}$$

    and update $S^{(k)} = S^{(k-1)} \cup \{a_k\}$

- Repeat the above step until the BIC stops being smaller. We then collect those models by a solution path $S = \{S^{(k)}: 1 \leq k \leq n\}$ with $S^{(k)} = \{a_1 \dots a_k\}$.

However, although the paper shows both theoretically and numerically that FR can discover all relevant predictors consistently even in a small-n-large-q data structure, we should not claim FR as the only good method for variable selection based on the simulation result. We will discuss some other sequential procedures below.

2.2.2 L_2Boosting

Another sequential procedure called L_2Boosting was developed by Peter Buhlmann in 2006. The paper proposes a computationally efficient approach for the tuning parameter in boosting and corrected AIC criterion is applied in this algorithm.

Suppose a regression model

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i, i = 1 \dots n,$$

Boosting using the squared error loss, L_2Boosting, has a simple structure. L_2Boosting algorithm is described below:

- (initialization). Apply the base procedure yielding the function estimate

  $$\hat{F}^{(1)}(\cdot) = \hat{g}(\cdot),$$

  where $\hat{g} = \hat{g}(X, Y)$ is estimated from original data. Set $m = 1$.

- Compute residuals $U_i = Y_i - \hat{F}^{(m)}(X_i)$, fit the real-valued base procedure to the current residuals. $\hat{g}^{(m+1)}(\cdot)$ is defined as an estimate based on the predictor variables and the current residuals.

  Update $\hat{F}^{(m+1)}(\cdot) = \hat{F}^{(m)}(\cdot) + \hat{g}^{(m+1)}(\cdot)$.

- (iteration). Increase the iteration index m by one and repeat step 2 until it reaches a stopping iteration.

For linear models, L_2Boosting has the useful properties of variable selection and shrinkage. Based on the simulation result, L_2Boosting is able to consistently recover high-dimensional sparse functions. However, this paper only shows the situation of simple linear model and in fact it dose not always stable in case of complex situations.

2.2.3 Orthogonal Matching Pursuit (OMP)

Orthogonal matching pursuit (OMP) is the canonical greedy algorithm for sparse approximation that selects the feature X_i that has the highest correlation coefficient with the current residual at each step. Letting ϕ denote a matrix of size $M * N$ (where typically $M < N$) and y denote a vector in R^M, the goal of OMP is to recover a coefficient vector with roughly $K < M$ nonzero terms so that $\phi_{\hat{x}}$ equals y exactly or approximately. OMP is frequently used to find sparse representations for signals $y \in R^M$ in settings where ϕ represents an overcomplete dictionary for the signal space.

The attractive feature of OMP is its simplicity. The entire algorithm is specified below, and it requires approximately the same number of lines of code to implement in a software package such as Matlab. Despite its simplicity, OMP is empirically competitive in terms of approximation performance.

OMP Algorithm:

- Input ϕ, y and the stopping criterion such as BIC
- Initialize $r^0 = y, x^0 = 0, \Lambda^0 = \emptyset, l = 0$
- While stopping criterion does not occur do
  - Match $h^l = \phi^T r^l$
  - Identify $\Lambda^{l+1} = \Lambda^l \cup \{argmax_j |h^l(j)|\}$
  - Check the stopping criterion
  - Update $x^{l+1} = argmin_{z:supp(z) \subseteq \Lambda^{l+1}} ||y - \phi_z||_2$

$$r^{l+1} = y - \phi x^{l+1}$$

$$l = l + 1$$

- End while
- Output $\hat{x} = x^l = argmin_{z:supp(z)\subseteq\Lambda^{l+1}} ||y - \phi_z||_2$

However, the conditions for OMP are quite strict and a complete investigation of OMP is still needed.

2.2.4 Sequential LASSO

Luo and Chen (2014) proposed a sequential method called Sequential Lasso. It is a slightly modified version of OMP. At each step of the procedure, it solves partially penalized least squares problems when the features selected in earlier steps are not penalized. This method uses Extended Bayesian Information Criterion (EBIC) as a stopping rule and the steps stop as long as the value of EBIC reaches the minimum. Firstly the Sequential Lasso selects all the relevant features before any irrelevant features can be selected, and that the EBIC decreases until it attains the minimum at the model consisting of exactly all the relevant features and then begins to increase. This method also holds the oracle property as well as consistency. The algorithm is described below:

- Initial Step: Standardize $y, x_j, j = 1 \dots q$. Find the maximum of $|{x_j}^T y|$ as a singleton S_{temp}. Let $S_{*1} = S_{temp}$ and consider S_{*1} as the active set. Compute $I - H\ (S_{*1})$ and $EBIC\ (S_{*1})$, where H is the Hessian matrix.
- General step: In kth step $(k \geq 1)$, find the maximum of $|{\tilde{x}_j}^T \tilde{y}|$ as S_{temp}, where $\tilde{y} = [I - H\ (S_{*k})]\ y$ and $\tilde{x}_j = [I - H\ (S_{*k})]\ \tilde{x}_j$. Let $S_{*k+1} = S_{temp} \cup S_{*k}$ as the active set. Compute $EBIC\ (S_{*k+1})$ and if $EBIC\ (S_{*k+1}) > EBIC\ (S_{*k})$, stop and go to next step; otherwise, compute $I - H\ (S_{*k+1})$ and repeat this step.
- Output step: the parameters in the selected set are estimated by their least square estimates (LSE).

2.2.5 Sequential Canonical Correlation Search Procedure(SCCS)

Luo and Chen introduce the sequential canonical correlation group search procedure and demonstrate its advantage in analysis of group structure effects. Similar sequential approaches also include sequential Lasso (Luo and Chen, 2014) and orthogonal matching pursuit (OMP) (Cai and Wang, 2011), which do not consider the group structure effect. The group structures in either the responses or the covariates contain the relationships between the variables within the groups. A statistic ignoring those correlation information fails to explain the truth. From the principle of sufficiency, by taking into account the group structures, the efficiency for detecting relevant features can be enhanced.

In the recent years, some other feature screening methods like Distance correlation screening method that I mentions later below and the group Lasso approach are proposed. However, those methods are mainly for feature screening instead of selection. Although feature screening and feature selection are similar in some characteristics, they are different to a large extent.

In principle, distance correlation can replace the role of the canonical correlation measure in SCCS. But the canonical correlation measure has certain edges over DC, which will be elaborated in the chapter four. The SCCS not only has an edge over the group Lasso approach of Li et al. (2015) but also has a great speed in computation, as our simulation studies show. The SCCS is not restricted by the dimension of the data while the group Lasso approach might not be implementable when the number of covariate groups is too large, which will be simulated and proved in the chapter four.

Similar to some other regularized regression approaches, SCCS aims to minimize the residual of sum square $Tr[(Y-XB)^T(Y-XB)]$ by imposing some constraints on B such as $C(B) \leq c$ for some constant c, $C(B)$ being a function of B, where Y is the response vector and X is the matrix of observed covariates. At each step, they start with a current estimated mean response $\hat{\mu}_A$, and select the next variable such that it has the largest correlation with the current residual $\hat{y} = y - \hat{\mu}_A$, then update the current estimate with the newly selected variable and proceed to the next step. Suppose there is only one variable that attains the largest correlation at each step. Then its nonzero coefficients are selected by EBIC (Chen and Chen, 2008). The EBIC also serves as the stopping rule for the procedure.

The selection consistency of the SCCS procedure is also shown by proving that

$$P\left(tr\left(\widehat{C_k}(\zeta)\right) = \max_{k=1\ldots p}\left\{tr\left(\widehat{C_k}(\zeta)\right)\right\}\right) \to 1$$

Where $\widehat{C_k}(\zeta)$ is the estimated canonical correlation matrix of X_k and the residual $\tilde{y}$ under model ζ.

The maximum number of the relevant features for each response variable p_0 is proved to diverge to infinity as the sample size n goes to infinity. This conclusion origins from the theorem: denoting by S_l^* the index set for the non-zero regression coefficients of the response variable y selected by the SCCS procedure, we have, as $n \to \infty$,

(1) For each $l \in C_0$,

$$P(S_l^* = S_{0l}) \to 1.$$

(2) For each $l \in C_0^C$,

$$P(S_l^* = \emptyset) \to 1.$$

With the above two properties, we can show that SCCS is a reliable approach of feature selection.

2.3 Dimensionality Reduction Method

2.3.1 Sure Independence Screening (SIS)

In a variable screening procedure, the probability of the survival of all the relevant covariates diverges to 1. Hence Fan and Lv (2008) introduce a simple sure screening method based on correlation learning. They call this correlation screening method SIS, since each feature is used independently as a predictor to decide how useful it is for predicting the response variable. The model we face on must be centered in advance, which means the observed mean is 0 and each standard deviation is 1. Suppose that

$M_* = \{1 \leq i \leq p: \beta_i \neq 0\}$ is the true sparse model and $s = |M_*|$ is the non-sparsity size. The left $p - s$ covariates can also be correlated with the response variable via linkage to the predictors that are contained in the model. Let $\omega = (\omega_1 \dots \omega_p)^T$ be a vector that is computed by componentwise regression $\omega = X^T y$ or Pearson correlation $\omega = cor\ (x_i, y)$.

where the $n * p$ data matrix X needs to be standardized in advance.

Therefore, we obtain a vector of marginal correlations of predictors with the response variable, rescaled by the standard deviation of the response. Then we sort the elements of this vector in a decreasing order. For any given $0 < \gamma < 1$, we define a submodel

$$M_\gamma = \{1 \leq i \leq p: the\ variables\ corresponding\ to\ the\ first\ [\gamma n]\ largest\ |\omega_i|\}$$

where $[\gamma n]$ denotes the integer part of γn. So we shrink the dimension of the full model down to a submodel with size $d = [\gamma n] < n$.

2.3.2 Sure Independence Screening Procedure Based on the Distance Correlation (DC-SIS)

The first method to screen non-zero features from a high-dimensional model is a new feature screening procedure for ultrahigh-dimensional data based on distance correlation (DC). This idea was first put forward in Szekely, Rizzo, and Bakirov (2007) and Szekely and Rizzo (2009). They showed that Distance correlation has properties of a true dependence measure, analogous to product moment correlation ρ. Distance correlation satisfies 0 ≤ R ≤ 1, and R = 0 only if X and Y are independent. In the bivariate normal case, R is a function of ρ, and $\mathrm{R}(\mathrm{X},\mathrm{Y}) \leq |\rho(\mathrm{X},\mathrm{Y})|$with equality when $\rho = \pm 1$, which is following the formula:

$$dcorr(u,v) = \frac{\rho\, arcsin(\rho) + \sqrt{1-\rho^2} - \rho\, arcsin\left(\frac{\rho}{2}\right) - \sqrt{4-\rho^2} + 1}{1 + \frac{\pi}{3} - \sqrt{3}}$$

Thus, the DC of two univariate normal random variables is also a strictly increasing function of the absolute value of the Pearson correlation of these two normal random variables.

Runze Li , Wei Zhong & Liping Zhu (2012) proved that the sure screening property based on the following theorem. Suppose two conditions:

(C1) Both x and y satisfy the sub-exponential tail probability uniformly in p. That is, there exists a positive constant S_0 such that for all $0 < S \leq 2\, S_0$,

$$\sup \max E\left\{\exp\left(s\left|\left|X_k\right|\right|_1^2\right)\right\} < \infty,\ and\ E\{\exp\left(s\left|\left|X_k\right|\right|_q^2\right)\} < \infty.$$

(C2) The minimum DC of active predictors satisfies

$$\min_{k \in D} w_k \geq 2\, cn^{-k},$$

for some constants c > 0 and 0 ≤ k < 1/2.

$Theorem. Under\ condition (C1), for\ any\ 0 < \gamma < \frac{1}{2} - k,$

$there\ exist\ positive\ constants\ c_1 > 0\ and\ c_1 > 0\ such\ that$

$Pr(\max_{1 \leq k \leq p} |\widehat{w_k} - w_k| \geq cn^{-k}) \leq O(p[exp\{-c_1 n^{1-2(k+\gamma)}\} + n\ exp(-c_2 n^{\gamma})]).$

$Under\ condition (C1)\ and\ (C2), we\ have\ that$

$Pr(D \subseteq \widehat{D^*}) \geq 1 - O(s_n[exp\{-c_1 n^{1-2(k+\gamma)}\} + n\ exp(-c_2 n^{\gamma})]).$

The sure screening property holds for the DC-SIS under milder conditions than that for the SIS (Fan and Lv 2008) in that we do not require the regression function of y onto x to be linear. Thus, the DC-SIS provides a unified alternative to existing model-based sure screening procedures. These two essential properties allow us to use the DC for feature screening in ultrahigh-dimensional data. The proposed DC-SIS can be directly employed for screening grouped variables, and can be directly applied for ultrahigh-dimensional data with multivariate responses, which is called a model-free screening procedure.

The distance is denoted by $||\varphi||^2 = \varphi\% * \%\varphi$, which φ is a complex-value function and $\bar{\varphi}$ is the conjugate of φ. Then the square of covariance between two random vectors u=$\varphi_u(t)$ and v=$\varphi_v(s)$ is given by

$$dcov^2(u,v) = \int_{R^{d_u+d_v}} ||\varphi_{u,v}(t,s) - \varphi_u(t)\varphi_v(s)||^2 / \{c_{d_u} c_{d_v} ||t||_{d_u}^{1+d_u} ||s||_{d_v}^{1+d_v}\} dtds$$

c_{d_u} is denoted as a constant which equals to $\pi^{(1+d)/2}/\Gamma\{(1+d)/2\}$.

Then the distance correlation coefficient can be defined as

$$dcorr(u,v) = \frac{dcov(u,v)}{\sqrt{dcov(u,u)dcov(v,v)}}$$

The value of $dcorr\ (u,v)$ will be used later.

Now we consider how to calculate $dcov(u,v)$. Szekely, Rizzo, and Bakirov (2007) stated that

$$dcov^2(u,v) = S_1 + S_2 - 2\,S_3$$

Where S_1, S_2, S_3 are defined as:

$$S_1 = E\left\{\left|\left|u-\tilde{u}\right|\right|_2 \left|\left|v-\tilde{v}\right|\right|_2\right\},$$

$$S_2 = E\left\{\left|\left|u-\tilde{u}\right|\right|_2\right\} * E\left\{\left|\left|v-\tilde{v}\right|\right|_2\right\},$$

$$S_3 = E\left\{E\left(\left|\left|u-\tilde{u}\right|\right|_2 \middle| u\right) * E(\left|\left|v-\tilde{v}\right|\right|_2 | v)\right\}.$$

Based on the correlation coefficients between y and x_i, without specifying a regression model, we define the index set of the active and inactive predictors by two sets X_D and X_I.

In an ultrahigh-dimensional setting, the dimensionality p greatly exceeds the sample size n. It is thus natural to assume that only a small number of predictors are relevant to y. Denote by F (y | x) the conditional distribution function of y given x.

So that elements in X can be divided into these two sets. Where D represents the set of k that the $F\ (y|x)$ functionally depend on x_k while I represents the set of k that the $F\ (y|x)$ does not functionally depend on x_k. We consider using ω_k as a marginal utility to rank the importance of X_k at the population level. We use the DC because it allows for arbitrary regression relationship of y onto x, regardless of whether it is linear or nonlinear.

The DC also permits univariate and multivariate responses, regardless of whether it is continuous, discrete, or categorical. Moreover, it allows for group wise predictors. cn^{-k} is denoted as a principle to distinguish those k features, where c and k are constant value and n is the number of equations. If $\widehat{\omega_k} \geq\ cn^{-k}$, this x_k can be put into the set of $\widehat{D^*}$, otherwise put it into I.

Chapter 3

Screening and Selection Procedure in Nonlinear Model

There are two crucial stages in the whole procedure to pick out important covariates from a nonlinear high-dimensional model. The first one is the screening stage, which could help us to reduce the dimension of the model and benefit to the speed of calculation in the second stage. In this stage, we use several measure approaches to obtain the correlation coefficients between each x_i and the response y. Then we rank all the correlation coefficients in the decreasing order and the $\frac{2*n}{\log(n)}$ top coefficients will be accept, where we let the constant γ in SIS equal to $\frac{2}{\log(n)}$. We will use some methods to record the root of the coefficients that enter into the second stage.

The second stage is the selection stage, which exactly shows how the covariates influence the response y. Here we use the residual of the current model $\tilde{y}(\zeta)$ to replace y, according to current sequential approaches. The measure methods we apply in the former can still be used here and now we only choose the largest correlation coefficient. The covariate corresponding to this coefficient will be added into the model set ζ. The new ζ should be compared with the former one under a stopping rule. If the new value is smaller, the procedure will continue and update the value of stopping rule. Otherwise, it stops and output the final set. In this article, we would use the Extended Bayesian Information Criterion proposed by Chen and Chen (2008).

3.1 Nonlinear Sparse Model

The model will be generated randomly and each covariate will follow the same distribution. We apply a bunch of nonlinear functions $\varphi_j(x_i)$ to connect the first p x_{ki} with y, which can be presented as following:

$$Y_i = \Sigma_k \Sigma_j \beta_{kj} \varphi_j(x_{ki}) + \varepsilon_i$$

where $i = 1 \dots n, k = 1 \dots p, j = 1 \dots m$. m means that there is a function space $\Psi = \{\varphi_j(x), j = 1 \dots m\}$ operating on x_{ki} and the dimension of the space is m. We require p is a small number and n is smaller than the number of candidates q as well. ε_i is the random error.

So the other way to demonstrate this equation is

$$Y_i = \Sigma_k \beta_k Z_{ki} + \varepsilon_i$$

Where $Z_{ki} = \Sigma_j \varphi_j(x_{mi})$, β_k is a vector of $(\beta_{i1} \dots \beta_{mi})$ if there are m functions in the bunch. Therefore it is a linear model of Y and Z, but a nonlinear sparse model of Y and X.

3.2 Extended Bayesian Information Criterion

Up to now, people have proposed various criteria in model selection approach. Akaike (1973) introduce a classical Akaike information criterion called AIC. Schwarz (1978) propose the Bayesian information criterion called BIC, Stone (1974) and Craven Wahba (1979) put forward cross-validation (CV) and generalized cross-validation (GCV), which are all widely used. But the drawback of those criteria is that in high-dimensional case, when the number of variables is extremely large while the number of observation is small, the probability of choosing irrelevant covariates could be high.

It is well known that BIC is consistent under some standard conditions such as that p is fixed. In non-regular problems such as changing point analysis, the root-n consistency of $\widehat{\theta}\ (s)$ may be violated, yet model selection using BIC is still consistent. Nevertheless, it still has some drawbacks. The precision of the Laplace approximation is influenced by the specific form of the prior density on $\theta\ (s)$ and the correlation structure between observations. The latter affects the interpretation of the sample size n in the definition of BIC (S). Clyde et al.(2007) and the work of Berger have focused on the marginal likelihood $m\ (Y|s)$ and rectified the problems caused by the Laplace approximation. However, they have not targeted the problems that could be caused by large model space. These criteria perform terribly in high-dimensional case, when the number of sample is much smaller than the number random variables. In this case, the probability of selecting wrong covariates would increase.

EBIC, proposed by Chen and Chen (2008), could make up the shortages of the several former criteria mentioned above. They propose an extended family of BIC called EBIC, which takes

number of unknown parameter and size of the model space into account together. This article not only proves the consistency of this criteria but also derives desirable properties especially in the case of large model space and those won't be influenced even covariates are heavily collinear.

Given n independent observations (y_i, x_i), $i = 1 \ldots n$. I denote The conditional density function of (y_i, x_i) is $f(y_i|x_i, \theta)$, here $\theta \in R^q$, R^q is a q dimensional space. Therefore the likelihood function of θ can be written as following

$$L_n(\theta) = \prod_{i=1}^{n} f(y_i|x_i, \theta)$$

Let s represent the subset of p features and denote $\theta\ (s)$ as corresponding parameters. In Schwarz's paper, the BIC procedure is to minimize

$$BIC(s) = -2lnL_n\{\hat{\theta}(s)\} + |s|\ln(n)$$

Here $\hat{\theta}(s)$ is the maximum likelihood estimator of $\theta\ (s)$ and $|s|$ is the number of the elements in set s.

Where $\hat{\theta}\ (s)$ is the maximum likelihood estimator of $\theta\ (s)$ and $|\mathrm{s}|$ is the length of s. Let S be the model space and let p(s) be the prior probability of model whose parameter is set s. Assume that $\pi\ \{\theta\ (s)\}$ is the density of $\theta\ (s)$ given s, then the posterior probability of s can be written as

$$p(s|Y) = \frac{m(Y|s)p(s)}{\sum_{s \in S} p(s)m(Y|s)}$$

Where $m(Y\ |s)$ is the likelihood of model with parameter s, and formed as

$$m(Y|s) = \int f\{Y; \beta(s)\}\pi\{\beta(s)\}d\beta(s)$$

By the theory of Bayes paradigm, our purpose is to find the s^* which maximize the posterior probability. Since the denominator of $p(s|\ Y)$ is a constant. $s^* = arg\ \max_{s\in S} m(Y|s)p(s)$.

Assume S_j is the class of models containing j variables. $p(s)$ is a constant and the probability assigned to S_j is proportional to their sizes. It is obvious that the size of S_j increases as j increases to $j\ = \frac{q}{2}$, thus the probability assigned to S_j by the prior increases almost exponentially. Models that have a larger number of covariates get much higher probabilities than models have fewer covariates. This obviously makes no sense.

In order to make a remedy, Chen and Chen (2008) consider a more reasonable prior during the Bayesian approach. Suppose the model space S is partitioned into $\bigcup_{j=1}^{q} S_j$, then each S_j has the same dimension. The extended BIC is expressed as following

$$BIC_\gamma(s) = -2\ln L_n\{\hat{\theta}(s)\} + |s|\ln(n) + 2\gamma\ln\ \tau(S_j)$$

Here partition the model space S into $U_{j=1}^{p} S_j$ and let $\tau\ (S_j)$ be the length of S_j. For instance, S_j is the set of all models with j covariates, $\tau(S_j) = \binom{q}{j} = \binom{q}{|s|}$.

The EBIC has a particular form in the context of the multi-response model with group structures as follows. Let m denote the number of nonzero entries in the whole model, p is

the total number of the covariates, n is the number of observations. Let ζ be the set of the covariates selected. Then

$$EBIC(\zeta) = n\ln\left|\left|\tilde{Y}\right|\right|^2 + m\ln(n) + 2\left(1 - \frac{\ln(n)}{2\ln(p)}\right) ln\binom{p}{m}.$$

Where $\tilde{Y}$ is the residual of the new model of $y = \beta^*\zeta$ and it is calculated by $\tilde{Y} = (I - \zeta\ (\zeta^T\ \zeta\ )^{-1}\zeta^T)y$. This formula we will use again and again in the procedure in simulation.

Let ζ_0 denotes the set of true features and ζ_* is the set of the set of features that we select under the principle of EBIC. Then it has been proven that EBIC enjoys a ideal property of Selection Consistency,

$$P\left(EBIC(\zeta_0) < \min_{\zeta_* \neq \zeta_0} EBIC(\zeta_*)\right) \to 1 \quad as\ the\ observation\ n \to \infty$$

With probability converging to 1, the set of the selected feature ζ_* according to EBIC converges to ζ_0, the true set of relevant features.

3.3 Distance Correlation Measurement

According to Szekely, Rizzo and Bakirov(2007), we use the estimator of distance correlation coefficient to replace the exact value. Here we first estimate the covariance. $\mathrm{dcov}^2(y,x)$ can be estimated as a new form that

$$\widehat{\mathrm{dcov}}^2(u,v) = \widehat{S_1} + \widehat{S_2} - 2\widehat{S_3}$$

where $\widehat{S_1}, \widehat{S_2}$ *and* $\widehat{S_3}$ *are also estimators*

$$\widehat{S_1} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} ||y_i - y_j||_2 ||x_i - x_j||_2$$

$$\widehat{S_2} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} ||y_i - y_j||_2 \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} ||x_i - x_j||_2$$

$$\widehat{S_3} = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} \left|\left|y_i - y_l\right|\right|_2 \left|\left|x_j - x_l\right|\right|_2$$

Based on $\widehat{\mathrm{dcov}}^2(y,x)$, $\widehat{\mathrm{dcov}}^2(y,y)$, $\widehat{\mathrm{dcov}}^2(x,x)$ we get in advance, the estimated distance correlation coefficient can be represented as

$$\widehat{dcorr}^2(y,x) = \frac{\widehat{\mathrm{dcov}}^2(y,x)}{\widehat{dcov}(y,y)\,\widehat{dcov}(x,x)} = \gamma^2$$

Here γ is the value we will contrast in the algorithms later.

3.3.1Distance Correlation Screening and Selection Approach via Covariates

This method includes two stages, screening and selection. In the first stage, we measure the relationship between each covariate $x_i (i = 1 \dots q)$ and response y by distance correlation approach. After getting q correlation coefficients, we rank them in the decreasing order and record their location in the set. The top $\frac{2\,n}{log(n)}$ covariates will be screened into the D set, which means y may depend on this covariate, and the chosen covariates will go to the next stage.

In the second stage, we use the $EBIC$ criterion as the stopping rule. The EBIC formula is following:

$$EBIC(\zeta) = n \ln \left|\left|\tilde{Y}\right|\right|^2 + p \ln(n) + 2\left(1 - \frac{\ln(n)}{2\ln(q)}\right) ln \binom{q}{p}$$

And we still use distance correlation approach as the measurement and find the maximum one each time and add the corresponding covariate to the final set ζ_*. If the $EBIC$ of the new ζ_* is smaller than that of former one, update the value of $EBIC$ and continue to select. Otherwise, this covariate should not stay in ζ_* and the whole procedure is stopped. These steps perform twice in this procedure. The first time is to pick out the relevant covariates and the second time is to check each element in covariate groups based on the result of the first time.

Initial step:

- Write the function of DC
- Repeat:
  - Compute the distance correlation coefficient between each covariate $x_i (i = 1 \dots q)$ and response y.
  - Put it into set DCORR
- Rank DCORR in the decreasing order
- Put the first $\frac{2\,n}{log(n)}$ covariates into the set x_list and record their location in the original set. (Screening stage is over.)
- Set $\zeta_* = \emptyset$ and $\tilde{Y} = Y$
- Set stop=0 as the stopping signal
- Repeat:
  - Add each covariate x_i in x_list to the ζ_* respectively.
  - Compute the residual $\tilde{Y} = \left(I - \zeta_* \left({\zeta_*}^T \zeta_* \right)^{-1} {\zeta_*}^T \right) Y$ and the DC coefficient between x_i and $\tilde{Y}$.
  - Choose the largest one and add the corresponding x_i to ζ_*.
  - If $EBIC\ (\zeta_*)$ is small than the former $EBIC$, then update the value of $EBIC$ and delete the x_i from set x_list.
  - Otherwise, delete x_i from ζ_* and set stop=1 to end the loop
- Now we get the candidate set, then we pick out the true feature from each candidate group
- Set $\zeta_* = \emptyset$ and $\tilde{Y} = Y$
- Set stop=0 as the stopping signal
- Repeat:

- Add each covariate x_{ij} in new_record to the ζ_* respectively.
- Compute the residual $\tilde{Y} = \left(I - \zeta_* \left({\zeta_*}^T \zeta_* \right)^{-1} {\zeta_*}^T \right) Y$ and the DC coefficient between x_i and $\tilde{Y}$.
- Choose the largest one and add the corresponding x_{ij} to ζ_*.
- If $EBIC\ (\zeta_*)$ is small than the former $EBIC$, then update the value of $EBIC$ and delete the x_i from set new_record.
- Otherwise, delete x_i from ζ_* and set stop=1 to end the loop

- Output the final set and compute the PDR and FDR.
- Output the time that the procedure use

3.3.2 Distance Correlation Screening and Selection Approach via Group Covariates

The group is different from x because here $x_i = (x_{i1} \dots x_{ik})$, which is a vector, and each element in this vector has n observations, so actually now x_i is a $n * k$ matrix. These elements are related to each other by a bunch of nonlinear functions such as $x^2 \dots x^k$. So we also change the estimator of DC coefficient in some place. Here in the formula, $||x_i - x_j||_2$ is no long the norm of two value. It is the 2-norm of two vectors and each vector is an observation of x_i.

The rest of the approach is similar to that via single x. It is also suitable to the EBIC criterion. But here since we focus on $3\ p$ important covariates, the EBIC formula should change to the following one:

$$EBIC(\zeta) = n \ln \left|\left|\tilde{Y}\right|\right|^2 + 3\ p \ln(n) + 2\left(1 - \frac{\ln(n)}{2\ln(q)}\right) ln \binom{q}{p}$$

However, in the last section of the whole procedure that is to pick out relevant covariates from the selected covariate groups, we still need to use the algorithm for covariate itself to find the final result. The algorithm via covariate group is just operated in the stage of screening and the process of picking out relevant covariate groups from the candidate set.

3.4 Canonical Correlation Screening and Selection Approach

3.4.1 the Screening Stage

This method also has two stages, screening and selection. In the first stage, we measure the relationship between each group covariate $x_i = (x_{i1} \dots x_{ik})(i = 1 \dots q)$ and response y by canonical correlation approach. Through the function of canonical correlation, we can obtain q correlation coefficients. We rank them in the decreasing order and record their location in the set. The top $\frac{2n}{log(n)}$ covariates will be screened into the D set, which means y may depend on this covariate, and the chosen covariates will go to the next stage.

Initial step:

- Import the function of Canonical correlation
- Repeat:
  - Compute the canonical correlation coefficient between each group $x_i(i = 1 \dots q)$ and response y.
  - Put it into set AA
- Rank AA in the decreasing order
- Put the first $\frac{2n}{log(n)}$ groups into the set x_list and record their location in the original set.

3.4.2 the Selection Stage

We need to divide this stage into two parts. In the first part, we focus on the whole group and tend to pick out the significant groups $x_i = (x_{i1} \dots x_{ik})$. We still use canonical correlation approach to measure the relationship between each group x_i and response y. We choose the biggest one each time and add the corresponding covariate to the final set ζ_*. If the $EBIC$ of the new ζ_* is smaller than that of former one, update the value of $EBIC$ and continue to select. Otherwise, this covariate should not stay in ζ_* and the whole procedure is stopped. Thus we attain a selection set.

In the second part, based on the selection set we get in the first part, we analysis every elements in each group and try to pick out the covariates that truly relate to the response y. It could share the same procedure with the part one but here we use $x_{ij}, j = 1 \dots k$, rather than the whole group. Repeat the algorithm, finally we can obtain an exact set of the covariates x_{ij} that relate to y.

Initial step:

- Set $\zeta_* = \emptyset$ and $\tilde{Y} = Y$
- Set stop=0 as the stopping signal
- Repeat:
  - Add each group x_i in x_list to the ζ_* respectively.
  - Compute the residual $\tilde{Y} = \left(I - \zeta_* \left({\zeta_*}^T \zeta_* \right)^{-1} {\zeta_*}^T \right) Y$ and the CC coefficient between x_i and $\tilde{Y}$.
  - Choose the largest one and add the corresponding x_i to ζ_*.

- If $EBIC\ (\zeta_*)$ is small than the former $EBIC$, then update the value of $EBIC$ and delete the x_i from set x_list.
- Otherwise, delete x_i from ζ_* and set stop=1 to end the loop

- Let $record = \zeta_*$ and extend $record$ to the size of $k * length\ (record)$ as new_record
- Set $\zeta_* = \emptyset$ and $\tilde{Y} = Y$
- Set stop=0 as the stopping signal
- Repeat:
  - Add each covariate x_{ij} in new_record to the ζ_* respectively.
  - Compute the residual $\tilde{Y} = \left(I - \zeta_* \left({\zeta_*}^T \zeta_* \right)^{-1} {\zeta_*}^T \right) Y$ and the CC coefficient between x_i and $\tilde{Y}$.
  - Choose the largest one and add the corresponding x_{ij} to ζ_*.
  - If $EBIC\ (\zeta_*)$ is small than the former $EBIC$, then update the value of $EBIC$ and delete the x_i from set new_record.
  - Otherwise, delete x_i from ζ_* and set stop=1 to end the loop
- Output the final set and compute the PDR and FDR.
- Output the time that the procedure spend

Chapter 4

Simulation Studies

4.1 Simulation Settings

In this chapter, we set four different conditions to explore the performance of these three algorithms. We will test the PDR and FDR of their output as well as the running time they spend. Here we focus on the quadratic problem, which can be expressed as

$$y_j = \sum_{i=1}^{p} \left(a_i \, x_{ij} + b_i \, {x_{ij}}^2 + c_i \;\; {x_{ij}}^3 \right) + \varepsilon_i$$

Here ε_i is the random error following the standard normal distribution $N\ (0,1)$ and parameters $\beta = (a_i, b_i, c_i)$ is a constant vector.

The number of relevant covariates is $p = 5$ and the total number of covariates is $q = 100$. The coefficients $\beta_i = (a_i, b_i, c_i)$ are generated in a special way. They are considered as independent random variables distributed as $(-1)^u (4n^{-0.15} + |z|)$, where $u \sim Bernoulli\ (0.4)$ and z is a random variable following normal distribution with mean 0 and the probability of $|z| \geq 0.1$ is 0.25, referring to (Luo and Chen (2014)). Thus we compute the standard deviation of z is 0.08693011. We repeat the above steps for 15 times to get 15 independent coefficients. Then we pack each three in a group as $\beta_i = (a_i, b_i, c_i)$.

Secondly, we need to randomly arrange those relevant covariates in each covariate group. We use a set of coefficient R that follows the Bernoulli distribution with the probability 0.6. The probability could also be changed to change the number of covariates that relevant to Y, so we also test the situation that probability is equal to 0.8. R is consisted of 1 and 0 and generated three elements each time as $r_i, i = 1,2,3,4,5$, for example $r_1 = (1,0,1)$. Particularly, if the r_i that we generated is $(0,0,0)$, we will generate it again. We do the above step for five times and combine them together to obtain $R = \{0,0,1,0,1,1,0,1,0,1,1,1,0,1,0\}$, whose size is $3 * p = 15$. So we renew the above equation to the new form

$$y_j = \sum_{i=1}^{p} \left(a_i \, r_{i_1} \, x_{ij} + b_i \, r_{i_2} \, {x_{ij}}^2 + c_i \; r_{i_3} \, {x_{ij}}^3 \right) + \varepsilon_i$$

So the relationship between X and Y is randomized by R in that some $r_{i_t} = 1$ and others are 0. In a X_i group, it's possible that only x_i is relevant to y while in another X_j group there may be $(x_j, {x_j}^2)$, $(x_j, {x_j}^3)$, $({x_j}^2, {x_j}^3)$ or $(x_j, {x_j}^2, {x_j}^3)$ that is relevant to Y. The true relationship will be recorded by a set in our simulation.

Then we consider the distribution of covariates. The covariates can be independently and identically distributed, which is the simplest condition. The covariate set is generated from standard normal distribution in the simulation.

It is also possible that there are some covariates that depend on some other covariates. Here we provide a matrix

$$\Sigma = \begin{pmatrix} 1 & 0.5 & 0 & & & & \\ 0.5 & 1 & 0.5 & \cdots & & 0 & \\ 0 & 0.5 & 1 & & & & \\ & \vdots & & \ddots & & \vdots & \\ & & & & 1 & 0.5 & 0 \\ & 0 & & \cdots & 0.5 & 1 & 0.5 \\ & & & & 0 & 0.5 & 1 \end{pmatrix}$$

We generate covariate set that follows the multivariate normal distribution $N\ (0, \Sigma)$.

We test how the algorithms perform in these two situations. For each situation, we will also perform the two different R sets mentioned in previous paragraph.

Under the settings above, we seek for the effect of sample size as well. For each setting, we change the sample size from a small one to a larger one. This sample size cannot be too small so that most true features should be detected. A large sample size is also inappropriate since we focus on the small-n-large-q structure. Here we set $n_1 = 50$ and $n_2 = 70$ for q=100 candidate covariates.

Thus totally we performs three algorithms under these 8 different settings and analysis the outputs to find their edges as well as drawbacks.

To assess the performance of feature selection, PCR (positive discovery rate) and FDR (false discovery rate) are often used. The PDR and FDP for a sample size of n is defined as following:

$$PDR = \frac{|\zeta_* \cap \zeta_0|}{|\zeta_0|}$$

$$FDR = \frac{|\zeta_* \cap \zeta_0{}^C|}{|\zeta_0|}$$

The ideal situation is that the selection method outputs a set ζ_* which has a large value of PDR and a small FDR. When $PDR \to 1$ and $FDR \to 0$, this selecting procedure is consistent. The former BIC criterion would cause high PDR and FDR at the same time in the simulation but extended BIC is able to reduce FDR in that it always puts fewer covariates into the ζ_*. It contains the edges of those classic feature selection criteria and overcome their drawbacks along with limitations. Therefore, EBIC is the superior feature selection criterion applied in a high dimensional model.

4.2 Simulation Results

Table 4.1 represents the values of PDRs and FDRs along with the computation time for canonical correlation method and two distance correlation methods. The number of relevant covariates here is around 9 and the sample size n=50. And Table 4.2 has the same setting as Table 4.1 except the sample size n=70. Table 4.3 and 4.4 represents PDRs, FDRs and the computation time for the condition that the number of relevant covariates is around 12. The exact number is not consistent in that it's generated randomly. The difference between table 4.3 and 4.4 is the sample size n. From table 4.1 to table 4.4, all the candidate covariates are independent.

Table 4.5 represents PDRs, FDRs and the computation time when dependent covariates have a quasi-diagonal matrix as the covariance matrix. The number of relevant covariates here is around 9 and the sample size n=50. Similarly to previous paragraph, Table 4.6 is the situation of sample size n=70. Table 4.7 and 4.8 represents the results of situations that the number of relevant covariates is around 12 and the covariates are dependent with their neighbours. Table 4.7 is the result of sample size n=50 and Table 4.8 is the result of sample size n=70.

Value \ Method	CC	DC via group	DC via covariate
PDR	0.8888889	0.6666667	0.6666667
FDR	0.6	0.7	0.7
Time	0.512	143.322	122.759

Table 4.1 n=50, covariates are i.i.d., the probability of Bernoulli distribution p=0.6

Value \ Method	CC	DC via group	DC via covariate
PDR	1	0.7	0.7
FDR	0.6153846	0.7692308	0.7692308
Time	0.716	646.529	568.896

Table 4.2 n=70, covariates are i.i.d., the probability of Bernoulli distribution p=0.6

Value \ Method	CC	DC via group	DC via covariate
PDR	0.75	0.6923	0.6923
FDR	0.6538462	0.55	0.55
Time	0.661	149.162	137.351

Table 4.3 n=50, covariates are i.i.d., the probability of Bernoulli distribution p=0.8

Value \ Method	CC	DC via group	DC via covariate
PDR	0.9166667	0.7545455	0.7545455
FDR	0.5769231	0.80769	0.80769
Time	0.638	748.596	633.926

Table 4.4 n=70, covariates are i.i.d., the probability of Bernoulli distribution p=0.8

Value \ Method	CC	DC via group	DC via covariate
PDR	0.8461538	0.7692308	0.7692308
FDR	0.45	0.5	0.5
Time	0.586	141.350	123.241

Table 4.5 n=50, covariates follow $N\ (0, \Sigma)$, the probability p=0.6

Value \ Method	CC	DC via group	DC via covariate
PDR	0.8888889	0.7777778	0.7777778
FDR	0.7241379	0.7307692	0.7307692
Time	0.734	657.201	570.006

Table 4.6 n=70, covariates follow $N\ (0, \Sigma)$, the probability p=0.6

Value \ Method	CC	DC via group	DC via covariate
PDR	0.8	0.4666667	0.4666667
FDR	0.5384615	0.65	0.65
Time	0.672	142.293	119

Table 4.7 n=50, covariates follow $N\ (0, \Sigma)$, the probability p=0.8

Value \ Method	CC	DC via group	DC via covariate
PDR	0.8333333	0.6666667	0.6666667
FDR	0.6153846	0.6923077	0.6923077
Time	0.694	673.756	643

Table 4.8 n=70, covariates follow $N\ (0, \Sigma)$, the probability p=0.8

4.3 General Comparison and Analysis

From the comparison between Table 4.1/4.2, Table 4.3/4.4, Table 4.5/4.6 and Table 4.7/4.8, PDR will increase significantly with the growth of the sample size in that there are more observations that can be referred. At the same time, FDR increase a little in all situations. It is likely because there is more noise in the model when the sample size becomes large. This property holds the robustness and fails to be influenced by the number of relevant covariates as well as the dependence of covariates. Furthermore, it is natural that when the sample size climbs, the time of algorithm will be longer due to more calculation tasks. However, the time of CC method just rises a little while the others surge significantly.

From Table 4.1/4.3 and Table 4.5/4.7, it is obvious that the size of relevant covariates will influence the accuracy of these algorithms when the sample size is not big enough. The algorithms can perfect pick out the features when there are only several relevant features. When it comes to a large number, their accuracy will decline in that the PDR is decreasing and FDR is rising in the table because for the same sample size, the increase of relevant covariates means fewer observations can be referred for each covariate.

Comparing Table 4.1-4.4 with Table 4.5-4.8, we find that for canonical correlation algorithm, the PDRs are always close to one and there is not a huge gap between situation of independent covariates and dependent covariates. The FDRs fluctuate from 0.4 to 0.7 and it seems not to exist an obvious trend.

Moreover, for the two types of distance correlation algorithm, the value of PDRs are not so stable and reach the bottom in table 4.7 which is below to 0.5 but it is still hard to say there are significant differences between these two conditions.

Besides, in each table, we find that the distance algorithm that analyze the covariates group and the algorithm that analyze candidate covariates itself always have the same results. However the later one spends a bit less computation time than the former one under the same setting due to the simplicity in its algorithm. Hence, we recommend to analysis each covariate directly rather than calculating the three covariates in a group together every time.

According to the data in each table, we find that CC approach has a higher PDR in all situations. Sometimes it is just a little high than that of DC approach and sometimes it has a significantly better performs better. The PDR of DC approach is also acceptable in most times but in table 4.1, 4.7 and 4.8 it does not have an ideal result. Meanwhile the values of FDRs of DC algorithm are also higher than that of CC algorithm in every table. That means CC approach can make less wrong choices during the process of selection. It can be said that the CC approach is superior to the others in accuracy and DC approach is not a terrible method in general as well.

The other advantage of canonical correlation method is the time complexity. When the sample size is large, the time that DC methods take will climb to hundreds of seconds, which is an incredible level. In each table, it is obvious that

under the same setting the canonical correlation approach is much more efficient as its running time can be limited in a second even in an extremely considerable sample size.

Overall, distance correlation method can be operated in feature selection in a high-dimensional nonlinear model. However, canonical correlation method is the superior way in any aspect in the sequential group search procedure. Therefore we recommend the sequential canonical correlation selection procedure as a better algorithm to deal with feature selection in nonlinear model.

Chapter 5

Conclusion and Further Studies

5.1 Conclusion

Feature selection in a nonlinear model for high dimensional data has been popular in recent years for it's widely use in genetic, medical research and other fields. Researchers have proposed various classic methods and criteria to select features in a small-n-large-q structure. Based on these achievements, we introduce the sequential selection procedure in nonlinear model via two measurements, distance correlation and canonical correlation.

From our simulation results, we found that (I) the accuracy of CC algorithm is more stable and better than that of DC algorithm in various situations. (II) The CC algorithm has much better efficiency than DC algorithm in running time. (III) If the sample size becomes larger, PDRs will grow and diverge to 1 but the FDRs could also increase a little. (IV) If the number of relevant covariates is rising, the FDR will decline in every algorithm. (V) The DC algorithm via covariates takes less running time in the same situation than The DC algorithm via the covariate groups. Meanwhile, it doesn't loss the accuracy and stability.

When the sample size n is not too large, the distance correlation and canonical correlation are both suitable in the selection procedure, but if the accuracy is required, canonical correlation method is superior. However, when it becomes extremely large, the distance correlation is computationally inefficient. Therefore the canonical correlation approach is the better method.

5.2 Further Studies

When I operate our simulation studies, I do not satisfy the time complexity of distance correlation. According to the formula, there are two or even three nested loops in the DC algorithm, which means if the sample size grows to 10 n, the running time will be 1000 times of the former one. In the future, we can try to improve it by some special methods such as pointer method in C# or creating more storage.

What's more, in practical problems, we could meet some values that make $X^T X$ be singular, where X is the covariate matrix. We can seek for some ways to regularize these values before the estimation so that the error of estimation can be reduced.

Finally, according to our simulation result, although the PDRs are high in most situations but the FDRs are still around 0.5. In the future, we may do some researches on the reduction of FDRs such as proposing a new selection criterion or improving current selection criteria. The reduction of FDRs could decline the variance of the new model in the test procedure.

Reference

[1] Chen, J., Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. Biometrika, 759-771.

[2] Runze Li , Wei Zhong & Liping Zhu (2012) Feature Screening via Distance Correlation Learning, Journal of the American Statistical Association, 107:499, 1129-1139

[3] Sz*é*kely, G. J., Rizzo, M. L. & Nail K. Bakirov (2007) Measuring and testing dependence by correlation of distances, The Annals of Statistics, Vol. 35, No. 6, 2769–2794

[4] Kosuke Yoshida , Junichiro Yoshimoto and Kenji Doya (2017) Sparse kernel canonical correlation analysis for discovery of nonlinear interactions in high-dimensional data, BMC Bioinformatics, 18:108

[5] S*z*ékely, G. J., and Rizzo, M. L. (2009), Brownian Distance Covariance, The Annals of Applied Statistics, 3, 1233–1303

[6] Luo, S. and Z. Chen (2014). Sequential lasso cum EBIC for feature selection with ultra-high dimensional feature space. Journal of the American Statistical Association 109(507), 1229–1240

[7] Cai, T. T. and L. Wang (2011). Orthogonal matching pursuit for sparse signal recovery with noise. IEEE Transactions on Information theory 57(7), 4680–4688.

[8] Fan, J., and Lv, J. (2008), Sure Independence Screening for Ultrahigh Dimensional Feature Space, Journal of the Royal Statistical Society, Series B, 70, 849–911

[9] Li, Y., B. Nan, and J. Zhu (2015), Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure, Biometrics, Vol. 71(2), 354–363

[10] Mark A. Davenport and Michael B. Wakin (2010), Analysis of Orthogonal Matching Pursuit Using the Restricted Isometry Property, IEEE Transactions on Information Theory, Vol. 56, NO. 9, 4395-4401

[11] Robert Tibshirani (1996), Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58, No. 1, 267-288

[12] Jianqing Fan and Runze Li (2001), Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties, Journal of the American Statistical Association, Vol. 96, No. 456, 1348-1360

[13] Jianqing Fan and Jinchi Lv (2010), A Selective Overview of Variable Selection in High Dimensional Feature Space, Statistica Sinica, Vol. 20, No. 1, 101-148

[14] Shan Luo and Zehua Chen (2018), Feature selection by canonical correlation search in high-dimensional multi-response models with complex group structures.

[15] Hansheng Wang (2009), Forward Regression for Ultra-High Dimensional Variable Screening, Journal of the American Statistical Association, Vol. 104, No. 488, 1512-1524

[16] Hui Zou (2006), The Adaptive Lasso and Its Oracle Properties, Journal of the American Statistical Association, Vol. 101, No. 476, 1418-1429

[17] Peter Buhlmann and Eth Zurich (2006), Boosting for High-dimensional Linear Model, The Annals of Statistics, Vol. 34, No. 2, 559–583