ENERGY DISTANCE CORRELATION WITH EXTENDED BAYESIAN INFORMATION CRITERIA FOR FEATURE SELECTION IN HIGH DIMENSIONAL MODELS

Isaac Xoese Ocloo

A Dissertation

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2021

Committee:

Hanfeng Chen, Advisor

Yuning Fu,
Graduate Faculty Representative

Wei Ning

Maria Rizzo

ABSTRACT

Hanfeng Chen, Advisor

In this research, we investigate the sequential lasso method for feature selection in sparse high-dimensional linear models. It was recently proposed by Luo and Chen (2014). In this project, we propose a new method by introducing the energy distance correlation by Székely et al. (2007) to replace the ordinary correlation in Luo and Chen's algorithm. We continue to adopt the extended Bayesian Information Criteria as the stopping criteria in the computing algorithm. The advantage of energy distance correlation is that it is able to detect linear and non-linear association between two variables, while the ordinary correlation can detect only linear part of association between two variables. As a result, it appears that the new method is shown to be more powerful than Luo and Chen's method for feature selections. This is demonstrated by simulation studies and illustrated by two real-life examples. It is shown that the proposed new algorithm is also selection consistent.

For the first part of our research we examine through simulations the model size selection by Adaptive Lasso and SCAD after a sure screening method proposed by Li et al. (2012) using distance correlation is applied to the data first. We observe that the average model size selected was quite high.

In the second part we describe the new sequential variable selection method which we call energy distance correlation with extended Bayesian Information Criteria (Edc+EBIC). At each stage of the sequential procedure we maximize the energy distance correlation between the response and each of the predictor variables. This maximization is done such that if a variable is selected in the previous stage, it's contribution to the response is removed so that it won't have a chance of being selected again. The active set of selected variables is updated once a variable is selected and the EBIC of the set is calculated. The process stops if the EBIC for the current active set is greater than the EBIC of the previous active set. We compare the performance of Edc+EBIC with sequential Lasso, Adaptive Lasso, SCAD and SIS+SCAD. We observed that our proposed method on average has a positive discovery rate close to 100%, a low false discovery rate and an average model size as expected in our simulation set-up.

I dedicate this dissertation to the memory of my father.

ACKNOWLEDGMENTS

Bless the Lord, Oh my soul and all that is within me, bless his holy name. I am thankful to you Lord for bringing me to Ebenezer.

I express my sincere gratitude to my advisor and mentor Dr. Henfeng Chen, whose guidance and supervision made this work a success. He didn't give up on me when I even gave up on myself. He is like a father to me, always encouraging and guiding me through my studies. I appreciate his timely response to all my emails.

I wish to thank the other members of my committee, Dr. Wei Ning, Dr. Yuning Fu and Dr. Maria Rizzo. Their guidance and assistance were invaluable for the completion of this work.

I also wish to thank Dr. Junfeng Shang, Dr. John Chen, Dr. Craig Zirbel, Dr. Jim Albert, Dr. Kimberly Rogers, Dr. Steven Seubert, and all members of the Department of Mathematics and Statistics at Bowling Green State University who taught and mentored me throughout my studies. Additionally and particularly, I appreciate the financial aids from the the Department during the entire period of my study that enabled me to complete my degree, and I wish to thank all professors in the Department under whom I worked as a Graduate Instructor or Research Assistant for all their supports.

I appreciate my wonderful family in Ghana. I am so sad my Dad couldn't live to see this day due to his sudden demise on 05/09/2021. His love and prayers made this journey a success. Thank you Michel Tornyeviadzi for being a good friend and support. To all well-wishers whose names are not mentioned due to limitation of space, I say God richly bless you".

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER 1 INTRODUCTION

1.1 Introduction

High dimensional data is a data set with more features (p or predictors) than the sample size n. It usually comes from genetic research, e-commerce, warehouse data in business, biomedical imaging, functional magnetic resonance imaging and longitudinal data, among many others.

In genomics, high dimensional data has become common due to improvements in single cell technology which has led to increased recognition that cellular heterogeneity is a universal feature of any cell population. In principle, one wants to know, for each single cell, the molecular code of the cell (the genome), the functionality of the cell (the proteome and metabolome) and the connection between the two - the transcriptome, Su et al. (2017). Data to answer these questions is high dimensional because a large number of parameters across thousands of single cells at a given time point are measured. The goal is to infer short DNA-words of approximate length 8 - 16 base pairs, e.g., "ACCGTTAC", where a certain protein or transcription factor binds to the DNA. The response Y_i, measures for example the binding intensity of the protein of interest in the ith region of the whole DNA sequence and X_i contains abundance scores of p candidate motifs (or DNA words) in the ith region of the DNA. The task is to infer which candidate words are relevant for explaining the response Y. Statistically, we want to find the features whose corresponding regression coefficients are substantial in absolute value or significantly different from zero.

High dimensional statistics refers to statistical inference when the number of unknown parameters p is of much larger order than sample size n, that is: $p \gg n$. This encompasses supervised regression and classification models where the number of covariates is of much larger order than n, unsupervised settings such as clustering or graphical modeling with more variables than observations or multiple testing where the number of considered testing hypotheses is larger than sample size. The methodological concepts for high-dimensional statistics share some common aspects with nonparametric statistics and machine learning, Bühlmann and Van De Geer (2011).

The largest development of the science of statistics occurred in the twentieth century Johnstone and Titterington (2009). During this period, most of the motivating practical problems consisted of large sample size (n) and small predictors (p). For a continuous response variable $y_i : i = 1, ..., n$ and predictor variables $X_1, X_2, ..., X_p$, a model to describe the linear relationship between the response variable and the predictor variables is the multiple linear regression equation. The regression equation is of the form

$$\begin{aligned} y_i =& \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i \\ =& \sum_{j=0}^{p} \beta_j x_{ij} + \epsilon_i \end{aligned} \tag{1.1.1}$$

In a matrix notation it can be written as $Y = X\beta + \epsilon$. The predictors are considered fixed while $\beta_0, \beta_1, ..., \beta_p$ are the unknown parameters that are usually estimated by the least-square method. By this

$$\hat{\beta} = argmin_\beta \sum_{i}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2$$

meaning $\hat{\beta}$ is the minimizer of the sum of squares function on the right hand side. In the vector-matrix notation, we can write this in terms of the l_2 norm as

$$\hat{\beta} = argmin \, \|Y - X\beta\|_2^2$$

In this problem $\hat{\beta}$ satisfies

$$X^T X \hat{\beta} = X^T Y$$

and

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

The estimate $\hat{\beta}$ is unique provided $X^T X$ can be inverted. For $X^T X$ to be invertible it requires that $n \geq p$, otherwise $X^T X$ will be singular and the estimate of β will not be unique.

One of the two approaches to the singularity problem and in general in dealing with the large-p-small-n problem is the method of regularization or penalized least squares. An early regularization method, ridge regression by Hoerl and Kennard (1970) estimates the parameter by $\hat{\beta}_{Ridge} = (X^T X + \lambda_2 I)^{-1} X^T Y$ and the positive scalar λ_2 is called the ridge parameter or regularization constant. The following are equivalent formulations of the ridge regression problem.

1. $\hat{\beta}_{Ridge} = argmin_\beta \{\|y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2\}$ for some λ_2
2. $\hat{\beta}_{Ridge}$ = minimizes $\|y - X\beta\|_2^2$ subject to $\|\beta\|_2^2 \leq c_2(\lambda_2)$, for some $c_2(\lambda_2)$, depending on λ_2.
3. $\hat{\beta}_{Ridge}$ = minimizes $\|\beta\|_2^2$ subject to $\|y - X\beta\|_2^2 \leq b_2(\lambda_2)$, for some $b_2(\lambda_2)$, depending on λ_2.

The ridge regression estimation method overcomes the challenge of invertibility but is unable to shrink coefficients to exactly zero and as a result can not perform variable selection. But looking at the problem of large-p-small-n broadly is to consider that it is intuitively plausible that among the larger number of predictors only a small proportion are likely to be influential in predicting the response variable, this is known as sparsity. To make use of this sparsity assumption in the estimation of $\hat{\beta}$, Tibshirani (1996) used a penalty function based on L_1 norm which is $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ leading to the following equivalent formulations.

1. $\hat{\beta}_{Lasso} = argmin_\beta \{\|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1\}$, for some λ_1.
2. $\hat{\beta}_{Lasso}$ = minimizes $\|y - X\beta\|_2^2$ subject to $\|\beta\|_1 \leq c_1(\lambda_1)$, for some $c_1(\lambda_1)$, depending on λ_1.
3. $\hat{\beta}_{Lasso}$ = minimizes $\|\beta\|_1$ subject to $\|y - X\beta\|_2^2 \leq b_1(\lambda_1)$, for some $b_1(\lambda_1)$, depending on λ_1.

The L_1 norm penalty allows coefficients to shrink towards exactly zero. Thus the LASSO usually results into sparse models, that are easier to interpret. Other penalty functions have been considered, such as SCAD (Fan and Li, 2001), which smoothly clips a L_1 penalty (for small $|\beta_j|$) and a

constant penalty (for large $\mid \beta_j \mid$'s), adaptive Lasso(Zou, 2006): $p_\lambda(\mid \beta_j \mid) = \lambda w_j \mid \beta_j \mid$, where w_j are given weights and Minimax concave penalty by Zhang et al. (2010). Cross-validation (CV) is commonly used in these methods for the choice of the regulating parameter.

With the invertibility problem taken care of, coupled with the assumption of sparsity the main goals in the analysis of high dimensional data is to identify the features which have coefficient estimates not equal to zero and are also highly correlated to the response variable, we would refer to these features as the relevant features. To state the main goal more succinctly is the so-called oracle property in feature selection. The oracle property as stated by, Luo and Chen (2014) refers to two asymptotic natures: (i) selection consistency, that is, the sparse relevant features can be exactly selected with probability converging to 1, and (ii) the effects of relevant features can be consistently estimated the same as they would be, were they obtained by knowing the relevant features in advance.

The second approach in analyzing the large-p-small-n problem is sequential variable selection. Various methods have been developed under the sequential approach. The nature of these proposed methods has been to reduce the dimension of the data to $d < p$. There are two forms of these, one is to select from the many features a subset of which we are sure contains the relevant features (predictors), this is referred to as the sure screening property. This is usually followed by a regularization method such as SCAD to identify and estimate the relevant predictors from the reduced feature space. The other form is to sequentially select the relevant features through a repetitive process which terminates when a stopping criteria is met.

A recent addition to sequential feature selection is Sequential Lasso Cum EBIC for feature selection with ultra high dimensional feature space (SLasso), (Luo and Chen, 2014) which solves a sequence of partially penalized least squares problems and uses the Extended Bayesian Information Criteria, (Chen and Chen, 2008) as a stopping criteria. Solving the partially penalized least squares problem reduces to selecting the feature(s) which maximizes the Pearson correlation coefficient between the features and the response variable at each step. It is well known that the Pearson correlation coefficient is for measuring the strength of linear associations. Thus maximiz-

ing the Pearson correlation coefficient may not work well for data structures where the relationship between at least one feature and the response variable is not linear.

Developing a sequential feature selection method which is able to identify and maximize both linear and non-linear relationships that might exist between the features and the response, sets the tone for this research work.

1.2 Background of Problem

In de Siqueira Santos et al. (2014) the Figure 1.1 is adapted which shows the correlation and dependence in biological data.

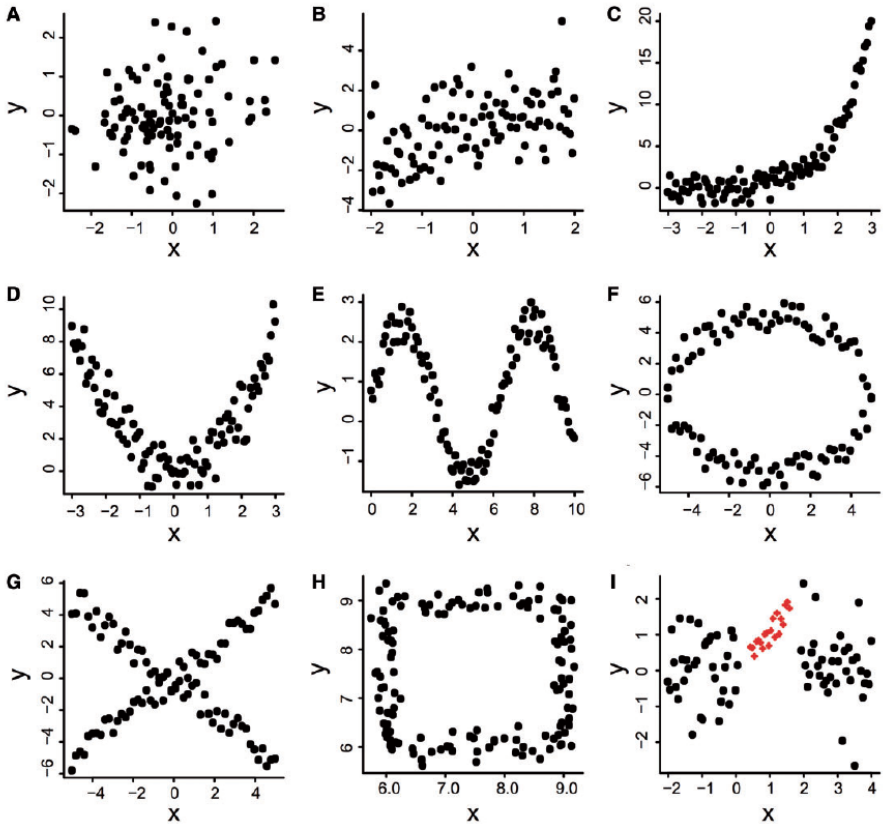

Figure 1.1 (A) Independent data, (B) linear association, (C) exponential association - non-linear monotonic association, (D) quadratic association - non-linear non-monotonic, (E) sine association-non-linear non-monotonic, (F) circumference-non-functional association, (G) cross-non-functional association, (H) square-non-functional association and (I) local correlation- only part of the data is correlated, which is represented by crosses.

With these many possible associations that could exist between any two variables, a method

which only maximizes the linear association between two variables will not be appropriate to detect other relations that may exist between a continuous dependent variable and each of the many features in a high dimensional data.

Aside the well known Pearson's product-moment correlation or simply Pearson's correlation (Pearson, 1920) various parametric and non-parametric measures of correlation coefficients have been developed. Some of them are Spearman's rank correlation coefficient (Spearman, 1904), Kendall τ rank correlation coefficient (Kendall, 1938), Distance correlaiton (Székely et al., 2007), Hoeffding's D measure (Hoeffding, 1948), Heller, Heller and Gorfine measure (Heller et al., 2013) and Maximal information coefficient (Reshef et al., 2011).

1.3 Problem Statement

Let $y_1, ..., y_n$ where n is the number of observations, $X_1, ..., X_p$ are p features, $\beta_0, \beta_1, ..., \beta_p$ are $p + 1$ coefficients of the features and $\epsilon_i, i = 1, ..., n$ iid errors of the response variable. Then a Sparse high-dimensional regression (SHR) model will be

$$y_i = \beta_o + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, i = 1, ..., n. \tag{1.3.1}$$

Maximizing the Pearson correlation between the response variable and the features at each step in the feature selection process by Luo and Chen (2014) does not take into account other forms of association and therefore not likely to select relevant variables which may be non-linearly related to the response.

1.4 Objective of the Research

The objective of this dissertation is to develop an algorithm to sequentially select the relevant variables in a sparse high dimensional linear model by maximizing the energy distance correlation which is able to detect correlations other than the linear correlation. The extended Bayesian Information Criteria will be used as stopping criteria in the feature selection process.

1.5 Significance of the Study

This study is important because it will develop a new algorithm which is able to select all relevant variables in a high dimensional data which may be either linearly or non-linearly related to the response variable.

1.6 Outline of the Dissertation

The remaining parts of the dissertation are organized as follows. Chapter 2 provides a literature review on the two broad approaches to analyzing high dimensional data. Literature on energy distance correlation and extended Bayesian Information criteria will also be reviewed. In chapter 3, we present the proposed algorithm for feature selection which maximizes the energy distance correlation where the extended Bayesian Information Criteria is used as the stopping criteria. In chapter 4, we present simulation results to compare the performance of our proposed method with results from other methods in the literature review. We also present two real-life data examples. Finally, we will discuss the results obtained from the simulation study and real data example, followed by our conclusion, and provide areas for future work in Chapter 5.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

The aim of this chapter is to present a literature review on analyzing high dimensional data. We review literature on the two approaches of analyzing high dimensional data, which are regularized regression approach and sequential approach. We would also review literature on Extended Bayes Information criteria and energy distance correlation.

2.2 High Dimensional Data

Let n be the sample size and p the number of predictors or features. A data set with $p > n$ is called high dimensional data. In many studies there are a large number of variables to measure on each experimental unit compared to the number of experimental units. For example in genetics studies there are many genes but just a relatively few patients to take the measurements on. Also there could be many samples of a person's speech but with a relatively few speakers sampled. In the next two sections we state the curses and blessings of dimensionality as discussed by Donoho et al. (2000).

2.2.1 Curse of Dimensionality

The phrase, "curse of dimensionality" was apparently coined by Richard Bellman, in connection with the difficulty of optimization by exhaustive enumeration on product spaces. Bellman reminded us that, if we consider a cartesian grid of spacing 1/10 on the unit cube in 10 dimensions, we have 10^{10} points; if the cube in 20 dimensions was considered, we would have of course 10^{20} points. His interpretation: if our goal is to optimize a function over a continuous product domain of a few dozen variables by exhaustively searching a discrete search space defined by a crude discretization, we could easily be faced with the problem of making tens of trillions of evaluations of the function. Bellman argued that this curse precluded, under almost any computational scheme then foreseeable, the use of exhaustive enumeration strategies, and argued in favor of his method of dynamic programming.

We can identify classically several areas in which curse of dimensionality appears.

1. *Optimization,* If we must approximately optimize a function of d variables and we know only that it is Lipschitz, say, then we need order $(1/\epsilon)^d$ evaluations on a grid in order to obtain an approximate minimizer within error ϵ.

2. *Function approximation*. If we must approximate a function of d variables and we know only that it is Lipschitz, say, then we need order $(1/\epsilon)^d$ evaluations on a grid in order to obtain an approximation scheme with uniform approximation error ϵ.

3. *Numerical integration*. If we must integrate a function of d variables and we know only that it is Lipschitz, say, then we need order $(1/\epsilon)^d$ evaluations on a grid in order to obtain an integration scheme with error ϵ.

2.2.2 Blessings of Dimensionality

Despite the challenges of high dimensionality, there are some silver linings which are stated below:

1. *Concentration of measure*. The "concentration of measure phenomenon" is a terminology introduced by V. Milman for a pervasive fact about probabilities on product spaces in high dimensions. Suppose we have a Lipschitz function f on the d-dimensional sphere. Place a uniform measure P on the sphere, and let X be a random variable distributed P. Then

$$P\{|f(x) - E(f(x))| > t\} \leq C_1 exp(-C_2 t^2).$$

where C_i are constants independent of f and of dimension. In short, a Lipschitz function is nearly constant. But even more importantly: the tails behave at worst like a scalar Gaussian random variable with absolutely controlled mean and variance. This phenomenon is by no means restricted to the simple sphere case just mentioned. It is also true, in parallel form,

for X taken from the multivariate Gaussian law with density

$$p(x) = (2-\pi)^{-d/2} exp(-\|x\|^2/2).$$

Variants of this phenomenon are known for many high-dimensional situations; e.g. discrete hypercubes Z_2^d and hamming distance. The roots are quite old: they go back to the isoperimetric problem of classical times. Milman credits the probabilist Paul Le'vy with the first modern general recognition of the phenomenon. A typical example is the following. Suppose I take the maximum of d i.i.d. Gaussian random variables $X_1, ..., X_d$. As the maximum is a Lipschitz functional, we know from the concentration of measure principle that the distribution of the maximum behaves no worse than a standard normal distribution in the tails. By other arguments, we can see that the expected value of $\max(X_1, ..., X_d)$ is less than $\sqrt{2log(d)}$. Hence the chance that this maximun exceeds $\sqrt{2log(d)} + t$ decays very rapidly in t.

2. *Dimension asymptotics*. A second phenomenon, well-exploited in analysis, is the existence of results obtained by letting the number of dimension go to infinity. This is often a kind of refinement of the concentration of measure phenomenon, because often when there is a dimension-free bound like the concentration of measure phenomenon, there is a limit distribution for the underlying quantity, for example a normal distribution. Return to the example of the maximum M_d of d i.i.d. Gaussian random variables. As remarked above, we know that the distribution of the maximum behaves no worse than a standard normal distribution in the tails. In fact, long ago Fisher and Tippett derived the limiting distribution, now called the extreme-value distribution. That is, they showed that

$$Prob\{M_d - \sqrt{2\log(d)} > t\} \to G(t)$$

where $G(t) = e^{-e^{-}t}$.

3. Many times we have high-dimensional data because the underlying objects are really continuous-space or continuous-time phenomena: there is an underlying curve or image that we are sampling. Typical examples cited earlier include measurements of spectra, gaits, and images. Since the measured curves are continuous, there is a underlying compactness to the space of observed data which will be reflected by an approximate finite-dimensionality and an increasing simplicity of analysis for large d. A classical example of this is as follows. Suppose we have d equispaced samples on an underlying curve B(t) on the interval [0, 1] which is a Brownian bridge. We have d dimensional data $X_{i,d} = B(i/d)$, and discuss two computations where the large d behavior is easy to spot. First, suppose we are interested in the maximum $max_i X_{i,d}$. Then quite obviously, this tends, for large d to the random variable $max_{t[0,1]} B(t)$, which has an exact distribution worked out by Kolmogorov and Smirnov. Second, suppose are interested in obtaining the principal components of the random vector. This involves taking the covariance matrix

$$C_{i,j} = Cov(X_i, X_j), 1 \leq i, j \leq d$$

and performing an eigenanalysis. On the other hand, the covariance kernel

$$\Gamma(s,t) = Cov(B(s), B(t)), s, t \in [0,1]$$

has the known form $min(s,t) - ts$ and known eigenfunctions $sin(\pi kt), fork = 1, 2, ...$ In this case, the first m eigenvalues of C tend in an appropriate sense to the first m eigenvalues of Γ and the eigenvectors of C are simply sampled sinusoids.

2.3 Regularized and Sequential Approach for High Dimensional Data Analysis

In the past only a few carefully chosen variables were measured for each observation, nowadays any variable that might plausibly have an effect on the response tends to be recorded, example in biological sciences, one may want to classify diseases and predict clinical outcomes using mi-

croarray gene expression or proteomics data, in which tens of thousands of expression levels are potential covariates but there are typically only tens or hundreds of subjects.

2.3.1 Regularized Approach

The regularized approach in the analysis of large-p-small-n problems consists of methods which are designed to penalize a regression equation. This approach selects the features and estimates the coefficients simultaneously by minimizing a penalized sum of squares of the form

$$\sum_{i=1}^{n}\left(y_i-\beta_o-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\sum_{j=1}^{p}p_\lambda(|\beta_j|), \tag{2.3.1}$$

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 $\beta_j's$ are estimated nonzero when λ is set at a certain value. Various penalty functions have been proposed and studied. The penalty function $p_\lambda(|\beta_j|)=\lambda|\beta_j|$ called Least Absolute Selection Shrinkage Operator (LASSO) (Tibshirani, 1996).This penalty function has an additional advantage of variable selection since it's able to shrink some of the coefficient estimates to exactly zero.

2.3.2 A Brief History About Lasso

The Least Absolute Selection Shrinkage Operator (LASSO) was developed by Tibshirani (1996). The motivation for the lasso came from an interesting proposal of Breiman (1993). Breiman's non-negative garotte minimizes

$$\sum_{i=1}^{N}\left(y_i-\alpha-\sum_{j}c_j\hat{\beta}_j^o x_{ij}\right)^2 \qquad \text{subject to} \quad c_j\geq 0,\ \sum c_j\leq t \tag{2.3.2}$$

The garotte starts with the Ordinary Least Squares (OLS) estimates and shrinks them by non-negative factors whose sum is constrained. In extensive simulation studies, Breiman (1993) showed that the garotte has consistently lower prediction error than subset selection and is competitive with ridge regression (a regularized regression method with penalty function $p_\lambda(|\beta_j|)=\lambda\beta_j^2$) except when the true model has many small non-zero coefficients. A drawback of the garotte is that its

solution depends on both the sign and the magnitude of the OLS estimates. In overfit or highly correlated settings where the OLS estimates behave poorly, the garotte may suffer as a result. In contrast, the lasso avoids the explicit use of the OLS estimates. The Lasso minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produre some coefficients that are exactly 0 and hence gives interpretable models. Simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge regression.

2.3.3 Description of LASSO Algorithm

Suppose that we have data (X^i, y_i), i=1,2,. . .,N, where $X^i = (x_{i1}, ..., x_{ip})^T$ are the predictor variables and y_i are the responses. As in the usual regression set-up, we assume either that the observations are independent or that the y_i's are conditionally independent given the x_{ij}'s. We assume that the x_{ij} are standardized so that $\sum_i x_{ij}/N = 0$,$\sum_i x_{ij}^2/N = 1$. Letting $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)$, the lasso estimate $(\hat{\alpha}, \hat{\beta})$ is defined by

$$(\hat{\alpha}, \hat{\beta}) = \text{argmin} \left\{ \sum_{i=1}^{N} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \right\} \tag{2.3.3}$$

$$\text{subject to } \sum_j |\beta_j| \leq t$$

Here $t \geq 0$ is a tuning parameter. Now, for all t, the solution for α is $\hat{\alpha} = \bar{y}$. We can assume without loss of generality that $\bar{y} = 0$ and hence ommit α. Computation of the solution to equation 2.3.3 is a quadratic programming problem with linear inequality constraints.

The criterion $\sum_{i=1}^{N} (y_i - \sum_j \beta_j x_{ij})^2$ equals the quadratic function $(\beta - \hat{\beta}^o)^T \mathbf{X^T X} (\beta - \hat{\beta}^o)$ (plus a constant). The elliptical contours of this function are shown by the full curves in figure 2.1; they are centered at the OLS estimates; the constraint region is the rotated square. The lasso solution is the first place that the contours touch the square, and this will sometimes occur at a corner, corresponding to a zero coefficient.

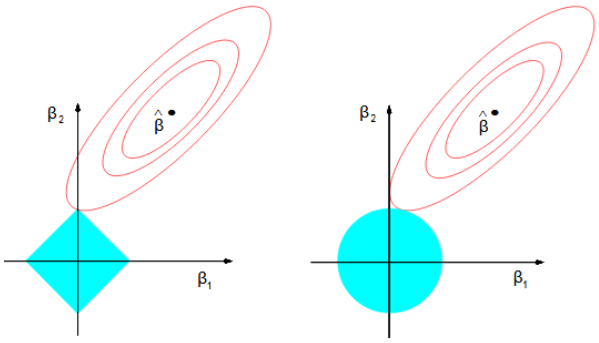

Figure 2.1 Estimation picture for the lasso (left) and ridge regression (right) for two parameters β_1 and $\beta 2$ (from Hastie et al. (2009))

2.3.4 Other Penalty Functions

SCAD, which smoothly clips a L_1 penalty (for small $|\beta_j|$) and a constant penalty (for large $|\beta_j|$'s), adaptive Lasso: $p_\lambda(|\beta_j|) = \lambda w_j|\beta_j|$, where w_j are given weights. Cross-validation (CV) is commonly used in these methods for the choice of the regulating parameter.

2.4 Sequential Methods

Sequential feature selection algorithms are a family of greedy search algorithms (algorithms which always make the choice that seems to be the best at that moment) that are used to reduce an initial p-dimensional feature space to a k-dimensional feature subspace where $k < p$. The motivation behind feature selection algorithms is to automatically select a subset of features that is most relevant to the problem.

A so-called oracle property is of major concern for any feature selection method. The oracle property refers to two asymptotic natures: (i) selection consistency, that is, the sparse relevant features can be exactly selected with probability converging to 1, and (ii) the effects of relevant features can be consistently estimated the same as they would be, were they obtained by knowing the relevant features in advance. For fixed p,it was shown that Lasso is consistent in estimating the regression coefficients but, in general, it does not have the oracle property. See Luo and Chen (2014).

In classical linear regression where $n > p$, various variable selection methods have been devel-

oped. There is the stepwise regression which is either forward, backward or both. Also there is a best subset selection which is designed to select one model among 2^p possible candidate models. To determine when the algorithm terminates, various stopping criteria such as Akaike information criteria or AIC (Akaike, 1998) or Bayes information criteria or BIC (Schwarz, 1978) have been develop. For example in a stepwise forward selection method the algorithm terminates if the AIC of the current model is greater than the AIC of the previous model.

For dimension reduction in high dimensional data analysis is the concept of sure screening which is the property that all the important (relevant) variables survive after variable screening with probability tending to one. In their work Fan and Lv (2008) proposed a Sure Independence Screening (SIS) to reduce dimensionality from high to a relatively large scale d that is below the sample size and use methods such as the Dantzig selector, SCAD, LASSO or Adaptive LASSO for variable selection and estimation. To perform the SIS, all the variables in the data are centered and standardized. A componentwise regression, that is, $\omega = X^T y$, is performed to obtain a p-vector $\omega = (w_1, w_2, ..., w_p)^T$. For any $\gamma \in (0, 1)$, the vector ω is sorted in decreasing order and define a submodel $M_\gamma = \{1 \leq i \leq p : |\omega_i| \text{ is among the first } [\gamma n] \text{ largest of all}\}$, where $[\gamma n]$ denotes the integer part of γn. This shrinks the full model $\{1, ..., p\}$ down to a submodel M_γ with size $d = [\gamma n] < n$.

For large-p-small-n problems, the stopping criterion: AIC or BIC tend to select a model with many spurious covariates. This was observed by Broman and Speed (2002) and Storey et al. (2004) in their work on quantitative trait loci mapping (genome-wide inference of the relationship between genotype at various genomic locations and phenotype for a set of quantitative traits in terms of the number, genomic positions, effects, and interaction of QTL) using the BIC. In response to this challenge, Chen and Chen (2008) proposed and extended Bayesian information criterion family that is particularly suitable for model selection for large model spaces. It includes the original BIC as a special case and retains its simplicity. Under some mild conditions, these new criteria are shown to be consistent. The result is particularly useful even when the covariates are heavily collinear. The extended Bayesian information criteria do not require a data adaptive tuning

parameter procedure in order to be consistent, and hence are easy to use in applications.

A recent addition to the set of sequential approach is the Sequential LASSO Cum EBIC (SLasso) by Luo and Chen (2014). This method solves a sequence of partially penalized least squares problems. The features selected in an earlier step are not penalized in the subsequent steps. The EBIC is used as the stopping rule. For each s_{*k}, the EBIC of the model with features in s_{*k} is computed. The procedure continues, if the EBIC keeps decreasing. If the EBIC attains a minimum at step k^*, the procedure stops and the set s_{*k^*} is taken as the final selected set.

The partially penalized squares problem at each step is simply finding the feature j which maximizes $|x_j^\tau y|$, the absolute value of an unstandardized Pearson correlation coefficient. According to de Siqueira Santos et al. (2014), one major task in molecular biology is to understand the dependency among genes to model gene regulatory networks. Pearson's correlation is the most common method used to measure dependence between gene expression signals, but it works well only when data are linearly associated. For other types of association, such as non-linear or non-functional relationships, methods based on the concepts of rank correlation and information theory-based measures are more adequate than the Pearson's correlation. It is therefore important to consider other measures of association which could measure other types of associations other than linear association.

2.4.1 Conceptual Description of SLasso

Luo and Chen (2014) proposed a sequential Lasso (SLasso) algorithm for high dimensioal data for sparse features. SLasso solves a sequence of partially penalized least squares problems. The features selected in an earlier step are not penalized in the subsequent steps. Let the vectors $y = (y_1, y_2, ..., y_n)^\tau, x_j = (x_1 j, ..., x_n j)^\tau$, be standized such that they have length $\sqrt{n}$ and are orthogonal to the vector with all elements 1. Thus the intercept β_o can be omitted. At the initial step, SLasso minimizes the following penalized sum of squares:

$$l_1 = \left\| y - \sum_{j=1}^{p} \beta_j x_j \right\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|,$$

where $\|.\|$ is the L_2-norm, and λ_1 is the largest value of the penalty parameter such that at least one of the $\beta_j' s$ 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 the step k+1, SLasso minimizes the following partially penalized sum of squares:

$$l_{k+1} = \left\| y - \sum_{j=1}^{p} \beta_j x_j \right\|^2 + \lambda_{k+1} \sum_{j \notin s_{*k}}^{p} |\beta_j|,$$

where no penalty is imposed on the $\beta_j' s$ for $j \in s_{*k}$ and λ_{k+1} is the largest of the penalty parameter such that at least one of the β_j's, $j \notin s_{*k}$, will be estimated nonzero. The selected set is then updated to s_{*k+1}.

The EBIC is used as the stopping rule for each s_{*k}, the EBIC of the model with features in s_{*k} is computed. The procedure continues, if the EBIC keeps decreasing. If the EBIC attains a minimum at step k^*, the procedure stops and the set s_{*k^*} is taken as the final selected set.

2.4.2 SLasso Algorithm

- Initial Step: Standardize $y, x_j, j = 1, ..., p$ such that $y^\tau 1 = 0, x_j^\tau 1 = 0$ and $y^\tau y = n, x_j^\tau x_j = n$. Compute $x_j^\tau y$ for $j \in S$.

  Let

  $$s_{TEMP} = \{j : |x_j^\tau y| = \max_{j' \in S} |x_{j'}^\tau y|\}$$

  Let $s_{*1} = s_{TEMP}$, be the active set.

  Compute $I - H(s_{*1})$ and EBIC(s_{*1}). Where $H(s) = X(s)[X^\tau(s)X(s)]^{-1}X^\tau(s)$.

- General Step: For k $\geq$ 1, compute $\tilde{x}_j^\tau \tilde{y}$ for $j \in s_{*k}^c$, where $\tilde{y} = [I - H(s_{*k})]y, \tilde{x}_j = [I - H(s_{*k})]x_j$. Let

  $$s_{TEMP} = \{j : |\tilde{x}_j^\tau \tilde{y}| = \max_{j' \in S_{*k}^c} |\tilde{x}_{j'}^\tau \tilde{y}|\}$$

  Let $s_{*k+1} = s_{*k} \cup s_{TEMP}$. Compute EBIC$(s_{*k+1})$. If EBIC$(s_{*k+1})$ > EBIC(s_{*k}), stop; otherwise, compute $I - H(s_{*k+1})$ and continue.

- When the process stops, the parameters in the selected model are estimated by their least-square estimates.

The EBIC for $s_{*k}, k = 1, 2, ...,$ in the above algorithm is given by

$$\text{EBIC}(s_{*k}) = \text{nln}\left(\frac{\|[I - H(s_{*k})]y\|_2^2}{n}\right) + |s_{*k}|\text{ln(n)} + 2\left(1 - \frac{\text{ln(n)}}{\text{rln(p)}}\right)\text{ln}\binom{p}{|s_{*k}|}$$

where r a positive number slightly bigger than 2, say r = 2.1 by Luo and Chen (2014)

2.5 Extended Bayesian Information Criteria (EBIC)

In a high-dimensional setting, the traditional Bayes information criterion (BIC) is inappropriate for feature selection. It tends to select too many features that are not necessarily causal. Chen and Chen (2008) have recently proposed a family of extended Bayes information criteria (EBIC). In EBIC, models are classified according to the number of features they contain, and the prior probability assigned to a model is inversely proportional to the size of the model class to which the model belongs. Let $\{(y_i, x_i) : i = 1, ..., n\}$ be independent observations. Suppose that the conditional density function of y_i given x_i is $f(y_i|x_i, \theta)$, where $\theta \in \Theta \subset R^P$, P being a positive integer. The likelihood function of θ is given by

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

Let s be a subset of $\{1, ..., P\}$. Denote by $\theta(s)$ the parameter θ with those components outside s being set to 0 or some prespecified values. The BIC proposed by Schwarz (1978) selects the model that minimizes

$$BIC(s) = -2logL_n\{\hat{\theta}(s)\} + \nu(s)\text{log}(n)$$

where $\hat{\theta}(s)$ is the maximun likelihood estimator of $\theta(s)$ and $\nu(s)$ is the number of components in s. The extended BIC family is defined as

$$BIC_\gamma(s) = -2logL_n\{\hat{\theta}(s)\} + \nu(s)\text{log}n + 2\gamma\text{log}\tau(S_j).0 \leq \gamma \leq 1.$$

where $\hat{\theta}(s)$ is the maximum likelihood estimator of $\theta(s)$ given model s.

2.6 Derivation of EBIC

Chen and Chen (2008) derived the extended Bayes information criteria which has special cases as AIC and BIC.The extended Bayes information criteria is shown to be selection consistent. We follow the derivation as derived in Chen and Chen (2008).

Let $\{(y_i, x_i) : i = 1, 2, ..., n\}$ be independent observations. Suppose that the conditional density function of y_i given x_i is $f(y_i|x_i, \beta)$, where $\beta \in \Theta \subset R^{p_n}, p_n$ being a positive integer. The likelihood function of β is given by

$$L_n(\beta) = f(x; \beta) = \prod_{i=1}^{n} f(y_i|x_i, \beta).$$

Denote $Y = (y_1, y_2, ..., y_n)$. Let s be a subset of $\{1, 2, ..., p_n\}$. Denote by $\beta(s)$ the parameter β with those components outside s being set to 0. Let S be the model space under consideration, i.e, $S = \{s : s \subseteq \{1, 2, ..., p_n\}\}$, let $p(s)$ be the prior probability of model s. Assume that, given s, the prior density of $\beta(s)$ is $\pi(\beta(s))$. The posterior probability of s is obtained as

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

where $m(Y|s)$ is the likelihood of model s, given by

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

The BIC selects the model that minimizes

$$BIC(s) = -2logL_n\{\hat{\beta}(s)\} + |s|log(n)$$

where $\hat{\beta}(s)$ is the maximum likelihood estimator of $\beta(s)$ and $|s|$ is the number of components in s. When $\hat{\beta}(s)$ is $\sqrt{(}n)$ consistent, $-2ln(m(Y|s))$ has a Laplace approximation given by the BIC(s)

up to an additive constant. In the derivation of BIC, this constant $p(s)$ is taken as a constant over all s. With this constant prior, BIC favors models with larger numbers of features in small-n-large-p problems.

Assume that S is partitioned into $\cup_{j=1}^{p} S_j$, such that models within each S_j, have equal dimension . Let $\tau(S_j)$ be the size of S_j. Assign the prior distribution $P(S_j)$ proportional to $\tau^{\eta}(S_j)$ for some η between 0 and 1. For each $s \in S_j$, assign equal probability, $p(s|S_j) = 1/\tau(S_j)$, this is equivalent to $P(s)$ for $s \in S_J$ proportional to $\tau^{-\gamma}(S_j)$ where $\gamma = 1 - \eta$. This extended BIC family is given by

$$EBIC_{\gamma}(s) = -2logL_n\{\hat{\beta}(s)\} + |s|log(n) + 2\gamma ln(\tau(S_{|s|})), 0 \leq \gamma \leq 1.$$

For details about the selection consistency of EBIC, the reader is encouraged to refer to SHAN (2012). The selection consistency of EBIC, was shown for a multiple linear regression model through simulations where different values of γ have been specified.

2.7 Energy Distance

Székely et al. (2007) proposed the energy distance between probability distributions. Suppose that $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ are two random vectors with $\mathbb{E}\|X\| < \infty$, and $\mathbb{E}\|X\| < \infty$. Let F and G be the cumulative distributions (CDF) of X and Y, respectively. The energy distance between F and G, denoted by $D^2(F, G)$, is defined as the square root of

$$D^2(F, G) := 2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\|, \tag{2.7.1}$$

where $\|.\|$ denotes the Euclidean norm of its argument, $\mathbb{E}$ denotes expected value, and (X', Y') is an independent copy of (X, Y). Note that the right-hand-side term of equation (2.7.1) is always non-negative.

For further details and a brief history of the energy distance , see the article Rizzo and Székely (2016).

In Székely et al. (2007), it is proved that the energy distance $D^2(F,G) = 0$ if and only if F = G, so it characterizes equality of distributions. According to *Wikipedia*, the energy distance for statistical applications was first introduced in 1985 by Professor Gabor J. Szekely, who proved that for real-valued random variables $D^2(F,G)$ is exactly twice Harald Cramers distance.

2.7.1 Energy Distance Covariance

The energy distance covariance between random vectors X and Y with finite first moments is the nonnegative number $\nu(X,Y)$ defined as follows (Székely et al., 2007):

$$\nu^2(X,Y) = \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|^2.$$

Similarly, distance variance is defined as the square root of

$$\nu^2(X) = \nu^2(X,X) = \|f_{X,X}(t,s) - f_X(t)f_X(s)\|^2.$$

For samples $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_m$ from X and Y, respectively. Let $A = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|x_i - y_j\|$, $B = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|x_i - x_j\|$, $C = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|y_i - y_j\|$ where A, B and C are averages of pairwise distance.

The formula for the sample distance covariance statistic is the square root of

$$V_n^2(X,Y) = \frac{1}{n^2}\sum_{i,j=1}^{n}\hat{A}_{ij}\hat{B}_{ij}$$

where $\hat{A}$ and $\hat{B}$ are the double-centered distance matrices of the X sample and the Y sample, respectively, and the subscript ij denotes the entry in the $i-th$ row and $j-th$ column. The double-centered distance matrices are computed as in classical multidimensional scaling. Given a random sample $(x,y) = \{(x_i, y_j) : i = 1, ..., n\}$ from the joint distribution of random vectors X in $\mathbb{R}^p$ and Y in $\mathbb{R}^q$, compute the Euclidean distance matrix $(a_{ij}) = \|x_i - x_j\|$ for the X sample and

$(b_{ij}) = \|y_i - y_j\|$ for the Y sample. The $ij - th$ entry of $\hat{A}$ is

$$\hat{A}_{ij} = a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}, \;\; i, j = 1, ..., n,$$

where $\bar{a}_{i.} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}, \quad \bar{a}_{.j} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}, \quad \bar{a}_{..} = \frac{1}{n^2}\sum_{j=1}^{n} a_{ij},$

Similarly, the ij-th entry of $\hat{B}$ is

$$\hat{B}_{ij} = b_{ij} - \bar{b}_{i.} - \bar{b}_{.j} + \bar{b}_{..}, \;\; i, j = 1, ..., n,$$

The sample distance variance is

$$V_n^2(X) = V_n^2(X, X) = \frac{1}{n^2}\sum_{i,j=1}^{n} \hat{A}_{ij}^2.$$

The distance covariance statistic is always non-negative, and $V_n^2(X) = 0$ only if all of the sample observations are identical. Furthermore, if $E|X|_p < \infty$ and $E|Y| < \infty$,, then almost surely,

$$\lim_{n\to\infty} V_n^2(X, Y) = V^2(X, Y).$$

2.7.2 Energy Distance Correlation

The distance correlation between random vectors X and Y with finite first moments is the nonnegative number $\mathcal{R}(X, Y)$ defined by

$$\mathcal{R}(X, Y) = \begin{cases} \frac{\nu^2(X,Y)}{\sqrt{\nu^2(X)\nu^2(Y)}}, & \nu^2(X)\nu^2(Y) > 0 \\ 0, & \nu^2(X)\nu^2(Y) = 0 \end{cases}$$

The sample distance correlation is defined by

$$\mathcal{R}_n(X, Y) = \begin{cases} \frac{\nu_n^2(X,Y)}{\sqrt{\nu_n^2(X)\nu_n^2(Y)}}, & \nu_n^2(X)\nu_n^2(Y) > 0 \\ 0, & \nu_n^2(X)\nu_n^2(Y) = 0. \end{cases}$$

It should be noted that $\mathcal{R}_n(X, Y)$ is the Distance Correlation with the empirical distributions of the observations.

Some basic properties of Distance Correlation are as follows:

(i) $0 \leq R_n(X, Y) \leq 1$,

(ii) If $E(|X|_p + |Y|_q) < \infty$, then $R(X, Y) = 0$ if and only if X and Y are independent.

(iii) If $R_n(X, Y) = 1$, then there exist a vector a, a nonzero real number b and an orthogonal matrix C such that $Y = a + bXC$.

For further details on the energy distance correlation, see the article, (Rizzo and Székely, 2016).

CHAPTER 3 METHODOLOGY

3.1 Introduction

In this chapter we perform some simulations and propose a method for sequential feature selection. We first check the performance of distance correlation in measuring linear relationships in comparison with the Pearson Correlation measure.

Secondly, we discuss various nonlinear relationships and compare through simulations the various methods for measuring nonlinear relationships and also present a result by Gorfine et al. (2012) on a power study for three correlation measures.

Thirdly, we examine how Adaptive Lasso and SCAD performed after the features are screened by a sure screening method using distance correlation by Li et al. (2012).

In the fourth place, we would propose a sequential variable selection method which we title "Energy distance correlation with EBIC" (Edc+EBIC) and attempt to give a theoretical prove. We also establish through some simulations the selection consistency of Edc+EBIC.

3.2 Correlation Comparisons

In this section we studied the performance of distance correlation in measuring linear and nonlinear relationships between two variables.

3.2.1 Linear Relationships

In this subsection we compare the performance of the Pearson correlation coefficient and the distance correlation in measuring the linear relationship between two variables which are linearly associated. We run a simulation with a random (standardized) normal data of size $n = 100$ which is generated 100 times for population correlations of $-0.8, -0.6, -0.4, 0.0, 0.4, 0.6$ and 0.8. We compare the performance of these two methods using their mean, standard deviations and quantiles at 2.5% and 97.5%

In Table 3.1 we present the results of this simulation. From the comparison we would observe that the negative population values were estimated as positive by the distance correlation, this is

because the distance correlation is defined between 0 and 1. The Pearson correlation was able to measure perfectly the linear association as expected. The distance correlation estimates were also good with small standard deviations. The $[2.5\%, 97.5\%]$ quantiles showed that many of the estimates were close to the population values. Thus the distance correlation is appropriate for measuring the strength of linear associations.

Table 3.1 Comparison of Pearson correlation coefficient and Distance correlation in measuring linear association using their mean, standard deviations and quantiles at 2.5% and 97.5%

ρ	Pearson Mean(SD)	Distance Correlation Mean(SD)	Distance Correlation $[2.5\%, 97.5\%]$ quantile
-0.8	-0.800(0.000)	0.762(0.014)	[0.740, 0.789]
-0.6	-0.600(0.000)	0.567(0.025)	[0.523, 0.611]
-0.4	-0.400(0.000)	0.386(0.023)	[0.335, 0.431]
0.0	0.000(0.000)	0.152(0.023)	[0.118, 0.198]
0.4	0.400(0.000)	0.386(0.025)	[0.344, 0.439]
0.6	0.600(0.000)	0.563(0.025)	[0.515, 0.611]
0.8	0.800(0.000)	0.765(0.018)	[0.735, 0.802]

3.2.2 Non-linear Relationships

Aside the linear relationship that could exist between two random variables X and Y, there are several non-linear relationships that could exist. In Figure 3.3 we present shapes of seven nonlinear relationships namely diamond, trapezoid, wave, quadratic, cluster, circle and cross (X). These shapes have two noise levels and are adapted from Clark (2013).

In detecting these non-linear associations, Reshef et al. (2011) proposed a measure which they referred to as a novel measure of dependence - the maximal information coefficient (MIC) aimed to capture a wide range of associations between pairs of variables and a statistical test for independence based on MIC. However in a simple power comparison by Gorfine et al. (2012), they showed that the conclusions by Reshef et al. (2011) about the performance of MIC were wrong. Also in a comment by Simon and Tibshirani (2014) about the MIC by Reshef et al. (2011), they compared the power of Pearson correlation, Distance correlation and MIC. We present the graph

Less Noisy Noisier

Figure 3.3 Non-linear relationships based on noise levels. Adapted from Clark (2013).

of their power comparison in Figure 3.4. In this power study they consider a linear, quadratic, cubic, sine: period 1/8, sine: period 1/2, $x^{1/4}$, circle and step function. The power was estimated via 500 simulations. MIC has lower power than distance correlation in every case except the somewhat pathological high-frequency sine wave.They concluded that the MIC has serious power deficiences and when used for large-scale exploratory analysis, it will produce too many false positives, and thus the distance correlation measure of Székely et al. (2009) is a more powerful technique that is simple, easy to compute and should be considered for general use.

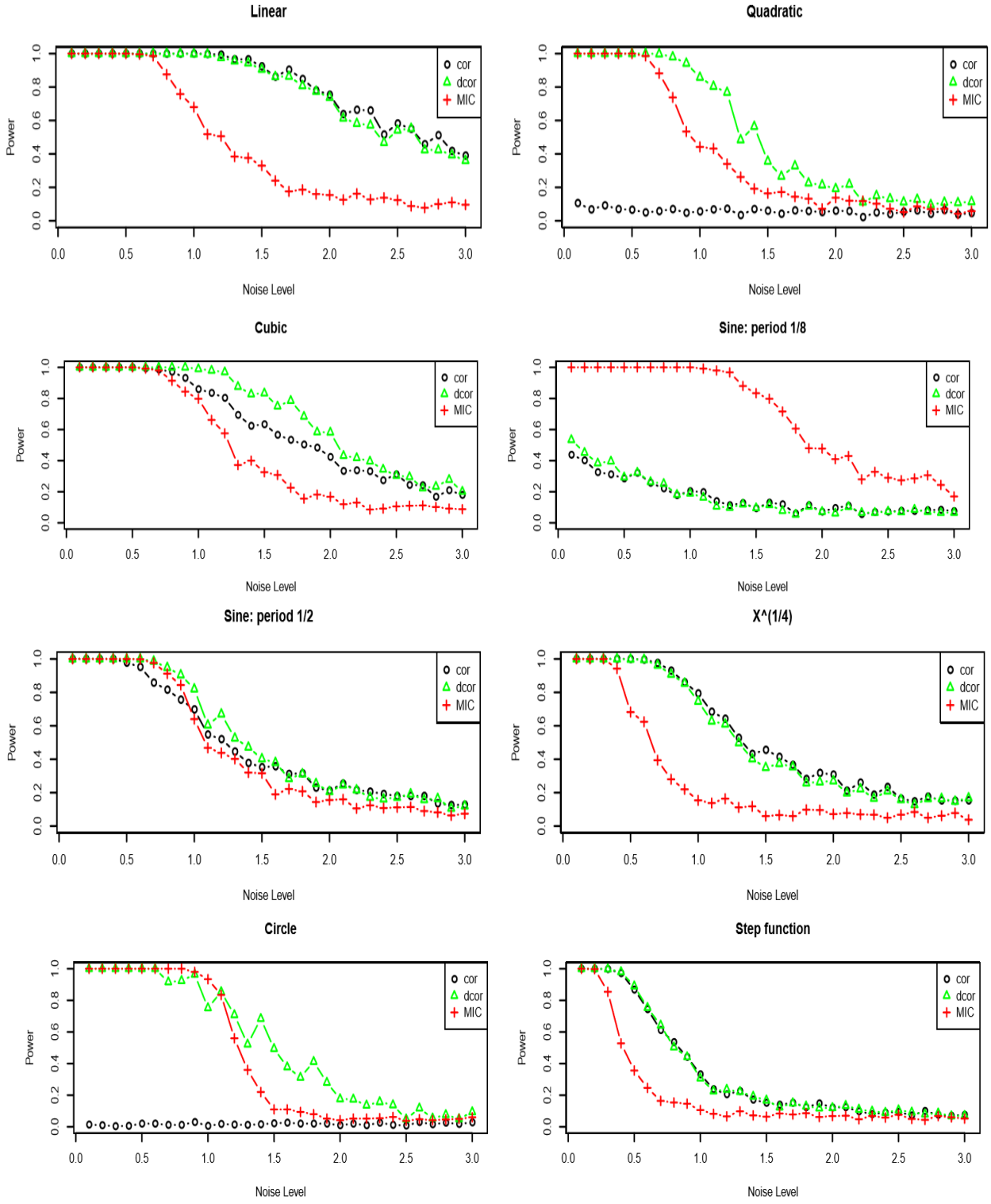

Figure 3.4 Power study of cor, dcor and MIC based on noise level. Adapted from Gorfine et al. (2012)

3.3 Sure Screening Using Energy Distance Correlation

In this section we examine the performance of a sure independence screening method introduced by Li et al. (2012) called DC-SIS. This is similar to the Sure Independence screening (SIS) introduced by Fan and Lv (2008).

In SIS, they perform a componentwise regression between each predictor and the response and select the first $n - 1$ or $[n/log(n)]$ predictors with the largest estimates. Performing a componentwise regression is equivalent to finding the Pearson correlation between the response and each predictor when the two variables are standardized. Hence in DC-SIS, they replaced the Pearson correlation with the distance correlation.

We examine the performance of DC-SIS through a simulation study. Our interest is to observe on average, the model size selected by SCAD or Adaptive Lasso if we screened the data first using DC-SIS. We present two simulation set-ups. For each simulation we generated two hundred datasets and for each data we run SCAD, Adaptive Lasso (ALasso), DC-SIS + SCAD, DC-SIS + ALasso and found the average model size and the standard deviation.

3.3.1 Simulation I: "Independent" Features

We adapt the simulation setup from Fan and Lv (2008). Data is simulated from the linear model 1.1.1 with i.i.d. standard Gaussian predictors and Gaussian noise with standard deviation $\sigma = 1.5$. We considered two such models with $(n, p) = (200, 1000)$ and $(800, 3000)$, respectively. The sizes s of the true models, i.e., the numbers of nonzero coefficients, were chosen to be 8 and 14, respectively, and the nonzero components of the p-vectors β were randomly chosen as follows. We set $a = 4\log n/\sqrt{n}$ and $5\log n/\sqrt{n}$, respectively, and picked nonzero coefficients of the form $(-1)^u(a + |z|)$ for each model, where u was drawn from a Bernoulli distribution with parameter 0.4 and z was drawn from the standard Gaussian distribution. In particular, the L_2- norms $\|\beta\|$ of the two simulated models are 6.795 and 8.908, respectively. For each model we simulated 100 data sets. SCAD and Adaptive LASSO were employed to estimate the sparse $p-$vectors β,. For the screening using the energy distance correlation we chose $d = [n/\log n]$ features and applied

SCAD or Adaptive Lasso.

In Tables 3.2 and 3.3, we report the average selected model size and their standard deviations. We observe that applying the sure screening by distance correlation before either SCAD or Adaptive Lasso in all cases did not lead to significant difference in the average model size when SCAD and ALasso were applied directly to the data.This suggests that either applying distance correlation before SCAD or Adaptive Lasso did not yield the intended result and thus needs some improvement.

Table 3.2 Comparing Model size selected with or without screening for $n = 200, s = 8, p = 1000$

Methods	MSize(SD)
SCAD	12.87(7.292)
DC-SIS + SCAD	10.7(3.1575)
ALasso	25.24(9.0365)
DC-SIS + ALasso	11.74(4.419)

Table 3.3 Comparing Model size selected with or without screening for $n = 800, s = 14, p = 3000$

Methods	MSize (SD)
SCAD	16.62(2.78807)
DC-SIS + SCAD	16.69(3.5525)
ALasso	14.78(0.7860)
DC-SIS + ALasso	14.78(3.8522)

3.3.2 Simulation II: "Dependent" Features

In this second simulation, we used similar models to those in simulation I except that the predictors are now correlated with each other. Three models are considered, $(n, p, s) = (200, 1000, 5)$, $(200, 1000, 8)$, and $(800, 3000, 14)$, respectively, where s denotes the size of the true model, i.e., the

number of nonzero coefficients. The three $p-$ vectors β were generated in the same way as in simulation I. We set $(\sigma, a) = (1, 2\text{log}n/\sqrt{n}), (1.5, 4\text{log}n/\sqrt{n}), (2, 4\text{log}n/\sqrt{n})$, respectively. We introduced a power decay correlation structure, i.e., $p_{i,j} = 0.5^{|i-j|}$, for $i, j = 1, ..., p$. $s_o = \{1, ..., p_o\}$ between the predictors. Then we took $Z_{s+1}, ..., Z_p \sim N(\mathbf{0}, I_{p-s})$ and defined the remaining predictors as $X_i = Z_i + rX_{i-s}, i = s+1, ..., 2s$ and $X_i = Z_i + (1-r)X_1, i = 2s+1, ..., p$ with $r = 1 - 4\text{log}n/p, 1 - 5\text{log}n/p$ and $1 - 5\text{log}n/p$ respectively. For each model we simulated 100 data sets. SCAD and Adaptive LASSO were employed to estimate the sparse $p-$vectors β,. For the screening using the energy distance correlation we chose $d = [n/\text{log}n]$.

In Tables 3.4 to 3.6, we report the selected model size and the standard deviation. We observe that applying the Distance correlation sure independence screening followed by either SCAD or Adaptive Lasso didn't yield any significant difference in the average model size just as in simulation I.

Table 3.4 Comparing Model size selected with or without screening for $n = 200, p = 1000, s = 5$

Methods	MSize (SD)
SCAD	12.215(12.2560)
DC-SIS + SCAD	7.335(2.6109)
ALasso	44.485(13.8041)
DC-SIS + ALasso	8.21(2.5844)

Table 3.5 Comparing Model size selected with or without screening for $n = 200, p = 1000, s = 8$

Methods	MSize (SD)
SCAD	14.625(10.3324)
DC-SIS + SCAD	10.905(2.3158)
ALasso	19.95(3.2789)
DC-SIS + ALasso	12.36(3.9875)

Table 3.6 Comparing Model size selected with or without screening for $n = 800, p = 3000, s = 14$

Methods	MSize (SD)
SCAD	19.185(5.0146)
DC-SIS + SCAD	17.675(4.6038)
ALasso	38.125(5.6372)
DC-SIS + ALasso	31.845(13.1011)

3.4 Proposed Method: Energy Distance Correlation with EBIC

In this section we propose a sequential model selection method. Let $y_i, i = 1, ..., n$ be a continuous response variable and $x_j, j = 1, ..., p$ be an $n \times p$ data matrix. Let S be the index set of all predictors. Let $s_0 = \{j : \beta_j \neq 0, j = 1, ..., p\}$. For $s \subset S$, let $s^- = s^c \cap s_0$. If $s \subset s_0$ then s^- is the complement of s in s_0. Let $p_0 = |s_0|$ be the number of elements in the set s_0.

At the initial stage we standardize all the variables. Next we find the distance correlation between the response variable and each of the predictor variables - $\{\mathcal{R}(x_j, y)\ j = 1, ..., p.\}$. We then select the predictor (feature) which has the highest distance correlation with the response and store it in the active set s_{*1}.

Let $\mathcal{L}(s)$ be the linear space spanned by the columns of $X(s)$ and $H(s)$ its corresponding projection matrix, i.e, $H(s) = X(s)[X^\tau(s)X(s)]^{-1}X^\tau(s)$. Next we compute $I - H(s_{*1})$, EBIC(s_{*1}), $\tilde{y} = [I - H(s_{*k})]y$ and $\tilde{x}_j = [I - H(s_{*k})]x_j$. The variable $\tilde{y}$ is the unexplained part of y by $X(s_{*1})$. This gives $X(s_{*1})$ close to a zero chance of been selected in the subsequent steps.

For the general step where $k > 1$ we calculate $\{\mathcal{R}(\tilde{x}_j, \tilde{y})\ j = 1, ..., p.\}$ and update the active set to s_{*k+1} which is the union of all the previous selected variables and the current one. We then compute EBIC(s_{*k+1}) and compare it with EBIC(s_{*k}). The procedure stops if EBIC$(s_{*k+1}) >$ EBIC(s_{*k}). The selected variables which we call the relevant variables will be $X(s_{*k})$. We can then fit a linear regression model between the response y and the relevant variables. We wish to note that care must be taken in fitting this model because some of the predictors might be non-linearly related to y and thus some of the predictors may have to enter into the model in their quadratic or cubic form etc. Alternatively is to perform a box-cox transformation on the data before fitting the model.

3.4.1 Energy Distance Correlation with EBIC (Edc+EBIC) Algorithm

We adapt the algorithm for SLasso by Luo and Chen (2014). We replace the maximization of $\{|x_j^\tau y|\ j = 1, ..., p\}$ with maximization of $\{\mathcal{R}(x_j, y)\ j = 1, ..., p.\}$

- Initial Step: Standardize $y, x_j, j = 1, ..., p$ such that $y^\tau 1 = 0, x_j^\tau 1 = 0$ and $y^\tau y = n, x_j^\tau x_j = n$. Compute $\mathcal{R}(x_j, y)$ for $j \in S$.

  Let

  $$s_{TEMP} = \{j : \mathcal{R}(x_j, y) = \max_{j' \in S} \mathcal{R}(x_{j'}, y)\}$$

  Let $s_{*1} = s_{TEMP}$, be the active set.

  Compute $I - H(s_{*1})$ and EBIC(s_{*1}). Where $H(s) = X(s)[X^\tau(s)X(s)]^{-1}X^\tau(s)$.

- General Step: For k $\geq$ 1, compute $\mathcal{R}(\tilde{x}_j, \tilde{y})$ for $j \in s_{*k}^c$, where $\tilde{y} = [I - H(s_{*k})]y, \tilde{x}_j = [I - H(s_{*k})]x_j$. Let

  $$s_{TEMP} = \{j : \mathcal{R}(\tilde{x}_j, \tilde{y}) = \max_{j' \in S_{*k}^c} \mathcal{R}(\tilde{x}_j, \tilde{y})\}$$

  Let $s_{*k+1} = s_{*k} \cup s_{TEMP}$. Compute EBIC$(s_{*k+1})$. If EBIC$(s_{*k+1})$ $>$ EBIC(s_{*k}), stop; otherwise, compute $I - H(s_{*k+1})$ and continue.

- When the process stops, the parameters in the selected model are estimated by their least-square estimates.

The EBIC for $s_{*k}, k = 1, 2, ...,$ in the above algorithm is given by

$$\text{EBIC}(s_{*k}) = \text{nln}\left(\frac{\|[I - H(s_{*k})]y\|_2^2}{n}\right) + |s_{*k}|\text{ln(n)} + 2\left(1 - \frac{\text{ln(n)}}{\text{rln(p)}}\right)\text{ln}\binom{p}{|s_{*k}|}$$

where r a positive number slightly bigger than 2, say r = 2.1 Luo and Chen (2014)

3.5 Selection Consistency

We attempt to establish the large sample property for the Edc+EBIC. We will show that under regular conditions, the Edc+EBIC is selection consistent. The proof essentially follows the approach in Li et al. (2012). We proceed with the regularity conditions.

C1: Both x and y satisfy the subexponential tail probability uniformly in p. That is there exist a positive constant a_0 such that for all $0 < a \leq 2a_0$, $sup_p max_{1\leq k\leq p} E\{exp(a\|X_k\|_1^2)\} < \infty$ and $E\{exp(a\|y\|_q^2)\} < \infty$

C2: The minimum distance correlation of predictors on which y functionally depends satisfies $\min_{j\in s_0}\mathcal{R}(\tilde{x}_j, \tilde{y}) \geq 2cn^{-\kappa}$, for some constants $0 < c < 1$ and $0 \leq \kappa < 1/2$.

C3: Let S be the index set of all predictors. Let $s_0 = \{j : \beta_j \neq 0, j = 1, ..., p\}$ and $p_0 = |s_0|$ (p_0 is the number of elements in the set s_0). For $s \subset S$ let $s^- = s^c \cap s_0$. If $s \subset s_0$ then s^- is the complement of s in s_0. For $s \subset s_0$, $\max_{j\in s_0^c}\mathcal{R}(\tilde{x}_j, \tilde{y}) < q\max_{j\in s^-}\mathcal{R}(\tilde{x}_j, \tilde{y})$ for some $0 < q < 1$. Where $\tilde{y} = [I - H(s_{*k})]y$, $\tilde{x}_j = [I - H(s_{*k})]x_j$. For $k = 0$, s_{*0} is taken as the empty set $\emptyset$.

Theorem 3.5.1. *Suppose that conditions C1 - C3 hold. The Edc+EBIC is selection consistent in the sense that*

$$P(s_{*k^*} = s_{0n}) \to 1, \textit{as } n \to \infty,$$

where s_*k^* is the set of features selected at the k^{*th} step of Edc+EBIC such that $|s_*k^*| = p_{0n}$, s_{0n} is the set of relevant features and $p_{0n} = |s_{0n}|$.

Proof. Suppose that $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with cumulative distribution function (CDF) F and G, respectively where $\mathbb{E}\|X\| < \infty$, and $\mathbb{E}\|X\| < \infty$. The population distance correlation $\mathcal{R}(X, Y)$ is the square root of the standardized coefficient:

$$\mathcal{R}(X,Y) = \begin{cases} \frac{\nu^2(X,Y)}{\sqrt{\nu^2(X)\nu^2(Y)}}, & \nu^2(X)\nu^2(Y) > 0 \\ 0, & \nu^2(X)\nu^2(Y) = 0 \end{cases}$$

where $0 \leq \mathcal{R}(X, Y) \leq 1$. In the numerator is the distance covariance defined by Székely et al.

(2007), as

$$dcov^2(x,y) = S_1 + S_2 - 2S_3,$$

where $S_j, j = 1, 2, \text{and } 3$, are defined as:

$$\begin{aligned} S_1 =& \mathbb{E}\|X - X'\|\|Y - Y'\| \\ S_2 =& \mathbb{E}\|X - X'\|\mathbb{E}\|Y - Y'\| \\ S_3 =& \mathbb{E}\|X - X'\|\|Y - Y^{''}\| \end{aligned} \tag{3.5.2}$$

where $(X, Y), (X', Y')$, and $(X^{''}, Y^{''})$ are independently and identically distributed.

For a random sample $\{(x_i, y_i), i = 1, ..., n\}$ from (x, y), Székely et al. (2007) estimated S_1, S_2, S_3 as:

$$\begin{aligned} \hat{S}_1 =& \frac{1}{n^2}\sum_{k,l=1}^{n} |x_k - x_l|_p |y_k - y_l|_q \\ \hat{S}_2 =& \frac{1}{n^2}\sum_{k,l=1}^{n} |x_k - x_l|_p \frac{1}{n^2}\sum_{k,l=1}^{n} |y_k - y_l|_q \\ \hat{S}_3 =& \frac{1}{n^3}\sum_{k=1}^{n}\sum_{l,m=1}^{n} |x_k - x_l|_p |y_k - y_l|_q \end{aligned}$$

with sample, the distance covariance $\widehat{dcov}^2 = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3$.

The remaining part of the proof is to show that the energy distance correlation is uniformly consistent and has the sure screening property. The numerator and denominator of the energy distance correlation are similar so to show the uniform consistency of the energy distance correlation it suffices to show that both the numerator and the denominator are uniformly consistent.

The uniform consistency of the numerator, $\widehat{dcov}^2 = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3$ of the energy distance correlation between the random vectors (x, y) is shown by Li et al. (2012). However in the general step of the sequential algorithm for Edc+EBIC, the energy distance correlation is calculated between the residuals $\tilde{y} = [I - H(s_{*k})]y, \text{ and } \tilde{x}_j = [I - H(s_{*k})]x_j$ at each step of the algorithm. Thus to show the uniform consistency of Edc+EBIC it is equivalent to follow the proof by Li et al. (2012).

Also in Li et al. (2012) they showed that the energy distance correlation has the sure screening property. They showed that the energy distance is able to select a subset of the features which contains the relevant features. Their argument applies here because we used the energy distance correlation as well, thus the Edc+EBIC has the sure screening property.

Therefore the Edc+EBIC is selection consistent since it is uniformly consistent and has the sure screening property. The proof is complete. □

3.5.1 Simulation Study on Selection Consistency of Edc+EBIC

We conducted simulation studies to demonstrate how the selection consistency of Edc+EBIC works. For each simulation setup we record the following:

1. Model size (MSize), which is the total number of relevant variables selected.
2. Positive Discovery Rate (PDR); PDRn $= \frac{|s_{*k^*} \cap s_o|}{|s_o|}$
3. False Discovery Rate (FDR); FDRn $= \frac{|s_{*k^*} \cap s_o^c|}{|s_{*k^*}|}$

We consider the diverging pattern $(n, p, p_o) = (n, [5e^{n^{0.3}}], [4n^{0.16}])$ for $n = 100, 200, 300$ and, 500. The coefficients are generated as independent random variables distributed as $(-1)^u(4n^{-0.15} + |z|)$, where $u \sim PBernoulli(0.4)$ and z is a normal random variable with mean 0 and satisfies $P(|z| \geq 0.1) = 0.25$. The variance of the error term in the linear model is determined by

$$h = \frac{\beta^\tau \Sigma \beta}{\beta^\tau \Sigma \beta + \sigma^2} = 0.8$$

where Σ is the variance-covariance matrix of relevant features. Two settings of the covariance structure for the design matrix X are considered.

1. All the p features are generated as i.i.d. standard normal random variables.

   In Table 3.7 we show the selection consistency of Edc+EBIC for independent features. Considering the diverging pattern, where as the sample size increases the number of features also increases, we observed that Edc+EBIC selected on average model sizes very close to

the expected number of relevant features and with small standard deviations. The positive discovery rate was 100% meaning that Edc+EBIC always selected all the relevant variables. Edc+EBIC is also seen to have reducing false discovery rate as the sample size increases and does at smaller standard deviations.

Table 3.7 Selection consistency of Edc+EBIC for independent features.

n	p_{on}	p	MSize	PDR	FDR
100	8	276	8.420(0.712)	1.000(0.000)	0.045(0.071)
200	9	682	9.270(0.591)	1.000(0.000)	0.026(0.054)
300	10	1277	10.185(0.460)	1.000(0.000)	0.016(0.040)
500	11	3181	11.130(0.352)	1.000(0.000)	0.011(0.029)

2. The features have a power decay correlation structure, i.e., $p_{i,j} = 0.5^{|i-j|}$, for $i, j = 1, ..., p$. $s_o = \{1, ..., p_o\}$. Where s_o is the number of relevant variables and p_o are the relevant variables.

   In Table 3.8 we show the selection consistency of Edc+EBIC for power decay correlated features. We considered the diverging pattern. We observed that Edc+EBIC selected on average model sizes very close to the expected number of relevant features and with standard deviations a little higher than those in Table 3.7. It did not achieve a perfect positive discovery rate for all the samples. Edc+EBIC is also seen to have reducing false discovery rate as the sample size increases and does so at smaller standard deviations.

Table 3.8 Selection consistency of Edc+EBIC for power decay correlated features.

n	p_{on}	p	MSize	PDR	FDR
100	8	276	8.490(0.763)	1.000(0.000)	0.051(0.076)
200	9	682	9.19(0.798)	0.991(0.0652)	0.026(0.050)
300	10	1277	10.16(0.553)	0.996(0.032)	0.018(0.040)
500	11	3181	11.125(0.374)	1.000(0.000)	0.010(0.030)

CHAPTER 4 SIMULATION STUDIES AND DATA ANALYSIS

4.1 Introduction

In this section we present the results of our simulations on comparing our proposed method with other variable selection methods. We also present two real-life data examples. We considered two setups in our simulation study as used in Luo and Chen (2014). In setup 1, four settings of the covariance structure for the design matrix X namely GA1, GA2, GA3 and GA5 were considered. In setup 2, three settings of the covariance structure for the design matrix X namely GB1, GB2 and GB3 were considered.

In each setup and in the two real-life data examples we compared the performance of Adaptive Lasso (ALasso), SCAD, SIS+SCAD, SLasso and the Energy Distance correlation with EBIC (Edc+EBIC) based on the model size (MSize), positive discovery rate (PDR), PDR $= \frac{|s_{*k^*} \cap s_o|}{|s_o|}$ and false discovery rate, FDR $= \frac{|s_{*k^*} \cap s_0^c|}{|s_{*k^*}|}$ averaged over 200 and 500 simulations respectively. The R packages glmnet, ncvreg and SIS were used for the computation of ALasso, SCAD, and SIS+SCAD respectively.

4.2 Simulation Results for Setup 1

We consider the diverging pattern $(n, p, p_o) = (n, [5e^{n^{0.3}}]], [4n^{0.16})$ for $n = 100$. The coefficients are generated as independent random variables distributed as $(-1)^u(4n^{-0.15} + |z|)$, where $u \sim PBernoulli(0.4)$ and z is a normal random variable with mean 0 and satisfies $P(|z| \geq 0.1) = 0.25$. The variance of the error term in the linear model is determined by

$$h = \frac{\beta^\tau \Sigma \beta}{\beta^\tau \Sigma \beta + \sigma^2} = 0.8$$

where Σ is the variance-covariance matrix of relevant features. Four settings of the covariance structure for the design matrix X are considered. They are named GA1,GA2,GA3 and GA5.The

response variable is simulated from the sparse high-dimensional regression (SHR) model

$$y_i = \beta_0 + \Sigma_{j=1}^{p} \beta_j x_{ij} + \in_i, i = 1, ..., n$$

According to our simulation setup the number of relevant random variables s_0 is 8, thus we expect a well performing method to select the 8 relevant random variables. The number of observations $n = 100$ and the total number of predictors is 276 where 8 are the relevant predictors.

GA1. All the p features are generated as i.i.d. standard normal random variables.

In Table 4.1 the best performing models for GA1 are SIS + SCAD, SLasso and EdcEBIC. All three methods selected a little over the 8 relevant predictors on average. The SLasso had the highest positive discovery rate but with a slightly high false discovery rate compared to EdcEBIC. ALasso and SCAD had perfect positive discovery rate but recorded very high model size and very high false discovery rate.

Table 4.1 We compare the methods using PDR, FDR, and model size (MSize) averaged over 200 simulation replications. The relevant predictors are 8 and the sample size is 100.The standard deviations are in parenthesis.

Setting	Methods	MSize	PDR	FDR
GA1	ALasso	34.105(13.95)	1.000(0.000)	0.721(0.120)
	SCAD	25.735(5.020)	1.000(0.000)	0.676(0.065)
	SIS + SCAD	8.100(1.790)	0.866(0.239)	0.157(0.167)
	SLasso	8.565(0.848)	1.000(0.000)	0.058(0.081)
	EdcEBIC	8.365(1.375)	0.978(0.125)	0.056(0.085)

GA2. The features have a power decay correlation structure, i.e., $p_{i,j} = 0.5^{|i-j|}$, for $i, j = 1, ..., p$. $s_o = \{1, ..., p_o\}$. Where s_o is the number of relevant variables and p_o are the relevant variables.

In Table 4.2 under GA2, the best performing modesls are SIS + SCAD, SLasso and EdcEBIC. The SIS+SCAD performed better in terms of the MSize and PDR compared to SLasso and EdcEBIC. The EdcEBIC performed better in terms of a lower FDR. We reckon that it was difficult for

SIS+SCAD, SLasso and EdcEBIC to select on average the 8 relevant variables and achieve a high PDR when there is a power decay correlation structure among the variables.

Table 4.2 We compare the methods using PDR, FDR, and model size (MSize) averaged over 200 simulation replications. The relevant predictors are 8 and the sample size is 100.The standard deviations are in parenthesis.

Setting	Methods	MSize	PDR	FDR
GA2	ALasso	34.455(11.095)	1.000(0.000)	0.754(0.108)
	SCAD	25.650(6.720)	0.876(0.141)	0.709(0.075)
	SIS + SCAD	7.335(1.740)	0.813(0.182)	0.103(0.107)
	SLasso	6.055(1.725)	0.688(0.185)	0.080(0.104)
	EdcEBIC	6.075(1.713)	0.717(0.195)	0.050(0.082)

GA3. The features $X_1, ..., X_p$ are determined as follows. Let $Z_1, ..., Z_p$ and $W_1, ..., W_{p_o}$ be i.i.d standard normal random variables. Then

$$x_j = \frac{Z_j + W_j}{\sqrt{2}}, \text{for} j \in s_0;$$

$x_j = \frac{Z_j + \Sigma_{k \in s_o} Z_k}{\sqrt{1 + p_0}}$ for $j \notin s_0$.

In Table 4.3 under GA3, there was a very steep competition between SIS + SCAD, SLasso and EdcEBIC. SLasso however selected a little more variables on average compared to SIS+SCAD and EdcEBIC. The SLasso had a perfect PDR while EdcEBIC had the smallest FDR and also with a smaller standard deviation.

GA5. The set s_o is taken as $\{1, 2, ..., p_o\}$. The features in s_o has the power decay correlation $p_{ij} = 0.5^{|i-j|}$. For $j \notin s_o$, x_j is generated as:

$$x_j = \in_j + \frac{\Sigma_{k \in s_o} X_k}{p_o},$$

Table 4.3 We compare the methods using PDR, FDR, and model size (MSize) averaged over 200 simulation replications. The relevant predictors are 8 and the sample size is 100.The standard deviations are in parenthesis.

Setting	Methods	MSize	PDR	FDR
GA3	ALasso	14.710(3.847)	1.000(0.000)	0.423(0.131)
	SCAD	26.27(5.244)	1.000(0.000)	0.680(0.070)
	SIS + SCAD	8.165(1.160)	0.951(0.113)	0.062(0.078)
	SLasso	8.625(1.005)	1.000(0.000)	0.062(0.089)
	EdcEBIC	8.265(1.373)	0.976(0.132)	0.048(0.074)

where $\in_j' s$ are i.i.d, with distribution $N(0, 0.08)$. The variance 0.08 is chosen such that the second term, which is correlated with relevant features, dominates the variance of x_j.

In Table 4.4 under GA5,, the three best methods, SIS+SCAD, SLasso and EdcEBIC, did not perform well in the model size and positive discovery rate.However the SLasso performed better in terms of MSize and PDR. The EdcEBIC performed better by recording the smallest FDR.

Table 4.4 We compare the methods using PDR, FDR, and model size (MSize) averaged over 200 simulation replications. The relevant predictors are 8 and the sample size is 100.The standard deviations are in parenthesis.

Setting	Methods	MSize	PDR	FDR
GA5	ALasso	23.845(7.005)	0.964(0.057)	0.652(0.092)
	SCAD	24.070(6.147)	0.997(0.020)	0.642(0.102)
	SIS + SCAD	7.605(2.020)	0.832(0.245)	0.127(0.141)
	SLasso	7.650(2.182)	0.856(0.217)	0.089(0.113)
	EdcEBIC	7.180(2.453)	0.842(0.270)	0.050(0.081)

4.2.1 Simulation Results for Sample Size 200 for Setup 1 with 8 Relevant Predictors

In Table 4.5, we report the simulation results under the conditions for GA1, GA2, GA3 and GA5 except that we increased the sample size to 200. We observe that under all the setups, EdcEBIC improved in the average model size, PDR and FDR. This also shows the selection consistency of EdcEBIC under fix number of predictors with increased sample size.

Table 4.5 We compare the methods using PDR, FDR, and model size (MSize) averaged over 200 simulation replications. The relevant predictors are 8 and the sample size is 200.The standard deviations are in parenthesis.

Setting	Methods	MSize	PDR	FDR
GA1	ALasso	27.670(12.996)	1.000(0.000)	0.638(0.180)
	SCAD	17.035(7.746)	1.000(0.000)	0.454(0.168)
	SIS + SCAD	9.215(1.507)	1.000(0.000)	0.112(0.123)
	SLasso	8.710(0.0.944)	1.000(0.000)	0.072(0.088)
	EdcEBIC	8.42(0.712)	1.000(0.000)	0.045(0.071)
GA2	ALasso	27.92(9.686)	1.000(0.000)	0.675(0.124)
	SCAD	15.11(6.241)	1.000(0.000)	0.397(0.171)
	SIS + SCAD	9.16(1.509)	1.000(0.000)	0.108(0.171)
	SLasso	8.72(0.947)	1.000(0.000)	0.073(0.089)
	EdcEBIC	8.49(0.763)	1.000(0.000)	0.051(0.076)
GA3	ALasso	27.115(12.867)	1.000(0.000)	0.632(0.177)
	SCAD	16.245(7.770)	1.000(0.000)	0.434(0.162)
	SIS + SCAD	9.22(1.617)	1.000(0.000)	0.110(0.128)
	SLasso	8.70(0.857)	1.000(0.000)	0.072(0.084)
	EdcEBIC	8.47(0.694)	1.000(0.000)	0.050(0.071)
GA5	ALasso	38.95(8.308)	0.939(0.075)	0.797(0.054)
	SCAD	19.075(7.427)	1.000(0.000)	0.520(0.159)
	SIS + SCAD	9.975(1.858)	1.000(0.000)	0.174(0.132)
	SLasso	8.765(1.125)	0.989(0.061)	0.087(0.094)
	EdcEBIC	8.44(0.768)	0.998(0.025)	0.048(0.073)

4.3 Simulation Results under Setup II

In this section we considered three different covariance structures named GB1, GB2 and GB3 for the features (predictors). We also increase the signal to noise ratio by increasing the value of the expected predictors.

GB1. The setting is taken from Luo and Chen (2014). All the features have constant pair-wise correction $p_{ij} = 0.5$. $(n, p, p_0) = (100, 200, 15)$. $\sigma = 1.5$. The coefficients of the relevant features are specified as $|\beta_j| = 2.5$ for $1 \leq j \leq 5, 1.5$ for $6 \leq j \leq 10, 0.5$ for $11 \leq j \leq 15$. The signs of the coefficients are determined as $(-1)^{u_i}$ where the u_i's are i.i.d. Bernoulli random variables with probability of success $p = 0.5$.

In Table 4.6 we present the result for the setting where all the features are pair-wise correlated. We observe that SCAD, SLasso and EdcEBIC performed better. SLasso had the highest PDR

while EdcEBIC had the least FDR. Thus when all the features are pair-wise correlated, EdcEBIC still performs well.

Table 4.6 We compare the methods using PDR, FDR, and model size (MSize) averaged over 500 simulation replications. The relevant predictors are 15 and the sample size is 100.The standard deviations are in parenthesis.

Setting	Methods	MSize	PDR	FDR
GB1	ALasso	23.32(3.018)	0.766(0.062)	0.501(0.066)
	SCAD	14.08(1.644)	0.853(0.065)	0.085(0.068)
	SIS + SCAD	10.656(1.688)	0.694(0.112)	0.025(0.067)
	SLasso	14.916(2.194)	0.893(0.081)	0.092(0.089)
	EdcEBIC	14.094(2.035)	0.869(0.088)	0.067(0.076)

GB2. The setting is taken from Luo and Chen (2014). It is the same as in GB1 that $(n, p, p_0) = (100, 200, 15)$ and $\sigma = 1.5$. But the covariance structure of the features is specified such that the partially orthogonality condition Huang et al. (2008) is satisfied. Specifically, while s_0 is taken as $\{1, ...5, 11, ..., 15, 21, ..., 25\}$ the correlations are specified as $\rho_{ij} = 0.5^{|i-j|}$ for $1 \leq i \leq 215$ and $1 \leq j \leq 215$. The coefficients are specified as $|\beta| = 2.5$ for $1 \leq j \leq 5, 1.5$ for $10 \leq j \leq 15, 0.5$ for $21 \leq j \leq 25$. The signs of the coefficients are determined in the same way as in GB1.

In Table 4.7 we present the result for the setting where all the features have a power decay correlation structure. ALasso and SCAD are seen to have some challege is selecting the right number of expected relevant features. They also recorded poor PDR and FDR values. SLasso aside producing a better FDR compared to EdcEBIC didn't do well with model size and PDR. The average model size selected by EdcEBIC was closest to the expected number of relevant features (15). EdcEBIC had the highest PDR but didn't produce a good FDR.

GB3. The setting is taken from Luo and Chen (2014). $(n, p, p_0) = (100, 1000, 10)$ and $\sigma = 1$. The relevant features are generated as i.i.d. standard normal variables with coefficients

Table 4.7 We compare the methods using PDR, FDR, and model size (MSize) averaged over 500 simulation replications. The relevant predictors are 15 and the sample size is 100.The standard deviations are in parenthesis.

Setting	Methods	MSize	PDR	FDR
GB2	ALasso	40.474(11.7331)	0.447(0.0858)	0.710(0.0605)
	SCAD	20.966(7.6121)	0.517(0.0614)	0.315(0.1896)
	SIS + SCAD	10.314(1.0797)	0.403(0.0427)	0.042(0.0712)
	SLasso	13.65(2.038)	0.499(0.052)	0.077(0.081)
	EdcEBIC	14.006(1.657)	0.67(0.014)	0.273(0.0785)

$(3, 3.75, 4.5, 5.25, 6, 6.75, 7.5, 8.25, 9, 9.75)$. The irrelevant features are generated as

$$x_j = 0.25Z_j + \sqrt{0.75}\sum_{k\in s_0} X_k, j \notin s_0,$$

where $Z_j's$ are i.i.d. standard normal and independent from the relevant features.

In Table 4.8 under this setting we observe that EdcEBIC performed better than the other methods. The average model size selected by EdcEBIC was very close to the expected number of relevant features (10). EdcEBIC also recorded the high PDR and the least FDR.

Table 4.8 We compare the methods using PDR, FDR, and model size (MSize) averaged over 500 simulation replications. The relevant predictors are 10 and the sample size is 100.The standard deviations are in parenthesis.

Setting	Methods	MSize	PDR	FDR
GB3	ALasso	22.464(2.4414)	1.000(0.000)	0.5495(0.0498)
	SCAD	11.000(0.000)	1.000(0.000)	0.091(0.000)
	SIS + SCAD	9.964(0.6897)	0.992(0.0764)	0.107(0.050)
	SLasso	10.182(0.475)	0.667(0.006)	0.015(0.039)
	EdcEBIC	10.158(0.440)	1.000(0.000)	0.0139(0.038)

4.4 Real Data Examples

In this section we apply our method to two real-life datasets and compare our results with other researches which used these data for a similar purpose of variable selection.

Example 1: The description of the data as stated in Luo and Chen (2014) is as follows: the data, which were reported in Scheetz et al. (2006) , consist of the expression levels of over 31,042 different probes from 120 F2 male rats generated from an intercross experiment. A cross of SR/JrHsd male rats and SHRSP female rats was performed to generate F1 and the F1 rats were intercrossed to generate the F2 rats. The probes that were not expressed in the eye or that lacked sufficient variation were excluded. A probe was considered expressed if its maximum expression value observed among the 120 F2 rats was greater than the 25th percentile of the entire set of RMA (robust multichip averaging) expression values. A probe was considered "sufficiently variable" if it exhibited at least two-fold variation in expression level among the 120 F2 rats. A total of 18,976 probes that met these criteria were retained. Among the 18,976 probes, there is one, 1389163_at, from gene TRIM32. This gene was found to cause Bardet-Biedl syndrome (Bardet-Biedl syndrome (BBS) is a genetic condition that impacts multiple body systems.

The Bardet-Biedl syndrome is classically defined by six features. Patients with BBS can experience problems with obesity, specifically with fat deposition along the abdomen. They often also suffer from intellectual impairments. Commonly, the kidneys, eyes and function of the genitalia will be compromised. People with BBS may also be born with an extra digit on the hands. The severity of BBS varies greatly even among individuals within the same family).

Of interest is to find the probes among the remaining 18,975 probes that are most related to TRIM32. The response variable is the expression level of probe 1389163_at. The features are the expression levels of the remaining 18,975 probes. Of the 18,975 probes, the top 3000 probes with the largest variances were considered. The expression levels are standardized to have mean 0 and standard deviation 1.

In our analysis of the data, for each of the 100 replications we selected a random sample of size 100 from the 120 rats and apply the Edc+EBIC to it. From these 100 replications we selected the distinct probes that our method yielded.

In Table 4.9, we present the two probes selected by our method Edc+EBIC. In other to compare our result with other researches which used this data we adapt the result presented in Luo and Chen

(2014). We observe that SLASSO + EBIC selected two probes just as our method, Edc + EBIC. However these probes were different.

Table 4.9 Rat Data: The Gene Probes Selected by All Considered Methods.

Methods	ProbesID
ALasso+CV	1387060_at, 1388538_at, 1380070_at, 1370052_at, 1382452_at, 1379079_at, 1397489_at, 1374131_at, 1383110_at, 1389584_at, 1392692_at, 1379971_at 1385687_at, 1369353_at, 1374106_at, 1383673_at, 1379495_at, 1383749_at 1382835_at, 1395415_at, 1383996_at.
SCAD+CV	1394689_at, 1370434_a_at, 1375724_at, 1378765_at, 1375139_at, 1388538_at 1370052_at, 1382452_at, 1377781_at, 1383841_at, 1380311_at, 1379460_at, 1385921_at, 1384886_at, 1384136_at, 1387111_at, 1390789_at, 1376693_at, 1389584_at, 1389231_at, 1390788_a_at, 1367741_at, 1374106_at, 1387455_a_at, 1383749_at, 1379803_at, 1383996_at, 1382633_at
SIS+SCAD	1377546_at, 1396809_at, 1381430_at, 1393543_at, 1372481_at
SLasso+EBIC	1383110_at, 1392692_at
Edc+EBIC	1367728_at, 1367705_at

Example 2: Cardiomyopathy microarray data

In a study by Redfern et al. (2000) they generated the cardiomyopathy data. This data is a microarray data from a transgenic mouse of dilated cardiomyopathy. The mice overexpress a G protein-coupled receptor, designated Ro1, that is a mutated form of the human kappa opioid receptor, and that signals through the Gi pathway. When the receptor is overexpressed in the hearts of adult mice, the mice develop a lethal dilated cardiomyopathy that has many hallmarks of the human disease such as chamber dilation, left ventricular conduction delay, systolic dysfunction, and fibrosis. The sample size for the study was thirty (30). The thirty mice were divided into four groups. The control group was comprised of eight mice that were treated exactly the same as the eight-weeks experimental group except that they did not have the Ro1 transgene. A group of six transgenic mice expressed Ro1 for two weeks, which is approximately the amount of time required to reach maximal expression of Ro1 (Redfern et al., 1999). These mice did not show symptoms of disease. A group of nine transgenic mice expressed Ro1 for eight weeks and exhibited

cardiomyopathy symptoms. The recovery group of seven transgenic mice expressed Ro1 for eight weeks before expression was turned off for four weeks. To determine which gene expression changes were due to the expression of the Ro1 transgene, we want to find genes that correlate (positively or negatively) with the Ro1 expression profile as displayed.

For the cardiomyopathy study, available data consists of the $n \times p$ matrix of gene expression values $X = [x_{ij}]$ where x_{ij} is the expression level of the j^{th} gene $(j = 1, ..., p = 6,319)$ for the i^{th} mouse $(i = 1, ..., n = 30)$. Each mouse also provides an outcome (Ro1) measure y_i.

In our analysis of the data, we considered 100 replications by selecting a random sample of size 25 from the 30 specimens and apply the Edc+EBIC to it. We standardized the data set. From these 100 replications we selected the distinct genes that our method yielded. The Edc+EBIC selected genes Msa.10012.0, Msa.10044 and Msa.10108.0.

In Figure 4.1 - Figure 4.3, we show a scatterplot of the response variable Ro1 and each of the selected genes. We overlayed the scatterplots with a lowess curve to describe their relationships respectively.

Next we compare our result with results reported in some researches which used this data. These data were used by Hall and Miller (2009) and their method selected genes Msa.2877.0 and Msa.1166.0. Also Li et al. (2012) used it in feature screening via distance correlation learning and selected Msa.2134.0 and Msa.2877.0. Eventhough the genes selected by Edc+EBIC were different from those selected by these researches, the three methods agreed that the relationship between the response variable and the selected genes is nonlinear. If Pearson correlation were used in the feature selection process it may have missed these variables. No information was seen in the medical literature to associate the selected features with the Ro1 expression level.

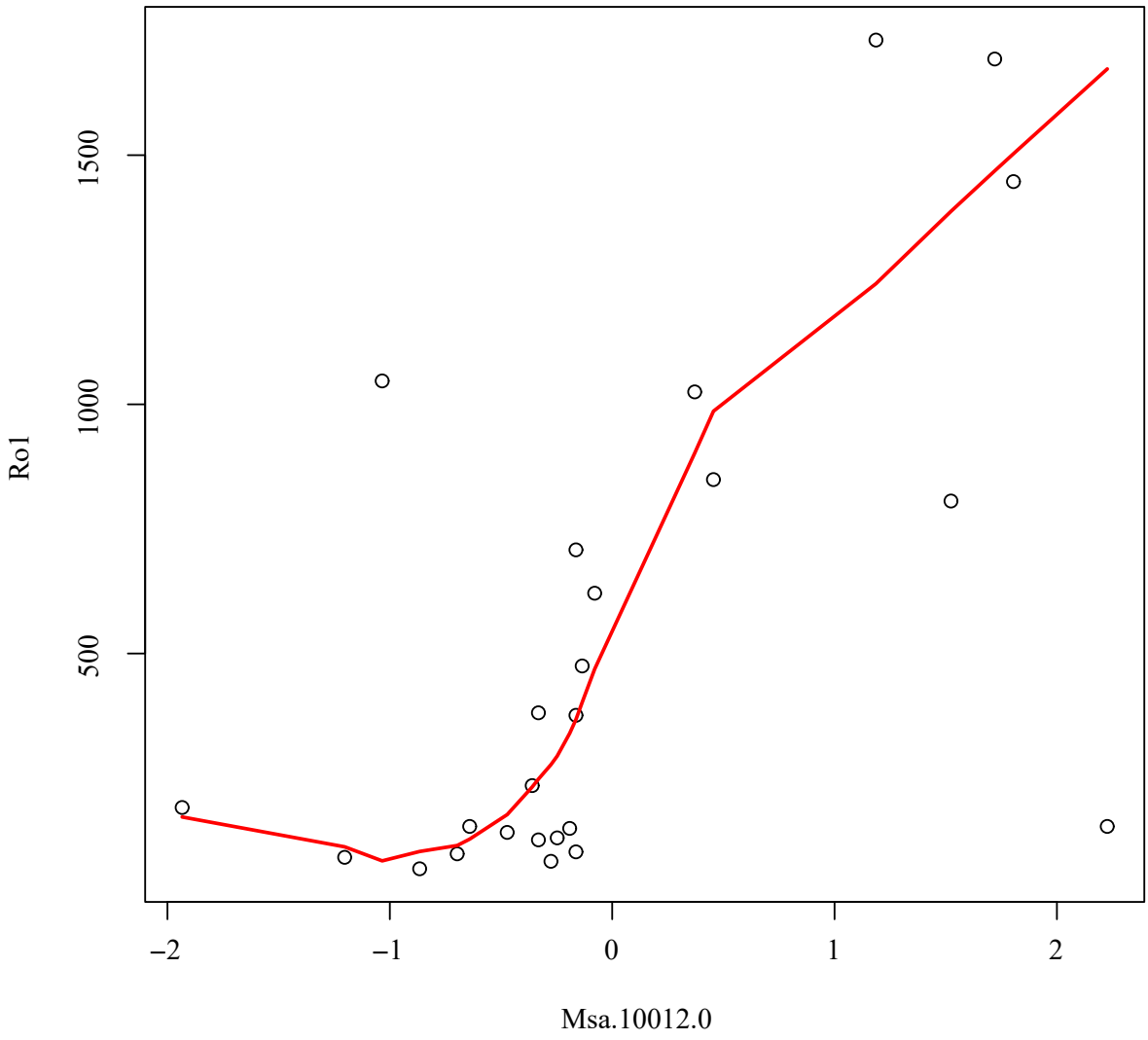

Figure 4.1 The scatterplot of Ro1 (Y) versus gene expression level Msa.10012.0 selected by Edc+EBIC, along with a fitting curve (in red). It is clear that the relationship between the variables is nonlinear.

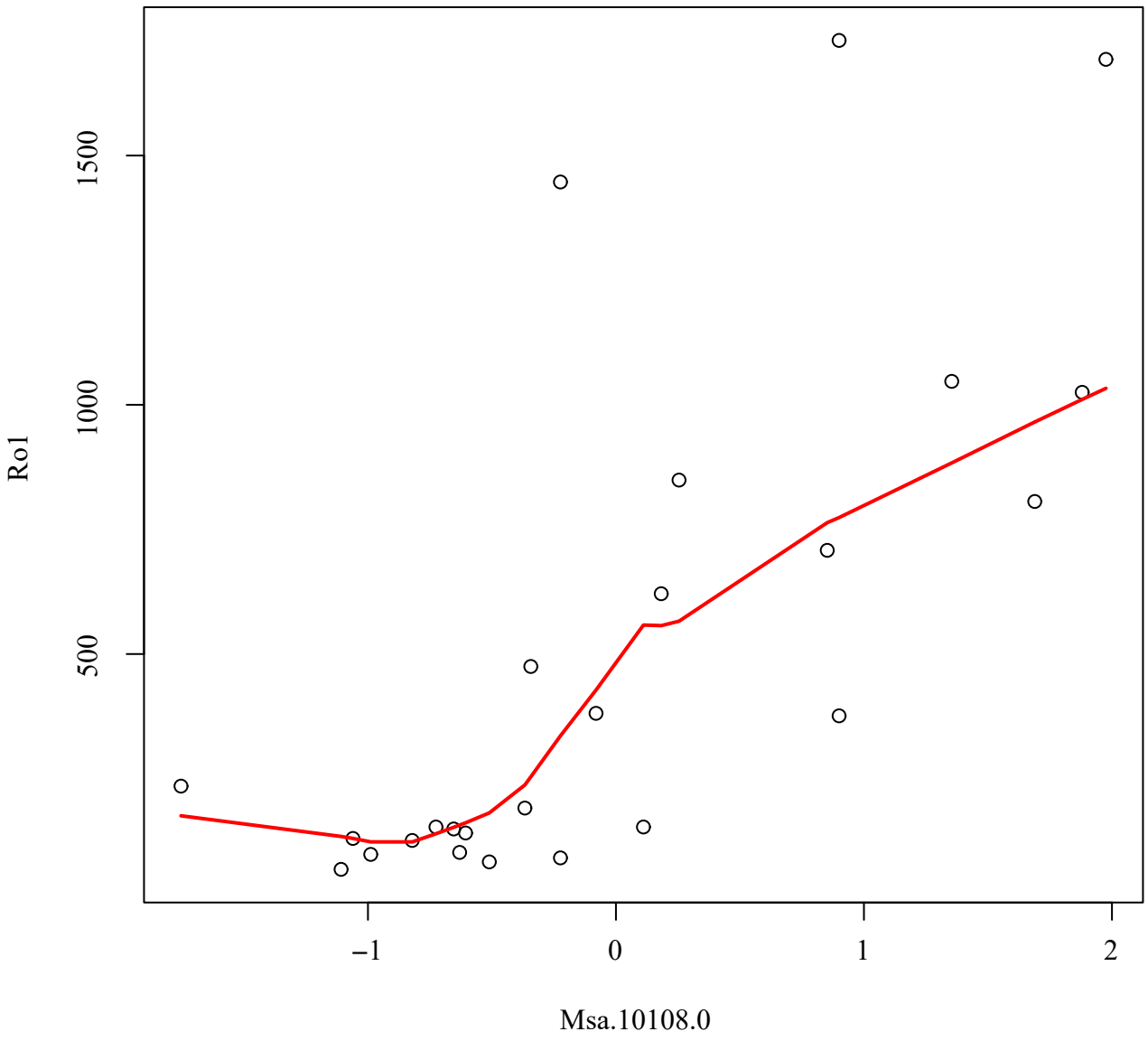

Figure 4.2 The scatterplot of Ro1 (Y) versus gene expression level Msa.10108.0 selected by Edc+EBIC, along with a fitting curve (in red). It is clear that the relationship between the variables is nonlinear.

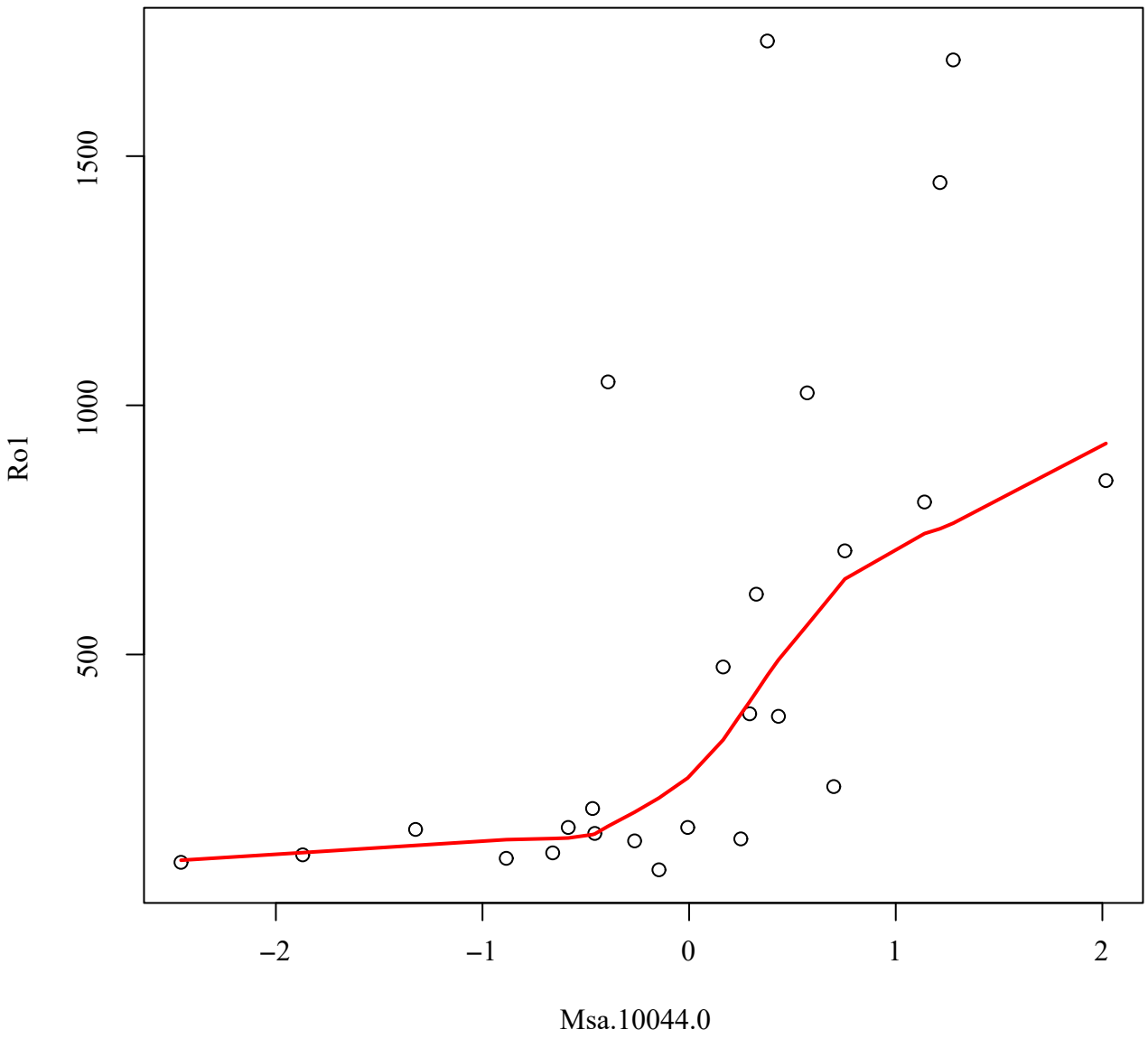

Figure 4.3 The scatterplot of Ro1 (Y) versus gene expression level Msa.10044.0 selected by Edc+EBIC, along with a fitting curve (in red). It is clear that the relationship between the variables is nonlinear.

CHAPTER 5 CONCLUSION AND DISCUSSION

In this chapter we summarize and discuss the results from the dissertation research,

5.1 Summarization of Dissertation Research

We have proposed a new method on sequential Lasso for feature selection in sparse high-dimensional linear models by introducing the energy distance correlation to replace the ordinary correlation in Luo and Chen (2014). The new sequential variable selection method which we call energy distance correlation with extended Bayesian Information Criteria (Edc+EBIC) is described in Chapter 3. At each stage of the sequential procedure we maximize the energy distance correlation between the response and each of the predictor variables. This maximization is done such that if a variable is selected in the previous stage, it's contribution to the response is removed so that it won't have a chance of being selected again. The active set of selected variables is updated once a variable is selected and the EBIC of the set is calculated. The process stops if the EBIC for the current active set is greater than the EBIC of the previous active set.

For the first part of our research we attempted to examine through simulations the model size selection by Adaptive Lasso and SCAD after a sure screening method proposed by Li et al. (2012) using distance correlation is applied to the data first. We observed that the average model size selected was quite high.

For the second part of our research we studied the properties of the new algorithm. It was shown that the new method is selection consistent. Two real-life data sets were analyzed by the new method to illustrate its use in applications. We compared through simulations the performance of Edc+EBIC with sequential Lasso, Adaptive Lasso, SCAD and SIS+SCAD. The simulation studies and the real data examples gave much insight into the performance of Edc+EBIC for sequential variable selection in high dimensional data analysis. Based on our findings we conclude the following.

1. The Edc+EBIC is equally a competitive method for variable selection in high dimensional

data. This is supported by the high PDR, low FDR and the average model sizes we observed in the simulations.

2. The Edc+EBIC performed equally well in the two health real datasets we considered. Our proposed method selecting very few features allows for model parsimony and easy of interpretation.

3. The selection consistency and sure screening property as evident in our simulations show that Edc+EBIC satisfies the oracle property. The oracle property means that the sparse relevant features can be exactly selected with probability converging to 1 and the effects of relevant features can be consistently estimated the same as they would be, were they obtained by knowing the relevant features in advance.

5.2 Discussion

As evident in the literature review several works have been done using the energy distance correlation as a means for variable selection. However as at the time of this research, to the best of our knowledge no research was done using energy distance correlation for sequential variable selection and used the extended Bayesian Information Criteria as the stopping criteria. We did followed the ideas of sequential Lasso and replaced their maximization of Pearson correlation by the energy distance correlation.

In the first place we established that the energy distance correlation was equally good for measuring the strength of linear associations just as the Pearson correlation. Thus it's advantageous to replace the Pearson correlation with distance correlation since the energy distance correlation was capable of measuring both linear and non-linear relationships.

Secondly we found out through a simulation exercise that when we applied the Distance Correlation - Sure Independence Screening (DC-SIS) proposed by Li et al. (2012) for variable screening followed by a regularization method such as SCAD and Adaptive Lasso the average model size selected was quite higher than expected and with high standard deviations.

Thirdly we proposed the Energy distance correlation with extended Bayesian Information Cri-

teria (Edc+EBIC) and examined through two simulation set-ups it's selection consistency. By the selection consistency of Edc+EBIC we expect that as the sample size increases our method selects the relevant variables in the data. We considered a diverging pattern i.e., as the sample size increases the number of predictors (features) also increases. From the results in Table 3.7 and Table 3.8 we observed that, as the sample size (n) increases, Edc+EBIC selected on average the expected number of predictors and did so with decreasing standard deviations meaning that through the simulation runs more and more of the selected predictors were close to the expected number of relevant predictors. We also observed the positive discovery rate which was 100% indicating that on average for each simulation run, out of the selected features all of the relevant features are selected. Additionally, of more importance is the small false discovery rates recorded as the sample size increases.

In the fourth place, we attempted to give a theoretical proof for the selection consistency and sure screening property of our method. We realized that the prove by Li et al. (2012) to show the selection consistency and sure screening of the energy distance is sufficient. This is because in our method we maximized the energy distance correlation betweeen $\tilde{y} = [I - H(s_{*k})]y$ and $\tilde{x}_j = [I - H(s_{*k})]x_j$ at each step of our procedure, while in their use of the energy distance correlation they maximized the distance correlation between (x, y) directly. For the second part of the prove for the selection consistency of EBIC, Luo and Chen (2014) gave a theoretical prove on the selection consistency of extended Bayesian Information Criteria as a stopping criteria in linear regression models.

In the fifth place, we compared our method with Adaptive Lasso, SCAD, SIS+SCAD and SLasso. We observed that in almost all simulations we considered, Edc+EBIC recorded the smallest false discovery rate. The SLasso and Edc+EBIC recorded on average model sizes which were close to the expected number of features. The positive discovery rate for Edc+EBIC is close to 100% in almost all the simulations.

In the sixth place, we considered two real-life data set problems. We examined gene expression data on an intercross experiment in rats. The gene TRIM32 is found to cause Bardet-Biedl

syndrome and the task was to select among 18,976 probes which are most correlated with it. Our method selected two probes just as SLasso but these probes were different. In the second example we considered the Cardiomyopathy microarray data where the task was to determine which genes were influential for overexpression of a G protein-coupled receptor, designated Ro1 out of 6319 genes. Our method selected three genes and we made a scatterplot of each selected feature and the response Ro1, which we observed showed a nonlinear relationship for all three, these three features may not have been selected if we had maximized the Pearson correlation.

5.3 Future Research Plan

In the near future, we will continue the research and work on the following topics.

1. The Edc+EBIC in this research was for a high dimensional data with a single response. The energy distance correlation is able to find the distance correlation between two vectors of unequal dimension. We plan to extend our Edc+EBIC to cover multiple response data. Technically the first part of applying our method to multiple responses data has been solved by Li et al. (2012) since their screening method with distance correlation is able to screen multiple response data. We just have to think through on how to sequentially select the features.

2. We observed in this research that many methods have been developed for analyzing high dimensional data. We intend to explore the large ocean of real datasets to see which methods are appropriate for data from a particular field.

3. We would also consider extending our Edc+EBIC to generalize linear models, survival data, time series data and longitudinal data.

BIBLIOGRAPHY

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*. Springer.

Breiman, L. (1993). Better subset selection using the non-negative garotte. Technical report, University of California, Berkeley.

Broman, K. W. and Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):641–656.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Clark, M. (2013). A comparison of correlation measures. https://m-clark.github.io/docs/CorrelationComparison.pdf.

de Siqueira Santos, S., Takahashi, D. Y., Nakata, A., and Fujita, A. (2014). A comparative study of statistical methods used to identify dependencies between gene expression signals. *Briefings in Bioinformatics*, 15(6):906–918.

Donoho, D. L. et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(2000):32.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Gorfine, M., Heller, R., and Heller, Y. (2012). Comment on "detecting novel associations in large data sets" by reshef et al, science dec 16, 2011. *Science*.

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18(3):533–550.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Heller, R., Heller, Y., and Gorfine, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510.

Hoeffding, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics*.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618.

Johnstone, I. M. and Titterington, D. M. (2009). Statistical challenges of high-dimensional data.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

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(507):1229–1240.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, 13(1):25–45.

Redfern, C. H., Coward, P., Degtyarev, M. Y., Lee, E. K., Kwa, A. T., Hennighausen, L., Bujard, H., Fishman, G. I., and Conklin, B. R. (1999). Conditional expression and signaling of a specifically designed g i-coupled receptor in transgenic mice. *Nature biotechnology*, 17(2):165–169.

Redfern, C. H., Degtyarev, M. Y., Kwa, A. T., Salomonis, N., Cotte, N., Nanevicz, T., Fidelman, N., Desai, K., Vranizan, K., Lee, E. K., et al. (2000). Conditional expression of a gi-coupled receptor causes ventricular conduction delay and a lethal cardiomyopathy. *Proceedings of the National Academy of Sciences*, 97(9):4826–4831.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524.

Rizzo, M. L. and Székely, G. J. (2016). Energy distance. *Wiley interdisciplinary reviews: Computational Statistics*, 8(1):27–38.

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434.

SHAN, L. (2012). Feature selection in high-dimensional studies.

Simon, N. and Tibshirani, R. (2014). Comment on" detecting novel associations in large data sets" by reshef et al, science dec 16, 2011. *arXiv preprint arXiv:1401.7645*.

Spearman, C. (1904). 'ôgeneral intelligence, õ objectively determined and measured. *American Journal of Psychology*, 15:201–93.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205.

Su, Y., Shi, Q., and Wei, W. (2017). Single cell proteomics in biomedicine: High-dimensional data acquisition, visualization, and analysis. *Proteomics*, 17(3-4):1600267.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.

Székely, G. J., Rizzo, M. L., et al. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

APPENDIX A SELECTED R PROGRAMS

```
#Distance Correlation with EBIC for GA1.
```

```{r}
library(energy)
#Distance Correlation with EBIC for GA1.
set.seed(02158)
n = 200; p = 276; rho = 0.5; real_p = 8
u <- rbinom(8,1,0.4)
z<-rnorm(8,mean=0,sd=(0.1/1.15))
beta<-((-1)^u)*(4*(n^(-0.15))+abs(z))


x = matrix(rnorm(p*n), nrow=n, ncol=p)
b = beta
sigm<-((t(b)%*%cov(x[,1:8])%*%b)-0.8*(t(b)%*%cov(x[,1:8])%*%b))/0.8


b <- as.matrix(b)
x <- as.matrix(x)



###Step 1
x <- scale(x,scale = TRUE)


selected  <- matrix(NA, nrow = 200, ncol = 1)
```

```
selectedD <- matrix(0, nrow = 200, ncol = 25)

for (j in 1:200) {
  EEBIC <- c(Inf, 10000)
  k = 2; s = 0; selectedC = NULL
  x1 <- x
  y <- x[,1:8]%*%b + rnorm(n,0,sqrt(sigm))
  y <- scale(y)
  idx <- seq(ncol(x))

  while (EEBIC[k] < EEBIC[k-1]) {

    out <- matrix(NA, nrow = nrow(x), ncol = 1)
    for (i in 1 : nrow(x)){
      out[i,1] = DCOR(x[,i],y)$dCor
    }

    stemp <- which.max(out)
    selectedC <- c(selectedC,idx[stemp])
    idx <- idx[-c(stemp)]



    sk <- as.vector(stemp)
    a <- as.matrix(x[,sk])
    H <- a%*%solve(t(a)%*%a)%*%t(a)

    p <- ncol(x)
```

```
      ynew <- (diag(n) - H)%*%y
      xnew <- (diag(n) - H)%*%x[,-c(sk)]
      s <- s+length(sk)
      EEBIC[k+1] <- n*log(sum((ynew)^2)/n) + (s)*log(n) +
      2*(1-log(n)/(2.1*log(p)))*log(choose(p,s))
      y <- ynew
      x <- xnew
      k = k + 1
    }
    x<-x1
    #sds<-print(s-1)
    sds<-s-1
    selected[j,1]<-sds
    selectedD[j,1:length(selectedC)-1]<-selectedC[1:length(selectedC)-1]

}

mean(selected)
sd(selected)

so <- seq(8);so


pdrout <- matrix(NA, nrow = 200, ncol = 1)
for(i in 1:200){
  pdrout[i,1] = length(intersect(selectedD[i,],so))/length(so)
}
mean(pdrout)
```

```
sd(pdrout)


sc <- seq(276)[-so]
fdrout <- matrix(NA, nrow = 200, ncol = 1)
for(i in 1:200){
   fdrout[i,1] = length(intersect(selectedD[i,], sc))/sum(selectedD[i,]!=0)
}
mean(fdrout)
sd(fdrout)

```