



W&M ScholarWorks

Undergraduate Honors Theses

Theses, Dissertations, & Master Projects


5-2018

Ultra-High Dimensional Statistical Learning

Yanxin Xu

College of William and Mary

Follow this and additional works at: <https://scholarworks.wm.edu/honorstheses>

 Part of the [Applied Mathematics Commons](#), and the [Statistics and Probability Commons](#)

Recommended Citation

Xu, Yanxin, "Ultra-High Dimensional Statistical Learning" (2018). *Undergraduate Honors Theses*. Paper 1196.

<https://scholarworks.wm.edu/honorstheses/1196>

This Honors Thesis is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Undergraduate Honors Theses by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.

Ultra-High Dimensional Statistical Learning

A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelor of Science in Mathematics Department from
The College of William and Mary

by

Yanxin Xu

Accepted for Honors
(Honors, High Honors, Highest Honors)

Guannan Wang 
Type in the name, Director

Ross Iaci 
Type in the name

Bin Ren 
Type in the name

Williamsburg, VA
May 4, 2018

Ultra-High Dimensional Statistical Learning

Yanxin Xu

Department of Mathematics,
College of William and Mary,
Williamsburg, VA 23187-8795, USA

Email: yxu08@email.wm.edu

Abstract

Advancements in information technology have enabled scientists to collect data of unprecedented size as well as complexity. Nowadays, high-dimensional data commonly arise in diverse fields as biology, engineering, health sciences, and economics. In this project, we consider both linear and non-parametric models with variable selection in the high-dimensional setting by assuming that only a small number of index coefficients influence the conditional mean of the response variable. Both the numerical results and the real data application demonstrate that the proposed approach selects the correct model with a high frequency and estimates the model coefficients accurately even for moderate sample size and ultra-high dimensionality.

Contents

1	Introduction	1
2	Bardet-Biedl Syndrome and Data	5
3	Methodology	7
3.1	Models	7
3.2	Model Selection Criteria	8
3.3	Variable Selection Techniques	10
3.3.1	Classical Variable Selection Technique	10
3.3.2	Shrinkage Methods	10
3.3.3	Sure Independence Screening (SIS)	12
3.3.4	Iterative Sure Independence Screening (ISIS)	13
4	Application and Results	15
4.1	Application of Methodology	15
4.2	Results	16
4.2.1	Linear Models Comparison	17
4.2.2	Additive Models Comparison	18
5	Conclusion	21
5.1	Acknowledgements	23

List of Figures

3.1	Sure Independence Screening + penalized regression	13
-----	--	----

Chapter 1

Introduction

Among numerous modern problems in multiple scientific fields, such as biology, engineering and health sciences, “high-dimensionality” is a noteworthy characteristic feature. Here, the word “high-dimensionality” refers to the case that the number of explanatory variables p is large, and is potentially much larger than the sample size n in the data. The analysis of high-dimensional data gives rise to many new challenges and opportunities for developing new statistical methodologies. Take a simple linear model for example, when having many more unknown parameters than the number of observations, the least-squares fitting is ill-posed.

As the number of explanatory variables increases, it is often useful and reasonable to assume that the p -dimensional parameters are sparse with many components being zero, and this assumption is well known as the “sparsity” assumption. With sparsity, it becomes a big challenge to identify the significant variables efficiently, and several procedures have been developed to conquer such challenge.

Variable selection plays an important role to overcome such challenge in high-dimensional data analysis. The two standard variable selection techniques usually being used are forward selection and backward selection. For forward selection, we sequentially adds important explanatory variables into the model one at a time, while for backward selection, we

successively eliminates insignificant variables out of the model. However, these stepwise selection methods can be computationally expensive and produces large error when the dataset is large.

Shrinkage methods can get more accurate estimations and work more efficiently as they simultaneously select variables and estimate coefficients. Tibshirani (1996) introduced the Least Absolute Shrinkage and Selection Operator (LASSO), which adds the sum of absolute values of the coefficients to the traditional sum of square error objective function and shrinks some coefficients to zero to get model with fewer explanatory variables. Fan and Li (2001) proposed another shrinkage method SCAD, which shares some properties with the LASSO but makes its penalty functions bounded by a constant to generate unbiased estimation for variables with larger coefficients. Moreover, MCP proposed by Zhang (2010) reduces the biasness to the variables with larger coefficients and is more likely to penalize smaller coefficients to zero.

All the above-mentioned shrinkage methods are proved to be theoretically stable and computationally efficient for selecting the correct model even when p is relatively large. However, when the dimension of the index parameters exceeds the sample size, it becomes a serious scientific endeavor to find the relationship between the response and the explanatory variables. Obstacles to high-dimensionality can be both theoretical and practical. To address these issues, several researchers have contributed some useful computational algorithms. For example, Fan and Lv (2008) introduced Sure Independent Screening (SIS), which is based on correlation learning, to reduce dimensionality from high to a moderate scale that is below the sample size. Moreover, the marginal regression can be nonlinear even when the true underlying model is linear. Fan, Feng and Song (2011) further extend the correlation learning in SIS to marginal nonparametric learning which can address the above-mentioned address this issue.

Nowadays, many problems in biology such as microarray and RNA-seq data analysis are involved in the case of $p \gg n$ since there are numerous biological features to be

estimated but very few samples that can be produced in lab. In this project, we aim to analyze a representative huge dataset with $p \gg n$ by applying the aforementioned techniques to conquer the challenge of how to accurately extract important information from a huge amount of data.

The article is organized as follows. In Chapter 2 we introduce the background and dataset of the study. In Chapter 3 we give a review of regression models, model selection criteria, and variable selection techniques. In Chapter 4 we implement the techniques in Chapter 3 and provide the results of our study, including the comparison of different models and variable selection results. Some concluding remarks are given in Chapter 5.

Chapter 2

Bardet-Biedl Syndrome and Data

Bardet-Biedl syndrome (BBS) is a human genetic disorder that can affect many body systems including the retina. One major feature of Bardet-Biedl syndrome is vision loss. According to Scheetz et al. (2006), gene *TRIM32* has been identified as the 11th member to the BBS associated gene family.

The goal of this project is to identify the genes which are statistically significantly related to gene *TRIM32* and build an accurate model. The results could help the researchers discover additional genes relevant to BBS or other human eye diseases.

The corresponding experiment selected 120 12-week-old male F2 offspring rats for tissue harvesting, microarray analysis, and genotyping. As discussed in Scheetz et al. (2006), among the 31,042 noncontrol gene probes on the array, probes that were not expressed in the eye or probes that lacked sufficient variation were excluded. A probe to be considered “expressed in the eye”, its maximum expression value among the 120 F2 rats must be greater than the 25th percentile of the entire set of expression values. For a probe to be considered as “sufficiently expressed”, it must have at least 2-fold variations in gene expression level.

As a result, the dataset records the gene expression values of 18,976 probes that were considered to be both “expressed in the eye” and “sufficiently expressed” for all the 120

F2 rats. The dataset is publicly available in the Gene Expression Omnibus repository, www.ncbi.nlm.nih.gov/geo (GEO accession id: GSE5680).

In the dataset, gene ***TRIM32*** (probe ID 1389163_at) is the response variable (Y) with the remaining 18,975 gene probes as explanatory variables (X_j 's). Scheetz et al. (2006) also suggest that among the 18,975 explanatory probes, there are about 3,057 gene probes having a 4-fold or greater change in expression. From biology point of view, the relatively high degree of expression variations indicates that these gene probes are potentially more important.

Having 120 samples (n) and 18,975 explanatory probes (p), the microarray dataset ($n \ll p$) raises an ultra-high dimensional problem. As pointed out in Fan and Lv (2008), the dimensionality is “ultra-high” when the number of explanatory variables, p , is one or several orders of magnitude larger than the sample size, n .

Chapter 3

Methodology

In this section, we introduce the methods we use to analyze the microarray data in this project.

3.1 Models

To identify important genes related to *TRIM32*, the most straight forward and intuitive method is linear regression and the model can expressed as

$$Y_i = \sum_{j=1}^p X_{ij}\beta_j + \epsilon_i, \quad i = 1, \dots, n,$$

where Y_i is the i th observation of the response variable, β_j , $j = 1, \dots, p$ is the unknown coefficient to estimate, X_{ij} , $i = 1, \dots, n$, $j = 1, \dots, p$ represents the explanatory variable and $\epsilon_i \sim N(0, \sigma^2)$ is the white noise which is independent from the explanatory variable.

The method of least squares is a standard approach in linear regression, and “Least squares” means that the overall solution minimizes the sum of the squares of the residuals. Specifically,

$$\boldsymbol{\beta} = \{\beta_j\}_{j=1}^p = \min_{\boldsymbol{\beta}} \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p X_{ij}\beta_j \right)^2.$$

The basis assumption for linear model is that the relationship between the response and the explanatory variable is linear. However, in practice, the relationship between the covariates and the response variable may not be linear. In contrast to the restricted form of linear models, additive models are more flexible by accommodating nonlinear functions for each covariate. The additive model is expressed as

$$Y_i = \sum_{j=1}^p m_j(X_{ij}) + \epsilon_i \quad (3.1)$$

where Y_i is the i th observation of the response variable, $m_j(\cdot)$ is the unknown smooth function for the j th feature, X_{ij} , $i = 1, \dots, n$, $j = 1, \dots, p$ represents the explanatory variable and $\epsilon_i \sim N(0, \sigma^2)$ is the white noise which is independent from the explanatory variable.

To estimate the unknown function $m_j(\cdot)$ in (3.1), we consider the use of polynomial spline smoothing in Xue and Yang (2006). The appeal of polynomial splines is that they often provide good approximations of smoothing functions with a simple linear combination of spline basis. Suppose that each X_{ij} , $j = 1, \dots, p$, takes value between a and b , where a and b are some finite numbers. We divide $[a, b]$ into $(N + 1)$ subintervals with a sequence of points given as

$$\{t_k\}_{k=1-r}^{N+r} = t_{1-r} = \dots = t_{-1} = t_0 = a < t_1 < \dots < t_N = b = t_{N+1} = \dots = t_{N+r}.$$

Let $B_r(u) = \{B_{k,r}\}_{k=1-r}^N$ be the spline basis functions of order r , and $m(\cdot)$ of which the k th order derivative of $m(\cdot)$ is continuous on $[a, b]$ can be estimated by

$$\hat{m} = B_r(x)\gamma,$$

where γ is the spline coefficient.

3.2 Model Selection Criteria

Model selection criteria plays an important role in selecting the best model from multiple candidates. Two traditional model selection criteria are **Akaike Information Cri-**

terion (AIC) by Akaike (1973) and **Bayesian Information Criterion (BIC)** by Schwarz (1978) , and they are defined as

$$\begin{aligned} \text{AIC} &= n \log (\text{MSE}) + 2 \times k; \\ \text{BIC} &= n \log (\text{MSE}) + \log (n) \times k, \end{aligned}$$

where $\text{MSE} = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p X_{ij} \hat{\beta}_j \right)^2$ and k represents the number of parameters in the model.

Another popular criteria for model selection is **k-fold cross validation (CV)** by Stone (1974) . We randomly partition the entire dataset into k equal sized subsets. For each iteration (k iterations in total), we pick a single subset as the validation data and use the remaining $k - 1$ subsets as training data to build a model. We collect the result of how well the model predicts the validation data for each iteration and calculate the average of k results, we can get a single estimation as the model selection criterion.

As pointed out in Wang, Li and Leng (2009) , the traditional BIC can identify the true model consistently, as long as the dimension of the explanatory variable is fixed. When the dimensionality is fixed dimension, the number of candidate models is also fixed. Thus, as long as the corresponding BIC can consistently differentiate the true model from an arbitrary candidate model, the true model can be identified with probability tending to 1. However, if the predictor dimension also goes to ∞ , the number of candidate models increases extremely fast. Thus, the traditional theoretical arguments for BIC are no longer applicable. To overcome such a challenging difficulty, Wang, Li and Leng (2009) proposed a **modified BIC (mBIC)** which is defined as

$$\text{mBIC} = n \log (\text{MSE}) + \log (n) \times k \times C_n,$$

where $C_n > 0$ is some positive constant. As one can see, when $C_n = 1$, $\text{mBIC} = \text{BIC}$ and Wang, Li and Leng (2009) suggested $C_n = \log (\log (d))$ based on extensive numerical studies.

When the model space is even larger, Chen and Chen (2008) reexamined the Bayesian paradigm for model selection and proposed an **extended family of Bayes information criteria (eBIC)**. eBIC takes into account both the number of unknown parameters and the complexity of the model space and is defined as

$$\text{eBIC} = n \log(\text{MSE}) + \log(n) \times k + 2\gamma \log \binom{p}{k}, \quad 0 \leq \gamma \leq 1.$$

3.3 Variable Selection Techniques

3.3.1 Classical Variable Selection Technique

Forward selection and backward selection are two standard variable selection techniques.

Forward selection starts from a null model (no variables in the model). For each step, we try all variables from the remaining predictors one at a time to test which variable statistically improves the model the most based on a certain model selection criterion such as AIC or BIC, and then add this variable to the model. We keep repeating the process until no more variables can improve the model.

Backward selection starts from a full model (all variables in the model). For each step, we remove a variable in the model that has the lowest statistical significance to the model based on AIC or BIC. We keep repeating the process until removing any of the remaining variables in the model will incur a significant loss.

3.3.2 Shrinkage Methods

Unfortunately, all the above-mentioned classical variable selection techniques are computationally expensive, particularly in high dimensional situations. Thus, in the past decade, various shrinkage methods, such as the LASSO by Tibshirani (1996) and the SCAD by Fan and Li (2001), have been proposed. The shrinkage methods can select variables and estimate coefficients simultaneously. It has been shown that when the tuning parame-

ters can be selected appropriately, the true model can be identified consistently. More discussion can be found in Fan and Li (2001), Fan and Peng (2004), and Zou (2006). Recently, similar results have also been extended to the situation with a diverging number of parameters by Fan and Peng (2004).

The main idea of shrinkage method is to add a penalty function $p_\lambda(\cdot)$ to the traditional sum of square error objective function, which can be expressed as

$$R = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p X_{ij} \beta_j \right)^2 + \sum_{j=1}^p p_\lambda(|\beta_j|).$$

Several penalty functions can be considered in the penalized regression:

- **Least Absolute Shrinkage and Selection Operator (LASSO)** by Tibshirani (1996)

$$p_\lambda(|\beta|) = \lambda|\beta|.$$

- **Smoothly Clipped Absolute Deviation (SCAD)** by Fan and Li (2001)

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta| & \text{if } |\beta| \leq \lambda; \\ -\left(\frac{|\beta|^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)}\right) & \text{if } \lambda < |\beta| \leq a\lambda; \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta| > a\lambda \end{cases}$$

where $a = 3.7$ is suggested in Fan and Li (2001).

- **Minimax Concave Penalty (MCP)** by Zhang (2010)

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta| - \frac{|\beta|^2}{2a} & \text{if } |\beta| \leq a\lambda; \\ \frac{a\lambda^2}{2} & \text{if } |\beta| > a\lambda \end{cases}$$

where $a = 3$ is suggested in Zhang (2010).

There are many other different penalty functions such as Adaptive Lasso by Zou (2006). In this project, we use the three penalty functions above (LASSO, SCAD, and MCP) to compare the results.

3.3.3 Sure Independence Screening (SIS)

In practice, it is often assumed that when the dimension p is very high, only a small number of explanatory variables among X_1, X_2, \dots, X_p contribute to the response Y . Moreover, in biology, a genetic disease is often related to one or a group of genes. Therefore, with the assumption of sparsity in our project, we can improve our estimation accuracy by choosing a subset of significant predictors.

In this project, we implement Sure Independence Screening (SIS) by Fan and Lv (2008) to screen the explanatory variables.

As pointed by Fan and Lv (2008), Sure Independence Screening has a property that all the important variables will survive after variable screening with a probability tending to one. Therefore, the screening process of applying componentwise marginal regression for each explanatory variable and selecting a set of variables by MSE or coefficient in SIS guarantees to keep the important variables after screening.

The basic procedures of Sure Independence Screening + penalized regression:

1. Standardize all explanatory variables X_j , for $j = 1, 2, \dots, p$, and the response variable Y .
2. Apply marginal regression on each explanatory variable X_j to Y ($Y \sim X_j$, for $j = 1, 2, \dots, p$).

Calculate the coefficient $\omega_j = X_j^T Y$ (linear SIS only) or the MSE for each $j = 1, 2, \dots, p$ (linear and nonparametric SIS).

3. Select a subset of d explanatory variables X_j with the first d largest coefficients $|\omega_j|$, or the first d smallest MSEs.

The new dimension of features is reduced from p to d , where $d < n$, usually pick $d = n - 1$ or $d = n/\log(n)$

4. After shrinking the data matrix to $n \times d$, we can then apply penalized regression,

using different model selection criteria (BIC/eBIC/CV) to do further variable selection.

Note that sorting by the coefficients in the process of variable screening is faster but it only applies to linear SIS. By contrast, screening by MSE is slower but it works for both linear and nonparametric SIS.

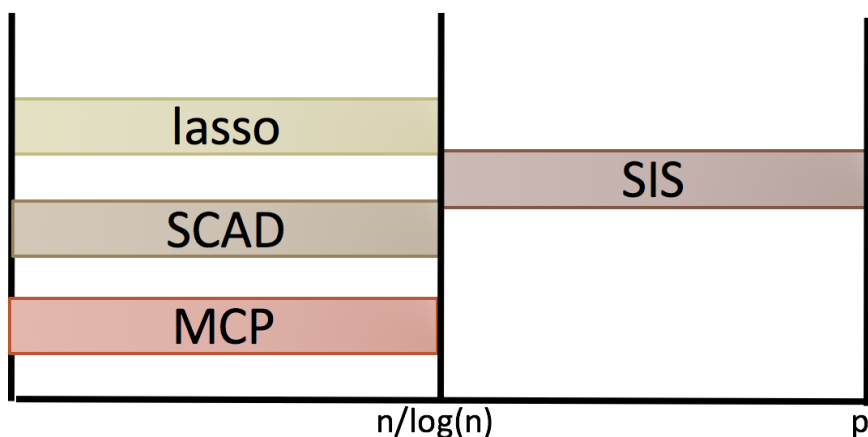


Figure 3.1: Sure Independence Screening + penalized regression

Figure 3.1 shows an example of using SIS to reduce the dimension of feature from p to $d = n/\log(n)$ and then applying penalized regression to select variables.

3.3.4 Iterative Sure Independence Screening (ISIS)

Fan and Lv (2008) also introduced an extension of Sure Independence Screening (SIS): iterative SIS (ISIS).

The basic procedures of Iterative Sure Independence Screening + penalized regression:

1. Standardize all explanatory variables X_j , for $j = 1, 2, \dots, p$, and the response variable Y .

2. For each step, use an SIS based variable selection method such as SIS-LASSO to select a subset of explanatory variables.
3. Get the residuals from regressing Y over the chosen subset of explanatory variables. ($residual = Y - \hat{Y}$).
4. Treat the residuals as the new responses and repeat from Procedure 2 by applying the same variable selection method to the remaining variables. ($residual \sim X_{remaining}$)
5. After several iterations, we get a subset of variables with size $d < n$. In practice, we can pick a maximum number of variables in the subset or a maximum number of steps to end the iterations.

An advantage of iterative SIS is that it not only keeps the marginally correlated explanatory variables as SIS, but it also considers variables that have high joint correlations with the response. Therefore, using iterative SIS can overcome some problems in the original SIS and make the variable selection more reasonable.

Chapter 4

Application and Results

In this chapter, we implement all the methodologies described in Chapter 3 to the Microarray dataset. To evaluate the different methods, we calculate mean squared prediction error (MSPE) which can be used to determine how well a model predicts unknown data. The lower MSPE suggests the better prediction for future data of a model.

4.1 Application of Methodology

In this project, we employ all the techniques in Chapter 3 to select variables and build models.

As mentioned in Chapter 2, about 3,000 gene probes exhibit relatively higher variances and are considered to be potentially more important variables. To speed up the performance, we select 3,000 explanatory variables with the largest variances instead of all the 18,975 predictors as a starting point. (we use $p = 18,975$ to make comparison later)

For building linear models, we directly apply SIS package in R, which iteratively screens the predictors and uses penalized linear regression to select variables. We could get 9 linear methods, the combination of choosing one from the three penalty functions

(LASSO/SCAD/MCP) and one from the three model selection criteria (BIC/eBIC/CV). We use the 9 linear methods to build linear models.

For building additive models, we implement nonparametric SIS by simulating the process of linear SIS. In nonparametric SIS, we first get the spline basis for each explanatory variable and apply marginal regression to each basis to select a group of the bases with the lowest marginal errors. The number of bases we select is $n - 1$, which is 119 in this dataset. We then get 9 nonparametric methods: choosing one from the three group penalty functions (group LASSO/group SCAD/group MCP) and one from the three model selection criteria (BIC/eBIC/CV). We use the 9 nonparametric methods to build additive models.

In practice, we use default tuning parameters for all penalty functions and model selection criteria. ($a = 3.7$ in SCAD and group SCAD, $a = 3$ in MCP and group MCP, $\gamma = 1$ in ebic, $\text{nfolds} = 10$ in CV)

4.2 Results

To calculate MSPE, we randomly pick 80 samples as the training data, apply **SIS** with a certain method (penalty function+model selection criterion) to get a new model, and use the parameters from the new model to predict the remaining 40 samples. We compare the prediction \hat{Y} with the response Y of the 40 samples and calculate the MSE. Repeating this process for 100 times, we get the averaged MSE as the MSPE.

4.2.1 Linear Models Comparison

MSPE	LASSO	SCAD	MCP
BIC	0.0148	0.0180	0.0207
eBIC	0.0148	0.0180	0.0207
CV	0.0136	0.0160	0.0155

Table 4.1: MSPE of the 9 linear models ($p = 3000$)

From Table 4.1, some partial conclusions are drawn:

1. **LASSO** predicts unknown data better than **SCAD** and **MCP**.
2. Using model selection criterion **CV** predicts unknown data better than using **BIC** and **eBIC** for each penalty function.

We select the linear model with the lowest MSPE for each penalty function. Among the three selected linear models, the model using method (LASSO+CV) has the lowest MSPE.

We further use $p = 18,975$ to select variables and build models using the three selected methods (LASSO+CV/SCAD+CV/MCP+CV):

	LASSO		SCAD		MCP	
	$p = 3000$	$p = 18,975$	$p = 3000$	$p = 18,975$	$p = 3000$	$p = 18,975$
MSE	0.0023	0.0013	0.0014	0.0006	0.0014	0.0006
MSPE	0.0136	0.0080	0.0160	0.0106	0.0155	0.0131
# of variables	25	25	25	25	25	25

Table 4.2: Comparison of 3 selected linear models

From Table 4.2, some partial conclusions are drawn:

1. The MSEs and MSPEs are lower than the those using $p = 3000$. Therefore, using the full dataset ($p = 18,975$) to build linear models and select variables may help get a better result.
2. **LASSO** predicts unknown data better than **SCAD** and **MCP**, which is identical to the partial conclusion when using $p = 3000$.
3. **LASSO+CV** might be the best method to apply for building linear model.

4.2.2 Additive Models Comparison

MSPE	group LASSO	group SCAD	group MCP
BIC	0.0205	0.0886	0.1463
eBIC	0.0222	0.0218	0.0380
CV	0.0136	0.0139	0.0274

Table 4.3: MSPE of the 9 additive models ($p = 3000$)

From Table 4.3, some partial conclusions are drawn:

1. **Group LASSO** and **group SCAD** predict unknown data better than **group MCP**.
2. Using model selection criterion **CV** predicts unknown data better than using **BIC** and **eBIC** for each group penalty function.

We select the model with the lowest MSPE for each group penalty function. Among the three selected additive models, the model using method **group LASSO+CV** and **group SCAD+CV** have low MSPEs.

We further use $p = 18,975$ to select variables and build models using the three selected methods (group LASSO+CV/group SCAD+CV/group MCP+CV):

	group LASSO		group SCAD		group MCP	
	$p = 3000$	$p = 18,975$	$p = 3000$	$p = 18,975$	$p = 3000$	$p = 18,975$
MSE	0.0018	0.0047	0.0084	0.0017	0.0112	0.0085
MSPE	0.0136	0.0184	0.0139	0.0165	0.0274	0.0236
# of variables	37	18	15	26	1	1

Table 4.4: Comparison of 3 selected additive models

From Table 4.4, some partial conclusions are drawn:

1. For **group LASSO** and **group SCAD**, the MSEs are lower than the those using $p = 3000$ while the MSPEs are higher, which raises a problem of “overfitting”. Therefore, using the full dataset ($p = 18,975$) cannot get better results.
2. **Group LASSO** and **group SCAD** each selects a larger number of variables while **group MCP** only selects one variable. Therefore, the **group MCP** model is not a relatively good estimation, but the only variable in this model needs to be considered.
3. The model using **group LASSO+CV** has the lowest MSPE but has too many variables. The model using **group SCAD+CV** has low MSPE and a moderate number of variables. Therefore, **group SCAD+CV** with $p = 3000$ might be the best method to apply for building additive model.

Chapter 5

Conclusion

From the results we get, linear models we built by using **SIS+penalized linear regression** can fit the dataset very well and can also predict future unknown data quite successfully. The linear model using shrinkage method **LASSO** with model selection criterion **CV** performs the best among all. In addition, using dataset starting from a larger dimension of p also improves data fitting and unknown data prediction. Therefore, the variables in the models with $p = 18,975$ need to be considered.

On the other hand, the additive models using **group LASSO** and **group SCAD** give good estimations and predictions. However, using the full dataset to build additive models does not improve unknown data prediction. We need to take the variables selected by these additive models with $p = 3000$ into consideration.

probe ID	gene name	$p = 3000$						$p = 18,975$		
		LASSO	SCAD	MCP	group LASSO	group SCAD	group MCP	LASSO	SCAD	MCP
1389584_at		✓	✓	✓	✓	✓	✓			
1374106_at		✓	✓	✓	✓	✓				
1383110_at		✓	✓	✓	✓	✓				
1367627_at	Gatm	✓	✓	✓						
1384466_at	Ispd	✓			✓	✓				
1385168_at	Lrif1	✓	✓	✓						
1393817_at	Wdr76	✓			✓	✓				
1379881_at				✓	✓	✓				
1372248_at	Sesn1							✓	✓	✓
1373887_at								✓	✓	✓
1389910_at	Tmem230							✓	✓	✓

Table 5.1: variables that appear in at least 3 models

Based on Table 5.1, the first 8 probes, especially 1389584_at, 1374106_at, and 1383110_at, are very likely to be both statistically and biologically significant genes. The other 3 probes, 1372248_at, 1373887_at, and 1389910_at, although do not have high degree of expression variations as Scheetz et al. (2006) suggests, might also be important genes.

After some careful studies of these selected genes in lab, there might be some further discussions about whether the high degree of gene expression variation is biologically significant.

5.1 Acknowledgements

Thank you to the support by William and Mary Charles Center Honors Fellowship. Thank you to Professor Guannan Wang for giving me so much support during the year and helping me get a much deeper understanding in statistics. Thank you to Professor Ross Iaci for supporting me during my entire four years of undergraduate study. Thank you to Professor Bin Ren for giving me constructive advice to my thesis and Honors defense.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, Ed. B. N. Petrox and F. Caski. Budapest: Akademiai Kiado. 267–281.

Breheny, P. and J, Huang. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing* 25:173–187.

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

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association* 96:1348–1360.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics* 32:928–961.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. *Proceedings of the International Congress of Mathematicians: Madrid* 3:595–622.

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

- Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association* 106:544–557.
- Scheetz, T., Kim, K., Swiderski, R., Philp, A., Braun, T., Knudtson, K., Dorrance, A., DiBona, G., Huang, J., Casavant, T., et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences* 103:14429–14434.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6:461–464.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with Discussion). *J. R. Journal of the Royal Statistical Society, Series B: Statistical Methodology* 39:111–147.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* 58:167–288.
- Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 71:671–683.
- Xue, L. and Yang, L. (2006). Additive coefficient modeling via polynomial spline *Statistica Sinica* 16:1423–1446.
- Zhang, N. and Siegmund, D. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* 63:22–32.
- Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* 38:894–942.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101:1418–1429.