2017-08-05

Some New Theories and Applications on L1 Shrinkage Estimation

Hongmei Liu
University of Miami, 1989luo@gmail.com

Follow this and additional works at: https://scholarlyrepository.miami.edu/oa_dissertations

Recommended Citation
https://scholarlyrepository.miami.edu/oa_dissertations/1954

This Open access is brought to you for free and open access by the Electronic Theses and Dissertations at Scholarly Repository. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of Scholarly Repository. For more information, please contact repository.library@miami.edu.
UNIVERSITY OF MIAMI

SOME NEW THEORIES AND APPLICATIONS ON L1 SHRINKAGE ESTIMATION

By

Hongmei Liu

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

August 2017
A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

SOME NEW THEORIES AND APPLICATIONS ON L1 SHRINKAGE ESTIMATION

Hongmei Liu

Approved:

J. Sunil Rao, Ph.D.  Hermant Ishwaran, Ph.D.
Professor of Biostatistics  Professor of Biostatistics

Lily Wang, Ph.D.  Guillermo Prado, Ph.D.
Associate Professor of Biostatistics  Dean of the Graduate School

Nagi Ayad, Ph.D.
Associate Professor of Psychiatry and Behavioral Sciences
In this thesis, I develop some new variable selection and statistical modeling techniques in the framework of L1 shrinkage estimation with applications to high dimensional genomic and pharmacogenomic datasets.

In the first part of the thesis, I revisit the problem of variable selection in linear regression models. While numerous variable selection procedures have been developed, their finite sample performance can often be less than satisfactory. I develop a new strategy for variable selection in the adaptive least absolute shrinkage and selection operator (Lasso) and adaptive elastic-net estimations with $p_n$ diverging. The basic idea first involves using the trace paths of their LARS solutions to bootstrap estimates of maximum frequency (MF) models conditioned on model dimension. Conditioning on dimension effectively mitigates overfitting. But to deal with underfitting these MFs are then prediction-wighted. I show that the new method is not only model selection consistent, but also has attractive convergence rate, which lead to outstanding finite sample performance.

In the second part, I propose a new statistical model to re-explore the Genomics of Drug Sensitivity (GDSC) study (Garnett et al., 2012). To link drug sensitivity with genomic profiles, the study screened 639 human tumor cell lines with 130 cancer drugs ranging from known chemotherapeutic agents to experimental compounds. However, the statistical challenges still exist in analyses of this dataset: i) biomarkers cluster
among the cell lines; ii) clusters can overlap (e.g. a cell line may belong to multiple clusters); iii) drugs should be modeled jointly. I introduce a new multivariate regression model with a latent overlapping cluster indicator variable to address these issues. I then propose the generalized mixture of multivariate regression (GMMR) models and build a connection with it to the new model. I develop a new EM algorithm for numerical computations in the GMMR model. The proposed new model can answer specific questions in the GDSC data: i) can cancer-specific therapeutic biomarkers be detected, ii) can drug resistance patterns be identified along with predictive strategies to circumvent resistance using alternate drugs?

In the third part of the thesis, I set out to tackle another challenging problem related to GDSC data – that of validating models built on one dataset but tested on similar datasets generated in other laboratories. The Genomics of Drug Sensitivity (GDSC) and Cancer Cell Line Encyclopedia (CCLE) are two major resources that can be used to mine for therapeutic biomarkers for cancers of a large variety. Recent studies found that while the genomic profiling seems consistent, the drug response data is not. As a result, both predictions and signatures do not validate well for models built on one dataset and tested on the other. I present a partitioning strategy based on a data sharing concept, which directly estimates the amount of discordance between datasets and in doing so, also allows for extraction of reproducible signals. I show both significantly improved test set prediction accuracy and signature validation as compared to other approaches that have been tried.
Acknowledgements

I wish to express my great appreciation to my advisor Professor J. Sunil Rao, who guided me through all stages of this thesis. I benefit so much from his deep thinking in biostatistics. I am truly grateful for his constant encouragement and generous help in every aspect.

I am deeply grateful to Professor Hermant Ishwaran for giving us excellent lectures, writing solid recommendation letter and giving insightful advices for my job application. My sincere thanks also go to Professor Lily Wang and Professor Nagi Ayad for joining my dissertation committee and providing valuable revision suggestions. In addition, I thank all other faculty members in the department for their great lectures, and my friends for their sincere help. Finally, I wish to dedicate this thesis to my husband Xi, my parents and sister for their love and support.
Contents

List of Figures vii

List of Tables ix

1 Introduction 1

2 Prediction Weighted Maximum Frequency Selection 4

2.1 Motivation and Literature Review . . . . . . . . . . . . . . . . . . . . 4

2.2 Bootstrapping the Adaptive Lasso and Adaptive elastic-net Estimators 11

2.3 The MF Procedure . . . . . . . . . . . . . . . . . . . . . . . . 13

2.3.1 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

2.3.2 Asymptotic properties . . . . . . . . . . . . . . . . . . . . 15

2.4 The WMF Procedure . . . . . . . . . . . . . . . . . . . . . . . 21

2.4.1 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . 21

2.4.2 Asymptotic properties . . . . . . . . . . . . . . . . . . . . 24

2.4.3 Computations . . . . . . . . . . . . . . . . . . . . . . . . . 25

2.5 Empirical Studies . . . . . . . . . . . . . . . . . . . . . . . . . 26

2.5.1 Simulations of the adaptive Lasso WMF procedure . . . . . . . 26

2.5.2 Simulations of the adaptive elastic-net WMF procedure . . . . 29

iv
2.5.3 Leukemia data analysis ........................................ 33
2.6 Extensions ........................................................... 34
  2.6.1 Generalized linear models ..................................... 34
  2.6.2 Ultra-high dimensional data .................................... 37
2.7 Discussions .......................................................... 38
2.8 Appendix ............................................................... 38
  2.8.1 Proofs of Lemma 2.1 and Theorems 2.1-2.3 ............... 38
  2.8.2 Additional simulation results ................................. 62

3 Precision Therapeutic Biomarker Identification 65
  3.1 Motivation .......................................................... 65
  3.2 Literature Review .................................................. 67
    3.2.1 Clustering and mixture model ................................. 67
    3.2.2 Overlapping clustering ......................................... 70
    3.2.3 Mixture regression models ................................... 72
    3.2.4 Mixture multivariate regression models .................... 76
  3.3 The Proposed Method .............................................. 77
    3.3.1 Statistical modeling ........................................... 77
    3.3.2 The generalized mixture of multivariate regression models .. 78
    3.3.3 Numerical solutions ............................................ 80
    3.3.4 Selections of tuning parameters and the $K$ .................. 84
    3.3.5 Asymptotic properties ......................................... 85
  3.4 Simulation Studies ............................................... 87
    3.4.1 Design .......................................................... 87
    3.4.2 Cluster recovery quality measures ............................ 88
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Introductory example</td>
<td>9</td>
</tr>
<tr>
<td>2.2</td>
<td>Simulation results for scenario 1</td>
<td>27</td>
</tr>
<tr>
<td>2.3</td>
<td>Simulation results for scenario 2</td>
<td>28</td>
</tr>
<tr>
<td>2.4</td>
<td>Simulation results for scenario 3</td>
<td>29</td>
</tr>
<tr>
<td>2.5</td>
<td>Simulation results for scenario 4</td>
<td>30</td>
</tr>
<tr>
<td>2.6</td>
<td>Simulation results for scenario 5</td>
<td>31</td>
</tr>
<tr>
<td>2.7</td>
<td>Simulation results for scenario 6</td>
<td>32</td>
</tr>
<tr>
<td>2.8</td>
<td>GLM example</td>
<td>36</td>
</tr>
<tr>
<td>2.9</td>
<td>Simulation results for scenario 1 using residual bootstrap</td>
<td>62</td>
</tr>
<tr>
<td>2.10</td>
<td>Simulation results for scenario 2 using residual bootstrap</td>
<td>62</td>
</tr>
<tr>
<td>2.11</td>
<td>Simulation results for scenario 3 using residual bootstrap</td>
<td>63</td>
</tr>
<tr>
<td>2.12</td>
<td>Simulation results for scenario 4 using residual bootstrap</td>
<td>63</td>
</tr>
<tr>
<td>2.13</td>
<td>Simulation results for scenario 5 using residual bootstrap</td>
<td>64</td>
</tr>
<tr>
<td>2.14</td>
<td>Simulation results for scenario 6 using residual bootstrap</td>
<td>64</td>
</tr>
<tr>
<td>3.1</td>
<td>Simulation results for scenario 1</td>
<td>89</td>
</tr>
<tr>
<td>3.2</td>
<td>Simulation results for scenario 2</td>
<td>90</td>
</tr>
<tr>
<td>3.3</td>
<td>Biomarkers from the GMR model</td>
<td>92</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>3.4</td>
<td>Biomarkers from the GMMR model</td>
<td>94</td>
</tr>
<tr>
<td>3.5</td>
<td>Mutation biomarkers from the GMR model</td>
<td>95</td>
</tr>
<tr>
<td>3.6</td>
<td>Drugs for resistance</td>
<td>96</td>
</tr>
<tr>
<td>3.7</td>
<td>Drugs for resistance</td>
<td>97</td>
</tr>
<tr>
<td>4.1</td>
<td>Schematic of data sharing</td>
<td>103</td>
</tr>
<tr>
<td>4.2</td>
<td>Test set error boxplots</td>
<td>108</td>
</tr>
<tr>
<td>4.3</td>
<td>Correlations between $\hat{\delta}$ and the true $\delta$</td>
<td>109</td>
</tr>
<tr>
<td>4.4</td>
<td>Correlations between $\hat{\beta}$ and the true $\beta$</td>
<td>110</td>
</tr>
<tr>
<td>4.5</td>
<td>Prediction validation plot</td>
<td>111</td>
</tr>
<tr>
<td>4.6</td>
<td>Prediction validation plot</td>
<td>112</td>
</tr>
<tr>
<td>4.7</td>
<td>Signature validation plot</td>
<td>114</td>
</tr>
<tr>
<td>4.8</td>
<td>Signature validation plot</td>
<td>115</td>
</tr>
</tbody>
</table>
List of Tables

2.1 The leukemia classification using adaptive Lasso . . . . . . . . . . . 34
2.2 The leukemia classification using adaptive elastic-net . . . . . . . . . 34
3.1 Informations of the 5 drug groups fitted in the GMMR model . . . . . 93
Chapter 1

Introduction

The process of extracting important patterns and trends from data and making precise predictions is called statistical learning, which plays an essential role in many subjects (Hastie et al., 2009a). For example, given performance measures and economic data for a particular company, one wants to predict the price of the company’s stock in 3 months. Similarly, in health studies it is highly in demand to identify the risk factors for heart or stroke disease based on clinical and demographic data.

The modern information era, where vast amounts of data are being collected at fairly low costs in diverse fields, has brought golden opportunities and remarkable challenges to statistics (Fan and Li, 2006). A particularly prominent problem is the high dimensionality of data frequently seen in a wide range of subjects from genomics and health sciences to engineering and economics. For example, in a leukemia classification study using microarray data (Golub et al., 1999) thousands of genes are candidate predictors for the type of leukemia. Fan and Fan (2008) demonstrated that classification using all features in a high dimensional dataset can perform as poor as a random guess.
Hence variable selection is essential in high dimensional statistical modeling. Traditional best-subset selection, in which the total number of candidate models grows exponentially with the dimension \( p \), is computationally infeasible when \( p > 30 \), not to mention for high dimensional datasets. To ease the computational costs, some greedy stepwise algorithms are often used as a substitute to best-subset selection. These kinds of algorithms however are highly variable in their selection.

The Lasso shrinkage estimation by adding an \( L_1 \) penalty function to the model fitting criterion has made far-reaching contributions to variable selection. The \( L_1 \) penalty can continuously shrink coefficients to exact 0 by using a large enough tuning parameter in the function. As a result, the Lasso can do simultaneous variable selection and estimation. Invention of the Lasso opened a Pandora box, a variety of its variants and generalizations were successively developed and widely used. The family includes adaptive Lasso, (adaptive) elastic-net, fused Lasso and SCAD. Since the tuning parameters control the variable selection in this kind of shrinkage estimations, the amount of penalty that should be added becomes an especially important issue.

In Chapter 2, I develop a new variable selection method in the family of \( L_1 \) shrinkage estimations. The method effectively mitigates overfitting by conditioning on the dimension. The underfitting is then handled by using a prediction-based weighting scheme. I show that the method is variable selection consistent and more importantly, it enjoys a fast convergence rate, which guarantees very good finite sample performance.

The (finite) mixture of regression models introduces more flexibility than standard regression to model unobserved cross-sectional heterogeneity in data. The model assumes that a sample of observations come from a number of latent sub-populations with each generated from a unique density function. But an underlying assump-
tion of the model is that the latent subpopulations are partitioned. In Chapter 3, I propose a generalized mixture of multivariate regression models to accommodate overlapping clustering and fit multiple response variables together. Modeling with multiple response variables induces improved estimation as well as clustering accuracy. I use a penalized maximum likelihood approach in estimation for concurrent variable selection in high dimensional feature space. The GDSC data screened 130 cancer drugs against a series of cell lines from 13 cancer tissue types to systematically link the cancer drugs to specific genomic alterations that can be used as biomarkers to indicate their selective therapeutic effectiveness (Garnett et al., 2012). By using the new model to fit the data, I find drug-cancer-specific therapeutic biomarkers and strategies to overcome drug resistance.

High throughput pharmacogenomic screening in cancer cell lines provides a powerful platform to discover predictive biomarkers to guide rational cancer therapeutic schemes. Given the importance of such assays, proper validation of the findings is necessary. A number of recent studies found that two major pharmacogenomic studies – the GDSC and CCLE – are highly inconsistent in their measured drug response data. Some insight as to why has been put forward, but the discordancies cannot be explained fully by these. In Chapter 4, I propose a discordancy partitioning approach by adapting a data sharing model, which was introduced to fit multi-sources datasets. The method partitions the genomic effects into two parts - the true underlying effect and the portion due to discordancy. By using the new method, I find many new reproducible biomarkers and test set prediction error rates are markedly lower than those in other compared approaches.
Chapter 2

Prediction Weighted Maximum Frequency Selection

2.1 Motivation and Literature Review

Two fundamental endeavors are generally pursued in statistical learning: prediction with a high level of accuracy and selection of a concise set of relevant predictors (Zou, 2006). While designing powerful predictive procedures remains a popular and critical research topic, demands of which have even generated a trendy branch of machine learning in recent decades, nevertheless variable selection plays a pivotal role for enhanced scientific discovery especially in frequent presence of high dimensional data in modern statistics. For example, the Framingham Offspring study in 1971 in efforts to identify significant risk factors associated with heart, stroke and other diseases in general population had revolutionized preventive medicine and radically changed the way the general public and medical community view the origin of disease; See Fan and Li (2006) for a comprehensive review of feature selection in knowledge discovery.
Consider the standard linear regression model

\[ y = X\beta + \varepsilon, \tag{2.1} \]

where \( y = (y_1, \ldots, y_n)^T \) is a vector of responses, \( X = (X_1, \ldots, X_{pn}) \) is an \( n \times p_n \) design matrix of predictors, \( \beta = (\beta_1, \ldots, \beta_{pn})^T \) is a vector of unknown regression parameters, \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \) is a vector of independent and identically distributed (i.i.d.) random errors. I allow \( p_n \) to increase with \( n \).

Because some elements of \( \beta \) might be 0, a family of penalized least squares (PLS) estimators were developed for variable selection and estimation,

\[ \hat{\beta} = \arg \min_\beta \| y - X\beta \|^2 + \sum_{j=1}^{p_n} \rho(|\beta_j|, \lambda), \tag{2.2} \]

where \( \| \cdot \| \) is the \( L_2 \)-norm, \( \lambda \geq 0 \) are regularization parameters, and \( \rho(|\beta_j|, \lambda) \) is positive valued for \( \beta_j \neq 0 \). Fan and Li (2006) pointed out that through variable selection one can focus on a small number of important predictors for enhanced scientific discovery and potentially improve prediction performance by removing noise variables.

Penalized \( L_q \)-regression is a special case of (2.2) with \( \rho(|\beta_j|, \lambda) = \lambda|\beta_j|^q, q \geq 0 \), which includes the best subset selection for \( q = 0 \); the Lasso (Tibshirani, 1996) for \( q = 1 \) and the ridge regression (Hoerl and Kennard, 1970) for \( q = 2 \). Best subset selection is known to be computationally infeasible for high dimensional data and inherently discrete in variable selection (Breiman, 1995). Ridge regression does not possess a variable selection property. The Lasso, however, can do simultaneous estimation and variable selection because its \( L_1 \) penalty is singular at the origin and
can shrink some coefficients to exact 0 with a sufficiently large \( \lambda \) (Fan and Li, 2001).

The Lasso estimate is

\[
\hat{\beta} = \arg \min_{\beta} \| y - X\beta \|^2 + \lambda \sum_{j=1}^{p} |\beta_j|,
\]

(2.3)

where \( \lambda \geq 0 \). Fan and Li (2001) advocated three properties for penalty functions: sparsity, unbiasedness and continuity, wherein the Lasso possesses the first and third properties. More specifically, the Lasso can continuously shrink all coefficients toward 0 as \( \lambda \) increases from 0 to \( \infty \). During the process, the estimated coefficients from small to large will be consecutively shrunk to exact 0. On one hand, the continuous shrinkage improves prediction accuracy due to the bias-variance trade-off. On the other hand however, the uniform shrinkage of Lasso cross all coefficients produces bias especially for large coefficients estimations.

Denote the true coefficients vector as \( \beta^0 \). Let \( A = \{ j : \beta^0_j \neq 0 \} \) and by postulate \( |A| = p_0 < p \). Fan and Li (2001) brought forward the concept of oracle properties that a good estimation procedure should satisfy:

I: Variable selection consistency, \( \{ j : \hat{\beta}_j \neq 0 \} = A \).

II: Enjoys the optimal root-\( n \) estimation rate, \( \sqrt{n}(\hat{\beta}_A - \beta^0_A) \rightarrow_d N(0, \Sigma_A) \),

where \( \Sigma_A \) is the covariance matrix knowing the true model. It has been shown by Zou (2006) that the Lasso is not an oracle procedure which does not fulfill the oracle properties; Also see Fan and Li (2001); Meinshausen and Buhlmann (2006) and Zhao and Yu (2006) for similar conclusions.

Other PLS estimators that can do simultaneous estimation and variable selection include the SCAD (Fan and Li, 2001) and adaptive Lasso (Zou, 2006) both enjoying the oracle properties; the elastic-net (Zou and Hastie, 2005) capable of detecting
grouped effects; and the adaptive elastic-net (Zou and Zhang, 2009) combining advantages of the adaptive Lasso and elastic-net.

Efficient algorithms are available for numerically computing the PLS estimators. Efron et al. (2004) developed the least angle regression (LARS) algorithm for variable selection in linear regression models. It starts with zero coefficients for all covariates and finds the covariate most correlated with the response variable, to which a coefficient is assigned such that a second covariate is found to have the same correlation (in magnitude) with the current working residual as the first one. The algorithm then proceeds in the direction equiangular between the two selected covariates until a third covariate satisfying the recruiting requirements enters and so on. I refer readers to Efron et al. (2004) for details. A simple modification of the LARS algorithm produces the full set of Lasso solutions \( \{ \hat{\beta}(\lambda) : \lambda > 0 \} \) in (1.3) with computation efforts in the same order of magnitude as a single ordinary least-squares (OLS) fit. The LARS algorithm can be easily modified to generate the entire solution paths for adaptive Lasso and (adaptive) elastic-net. Arise from the idea of the LARS algorithm, Zhang et al. (2010) introduced the PLUS algorithm for computations of the PLS estimations when the penalty function is a quadratic spline such as the SCAD. Also see Zou and Li (2008), Friedman et al. (2007) and Fan and Lv (2011) for the local linear approximation (LLA), the coordinate-wise descent and the iterative coordinate ascent (ICA) algorithms.

Selection of \( \lambda \) is essential in above penalized least squares estimation procedures. Although methods such as the SCAD, adaptive Lasso and adaptive elastic-net enjoy the oracle properties asymptotically, their optimal properties rely on particular specifications of the \( \lambda \), whose magnitude controls the complexity of a selected model and trade-off between bias and variance in estimators (Fan and Lv, 2010). Numerous
studies have been conducted for this purpose. The multi-fold cross-validation (CV) and
generalized cross-validation (GCV) are frequently applied for tuning parameters
selection (Tibshirani, 1996; Fan and Li, 2001; Zou, 2006; Zou and Hastie, 2005). But
they overfit the model asymptotically (Zhang, 1993). For consistent variable selec-
tion, Feng and Yu (2013) devised a consistent cross-validation procedure (CCV) for
the Lasso. The information criteria (IC) are by far the most explored approaches
wherein one seeks to minimize

\[
\text{measure of model fitting} + a_n \times \text{measure of model complexity},
\]

where \(a_n\) is a positive sequence depending on the sample size \(n\). The minus 2 times
log-likelihood is often used as a measure of model fitting, while in PLS estimations
the sum of squared errors is used instead.

Wang and Leng (2007) suggested to use BIC \((a_n = \log(n))\) in adaptive Lasso. A
modified BIC \((a_n = C_n \log(n))\) was later introduced for adaptive Lasso and SCAD
when \(p_n\) is diverging (Wang et al., 2009). Chen and Chen (2008) brought forward
a family of extended BIC (EBIC) for linear regression models and later generalized
it to sparse generalized linear models in a small-n-large-p problem (Chen and Chen,
2012). Fan and Tang (2013) developed a generalized information criterion (GIC) for
penalized likelihood methods with \(p_n\) growing exponentially with \(n\), where \(a_n\) was
taken as \(\log\{\log(n)\}\log p\). Finally Meinshausen and Bühlmann (2010) proposed the
stability selection (SS) for their randomized Lasso.

Although variable selection consistency has been established for these procedures,
their finite sample performance can often be less than optimal (Section 2.5 ahead
demonstrates this in simulation studies). This is a phenomenon commonly seen in
I propose a new method for tuning parameters selection, focusing in particular on the adaptive Lasso and adaptive elastic-net estimators. A simple example helps to illustrate the basic idea. Consider the adaptive Lasso estimator in following example.

**Example 1.** Data are drawn from model (2.1) with $\beta = (3, 1.5, 0, 0, 2, 0, \ldots)^T$, row vectors of the design matrix $\mathbf{x}_i \overset{iid}{\sim} \mathcal{N}_{10}(0, \Sigma)$ with $\Sigma(i, j) = 0.3| i - j |$ and $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, 3^2)$ for $i = 1, \ldots, 100$. So the true model size here is 3.

Figure 2.1 (top) shows the full adaptive Lasso solution path from the LARS algorithm. In the figure, each step indicates a dimension change in the estimator. These steps are called transition points in (Zou et al., 2007). They showed that if using
information criteria such as the AIC or BIC to identify the optimal $\lambda$ in adaptive Lasso, it lies in one of the transition points. This result helps to justify uses of the LARS algorithm and our subsequent focus on the transition points. The rest question is about how to choose from these transition points.

Figure 2.1 (middle and bottom) gives a brief look at our proposed method. The middle panel shows the estimated maximum frequency (MF) of a candidate model given the dimension. The MF estimation is done by a bootstrapping algorithm using the transition points. The strategy of conditioning on dimension has two important consequences: i) for overfit dimensions, the MFs are dramatically smaller than the true dimension MF (other than the full model of course), and ii) underfit dimensions can also produce large MF values. Point i) is important because for variable selection, overfitting is usually much more difficult to deal with. So one must now deal with underfitting. I do this by introducing a prediction-based weight to the MFs (labeled as WMF). The results are shown in bottom panel of Figure 2.1. As is evident, now the true dimension, which maps to the true model, stands out beautifully from all others.

In Section 2.2, I briefly review the adaptive Lasso and adaptive elastic-net estimators and introduce the bootstrapping approach for each. In Section 2.3, the MF procedure is introduced and its asymptotic properties are established. The WMF procedure and its variable selection consistency are presented in Section 2.4. Comprehensive simulation studies and the leukemia data analyses are shown in Section 2.5. Section 2.6 describes extensions of the MWF procedure to generalized linear models (GLMs) and to ultra-high dimensional data.
2.2 Bootstrapping the Adaptive Lasso and Adaptive elastic-net Estimators

Denote $\beta_0$ the true value of $\beta$ with model size $p_0$, and $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_{p_n})^T$ a consistent estimate of $\beta_0$. The adaptive Lasso (Zou, 2006) estimator is

$$\hat{\beta}_a = \arg \min_{\beta} ||y - X\beta||^2 + 2\lambda_n \sum_{j=1}^{p_n} \omega_j |\beta_j|,$$  
(2.5)

where $\omega_j = |\tilde{\beta}_j|^{-\gamma}$, $\gamma \geq 0$. It was suggested to use the ordinary least-squares (OLS) estimator or the best ridge estimator (if collinearity exists) for $\tilde{\beta}$. Under certain regularity conditions, $\hat{\beta}_a$ was shown to enjoy the oracle properties.

The elastic-net estimator (Zou and Hastie, 2005) is

$$\hat{\beta}_e = (1 + \frac{\lambda_{n2}}{n}) \left\{ \arg \min_{\beta} ||y - X\beta||^2 + \lambda_{n2} \sum_{j=1}^{p_n} |\beta_j|^2 + \lambda_{n1} \sum_{j=1}^{p_n} |\beta_j| \right\}.  
(2.6)$$

It overcomes several limitations pertaining to the Lasso: 1) the added $L_2$ penalty is strictly convex to allow grouping effects; 2) In a $p_n > n$ case, it can potentially estimate all $p_n$ predictors, while the Lasso can only find at most $n$ predictors.

(Zou and Zhang, 2009) proposed the adaptive elastic-net to combine strengths of the elastic-net and adaptive Lasso. The adaptive elastic-net estimator is

$$\hat{\beta}_{ae} = (1 + \frac{\lambda_{n2}}{n}) \left\{ \arg \min_{\beta} ||y - X\beta||^2 + \lambda_{n2} \sum_{j=1}^{p_n} |\beta_j|^2 + \lambda_{n1}^+ \sum_{j=1}^{p_n} \omega_j |\beta_j| \right\},  
(2.7)$$

where $\omega_j = |\tilde{\beta}_{ej}|^{-\gamma}$, $\gamma \geq 0$ and $\hat{\beta}_e = (\hat{\beta}_{e1}, \ldots, \hat{\beta}_{ep_n})^T$ is the elastic-net estimator in (2.6). Note that $\lambda_{n2}$ takes the same value for the $L_2$ penalty function in (2.6) and
(2.7), because the $L_2$ penalty contributes to the same kind of grouping effects. On the other hand, $\lambda_{n1}$ and $\lambda_{n1}^+$ are allowed to be different as they control the sparsity in estimators. Under some regularity conditions, $\hat{\beta}_{ae}$ was shown to enjoy the oracle properties.

I now detail bootstrapping for these two estimators. There are typically two ways of generating bootstrap observations for model (2.1).

1. *Bootstrapping pairs* (Efron and Efron, 1982). Let $\hat{F}(X, y)$ be the empirical distribution putting mass $n^{-1}$ on each data pair $(x_i, y_i), i = 1, \ldots, n$. Generate i.i.d. paired bootstrap data $\{(x_i^*, y_i^*), i = 1, \ldots, n\}$ from $\hat{F}(X, y)$. The bootstrap analog of $\hat{\beta}_a$, denoted as $\hat{\beta}_a^*$, is to replace $(X, y)$ with $(X^*, y^*)$ in (2.5) where $X^* = (x_1^*, \ldots, x_n^*)^T$ and $y^* = (y_1^*, \ldots, y_n^*)^T$. So is the bootstrap analog of $\hat{\beta}_{ae}$, denoted as $\hat{\beta}_{ae}^*$. Under the Iak condition that $X^T X \rightarrow \infty, X^T X (X^T X^*)^{-1} \rightarrow 1$ almost surely (Shao, 1996).

2. *Bootstrapping residuals* (Efron, 1979). Calculate the $i$th residual

$$\hat{\varepsilon}_i = y_i - x_i^T \hat{\beta},$$

where $\hat{\beta}$ is a ridge estimate of $\beta_0$. Generate i.i.d. bootstrap residuals $\{\varepsilon_i^*, i = 1, \ldots, n\}$ from the empirical distribution that puts mass $n^{-1}$ on each centered residual, $\hat{\varepsilon}_i = \hat{\varepsilon}_0 - \bar{\varepsilon}_0$, where $\bar{\varepsilon}_0$ is the average of $\hat{\varepsilon}_0, i = 1, \ldots, n$. Then the i.i.d. residual bootstrap data is $\{(x_i, y_i^*), i = 1, \ldots, n\}$ where $y_i^* = x_i^T \hat{\beta} + \varepsilon_i^*$. The bootstrap analog of $\hat{\beta}_a$, denoted as $\tilde{\beta}_a^*$, is to substitute $y$ with $y^*$ in (2.5). So is the bootstrap analog of $\hat{\beta}_{ae}$, denoted as $\tilde{\beta}_{ae}^*$.

In next section, I introduce the MF procedure which takes use of above bootstrap estimators.
2.3 The MF Procedure

2.3.1 Methods

Denote a $j$-dimensional candidate model from the $i$th bootstrap data as $M_j^i$.

**Algorithm 1:** The MF procedure for adaptive Lasso

1. Draw $B$ (residual or paired) bootstrap data;
2. Use the LARS algorithm to fit each bootstrap data, then get $B$ collections of candidate models, \{${M}_1^i, \ldots, {M}_{p_n}^i$\}, $i = 1, \ldots, B$;
3. At each dimension $j$, count the frequency of each unique model in \{${M}_1^j, \ldots, {M}_B^j$\}, denoted as \{${c}_1^j, \ldots, {c}_t^j$\} where $t$ is the number of unique models. Let $MF_j = \max\{c_1^j, \ldots, c_t^j\}$ corresponding to model $M_j$;
4. Select the dimension $r^*$ and model $M_{r^*}$ s.t.

$$r^* = \max\{j : j = \arg \max_{1 \leq i \leq p - 1} MF_i\}.$$

**Remark 2.1.** In the 4th step, the full model is excluded because it will destroy the maximum frequency rule by having the highest frequency, $B$, all the time. If there is a tie at the maximum of $MF_i, 1 \leq i \leq p - 1$, I select the one at the highest dimension. This strategy guarantees asymptotic variable selection consistency of the MF procedure, which will be discussed in Section 2.3.2.

The MF procedure for adaptive elastic-net is in parallel. In the 2nd step however, the LARS-EN algorithm (Zou and Hastie, 2005) is used instead to fit each bootstrap data.

I discussed in Introduction consequences of the MF procedure by conditioning on dimension. Here I use a simple orthogonal design with i.i.d. normal random errors to study underlying properties driving that performance. In this case, I have $X^TX = I$ and the adaptive elastic-net reduces automatically to the adaptive Lasso (Zou and
Zhang, 2009). Denote $X_j$ the $j$th column of $X$. Then the adaptive Lasso estimator is

$$\hat{\beta}_j = \left\{ |X_j^T y| - \frac{\lambda_n}{|\hat{\beta}_j|} \right\}_+ sgn(X_j^T y), \quad j = 1, \ldots, p, \quad (2.8)$$

where $z_+$ equals to $z$ if $z > 0$ otherwise 0. I can expand the $X_j^T y$ by

$$X_j^T y = \beta_{0j} + X_j^T \varepsilon,$$

where $X_j^T \varepsilon \sim N(0, \sigma^2)$. The following Lemma gives an order relationship for $X_j^T y$’s.

**Lemma 2.1.** Suppose $X^T X = I$, then I have

$$P \left( |X_i^T y| > |X_j^T y| \right) > 0.5 \quad \text{if} \quad |\beta_{0i}| > |\beta_{0j}|,$$

$$P \left( |X_i^T y| > |X_j^T y| \right) = 0.5 \quad \text{if} \quad |\beta_{0i}| = |\beta_{0j}|,$$

for $i, j \in \{1, \ldots, p\}$.

In combine with the fact that $\frac{\lambda_n}{|\beta_i|} > \frac{\lambda_n}{|\beta_j|}$ asymptotically for $\beta_{0i} < \beta_{0j}$, it is easy to deduce from (2.8) that given a $\lambda_n$ adaptive Lasso tends to select those variables, corresponding to the first $k_{\lambda_n}$ largest $|\beta_j|$’s, with the highest probability.

Without loss of generality, suppose $|\beta_0|$ is decreasingly ordered. Denote $S_r$ a $r$-dimensional model containing the first $r$ elements of $|\beta_0|$, and denote $W_r$ any other $r$-dimensional models. Let $\hat{A}_r$ be an adaptive Lasso model estimate given the model size is $r$, $P(\hat{A}_r = S_r \mid r)$ indicates the conditional probability of $\hat{A}_r = S_r$ given the
model size. Then preceding deductions from (2.8) can be formalized as

(1). \[ P(\hat{A}_r = S_r \mid r) > P(\hat{A}_r = W_r \mid r), \quad 0 < r \leq p_0, \]  
(2.9)

(2). \[ P(\hat{A}_r = W^1_r \mid r) = P(\hat{A}_r = W^2_r \mid r), \quad p_0 < r < p_n, \]  
(2.10)

where \( W^1_r \) and \( W^2_r \) are two \( r \)-dimensional models s.t. \( S_{p_0} \subset W^1_r, W^2_r \).

Above properties of the adaptive Lasso coincides to some extent with the results of Theorem 2 in (Dey et al., 2008). By (2.10), zero predictors will be equally likely selected at an overfit dimension. As a result \( P(\hat{A}_r = M_r \mid r) \) (see Algorithm 1 for definition of \( M_r \), \( p_0 < r < p_n \), drops down dramatically, which is why I see a huge gap between the true dimension and overfit dimensions in Figure 2.1 (middle). On the other hand, \( P(\hat{A}_r = S_r \mid r) \) at some underfit dimensions can be as competitive as \( P(\hat{A}_r = S_{p_0} \mid p_0) \). I propose a WMF procedure to tackle the underfitting in Section 2.4.

In next section, I show asymptotic variable selection properties for \( \hat{\beta}^*_a \) and \( \tilde{\beta}^*_ae \) in general settings, from which variable selection consistency of the MF procedure can be deduced.

### 2.3.2 Asymptotic properties

Let \( \mathcal{A} = \{ j : \beta_{0j} \neq 0 \} \) be the true model. I assume following regularity conditions for subsequent theoretical studies:

(A1) Denote \( \zeta_{\min}(C) \) and \( \zeta_{\max}(C) \) the minimum and maximum eigenvalues of a positive definite matrix \( C \). I assume

\[ d \leq \zeta_{\min}(\frac{1}{n}X^TX) \leq \zeta_{\max}(\frac{1}{n}X^TX) \leq D, \]
where \(d\) and \(D\) are two positive constants.

(A2) \(p_n = n^\varrho, 0 \leq \varrho < 1\) and \(\gamma > \frac{\varrho}{1-\varrho}\). The last inequality is to ensure \((1-\varrho)(1+\gamma) > 1\) in (A3)–(A4). Moreover,

\[
\lim_{n \to \infty} \frac{p_n}{n} \frac{1}{\min_{j \in A} |\beta_{0j}|^2} \to 0.
\]

(A3) In adaptive Lasso,

\[
\lim_{n \to \infty} \frac{\lambda_n}{\sqrt{n}} \to 0, \quad \lim_{n \to \infty} \frac{\lambda_n}{\sqrt{n}} \frac{(1-\varrho)(1+\gamma)^{-1}}{2} \to \infty,
\]

and

\[
\lim_{n \to \infty} \left(\frac{\lambda_n}{\sqrt{n}}\right)^{\frac{1}{\gamma}} \frac{1}{\min_{j \in A} |\beta_{0j}|} \to 0.
\]

(A4) In adaptive elastic-net,

\[
\lim_{n \to \infty} \frac{\lambda_{n1}}{\sqrt{n}} \to 0, \quad \lim_{n \to \infty} \frac{\lambda_{n2}}{\sqrt{n}} \to 0,
\]

and

\[
\lim_{n \to \infty} \frac{\lambda_{n1}^+}{\sqrt{n}} \to 0, \quad \lim_{n \to \infty} \frac{\lambda_{n1}^+}{\sqrt{n}} \frac{(1-\varrho)(1+\gamma)^{-1}}{2} \to \infty,
\]

\[
\lim_{n \to \infty} \left(\frac{\lambda_{n1}^+}{\sqrt{n}}\right)^{\frac{1}{\gamma}} \frac{1}{\min_{j \in A} |\beta_{0j}|} \to 0.
\]

(A5) The errors \(\{\varepsilon_i, i = 1, \ldots, n\}\) are i.i.d. with mean 0 and variance \(\sigma^2 < \infty\).

Denote \(A_n^* = \{j : \hat{\beta}_{aj}^* \neq 0\}\) an adaptive Lasso estimate of \(A\) using paired bootstrap data. Let \(P^* = P(\cdot | E)\) and \(E^* = E(\cdot | E)\) where \(E = \sigma ((x_i, y_i), i = 1, \ldots, n)\). Then \(P^*(A_n^* = A | \lambda_n)\) indicates the conditional probability of \(A_n^* = A\) given \(E\) and \(\lambda_n\).
Theorem 2.1. Suppose conditions (A1)–(A3) and (A5) hold, then

$$\lim_{n \to \infty} P^*(A_n^* = A \mid \lambda_n) = 1.$$  

Moreover, let $\lambda_n'$ be another tuning parameter such that the adaptive Lasso estimator under $\lambda_n'$ is of dimension $r$, $p_0 < r < p_n$, then

$$\lim_{n \to \infty} P^*(A_n^* = M_r \mid \lambda_n') < 1,$$

where $M_r$ is any $r$-dimensional model.

Proofs of Theorem 2.1 are included in Section 2.8.1.

In adaptive Lasso, given a $\lambda_n$ is equivalent to given a dimension, but the converse is not true. One dimension can be mapped to numerous models, as a result to numerous tuning parameters. Fortunately however, the LARS algorithm enables us to map a dimension to an optimal $\lambda_n$. Recall the adaptive Lasso solution path from the LARS in top panel of Figure 2.1. Transition points (e.g. steps) from 0 to 10 corresponds to a sequence of $\lambda_n$’s:

$$\lambda_n(0) > \lambda_n(1) > \cdots > \lambda_n(10) = 0.$$  

Note that $\hat{\beta}_a(\lambda_n) = 0$ for $\lambda_n > \lambda_n(0)$ where $\hat{\beta}_a(\lambda_n)$ is the adaptive Lasso estimator under $\lambda_n$. By Theorem 5 in (Zou et al., 2007),

$$\lambda_n(m + 1) = \arg \min_{\lambda_n} \|y - X\hat{\beta}_a(\lambda_n)\|^2 + a_n \hat{df}(\lambda_n), \quad \lambda_n(m + 1) \leq \lambda_n < \lambda_n(m),$$

where $\hat{df}(\lambda_n)$ is the number of non-zero elements in $\hat{\beta}_a(\lambda_n)$ and $a_n$ is a positive sequence depending on $n$. It is worth mentioning that $\lambda_n(m + 1)$ is optimum in
\[ \lambda_n(m + 1), \lambda_n(m) \] by producing the minimum sum of squared errors (SSE) and the smallest model size concurrently.

Also note that the number of steps can exceed the full model size — different steps may have a same model size. Denote \( m_k \) the last step having a model size \( k \), and \( m'_k \) is another step having the same model size. The theorem also showed that

\[
\|y - X \hat{\beta}_a(\lambda_n(m_k))\|^2 < \|y - X \hat{\beta}_a(\lambda_n(m'_k))\|^2.
\]

Therefore, \( \lambda_n(m_k) \) is the overall optimum in \( \{ \lambda_n : \hat{df}(\lambda_n) = k, \lambda_n \in [0, \infty] \} \). So the LARS algorithm enables us to create a one-to-one map between a dimension \( k \) and the optimum \( \lambda_n(m_k) \),

\[
k \iff \lambda_n(m_k).
\]

It is easy to see that \( \lambda_n(m_{p_0}) \) will satisfy condition (A3). Hence, I have the following corollary from Theorem 2.1.

**Corollary 2.1.** Suppose conditions (A1)–(A2) and (A5) hold, then

\[
\lim_{n \to \infty} P^*(\mathcal{A}_n^* = \mathcal{A} \mid p_0) = 1,
\]

\[
\lim_{n \to \infty} P^*(\mathcal{A}_n^* = \mathcal{M}_r \mid r) < 1, \quad p_0 < r < p_n,
\]

where \( \mathcal{M}_r \) is any \( r \)-dimensional model.

This result can also be established for adaptive elastic-net. Denote \( T_n^* = \{ j : \hat{\beta}_{aej} \neq 0 \} \) an adaptive elastic-net estimate of \( \mathcal{A} \) using paired bootstrap data.
Corollary 2.2. Suppose conditions (A1)–(A2) and (A5) hold, then

\[
\lim_{n \to \infty} P^* (T_n^* = A \mid p_0) = 1,
\]

\[
\lim_{n \to \infty} P^* (T_n^* = M_r \mid r) < 1, \quad p_0 < r < p_n,
\]

where \( M_r \) is any \( r \)-dimensional model.

Proof. It can be proved by using the techniques for deriving Theorem 2.1, Corollary 2.1 and Theorem 2.2. I bypass here. \( \square \)

I now study the estimation properties for using residual bootstrap data. Denote \( T_n^* = \{ j : \bar{\beta}_{aej} \neq 0 \} \) an adaptive elastic-net estimator of \( A \) using residual bootstrap data.

Theorem 2.2. Suppose conditions (A1)–(A2) and (A4)–(A5) hold, then

\[
\lim_{n \to \infty} P^* (T_n^* = A \mid \lambda_{n1}^+) = 1.
\]

Moreover, let \( \lambda_{n1}' \) be another tuning parameter such that the adaptive elastic-net estimator under \( \lambda_{n1}' \) is of dimension \( r \), \( p_0 < r < p_n \), then

\[
\lim_{n \to \infty} P^* (T_n^* = M_r \mid \lambda_{n1}') < 1,
\]

where \( M_r \) is any \( r \)-dimensional model.

Proofs of Theorem 2.2 are included in Section 2.8.1. The LARS-EN algorithm for adaptive elastic-net estimations is an extension of the LARS algorithm, which shares the same properties of the LARS for deriving Corollaries 2.1–2.2. Hence I obtain the following corollary from Theorem 2.
Corollary 2.3. Suppose conditions (A1)–(A2) and (A5) hold, then

$$\lim_{n \to \infty} P^*(T^*_n = \mathcal{A} \mid p_0) = 1,$$

$$\lim_{n \to \infty} P^*(T^*_n = \mathcal{M}_r \mid r) < 1, \quad p_0 < r < p_n,$$

where $\mathcal{M}_r$ is any $r$-dimensional model.

This result can also be established for adaptive Lasso. Denote $\mathcal{A}^*_n = \{j : \hat{\beta}_{aj}^* \neq 0\}$ an adaptive Lasso estimate of $\mathcal{A}$ using residual bootstrap data.

Corollary 2.4. Suppose conditions (A1)–(A2) and (A5) hold, then

$$\lim_{n \to \infty} P^*(\mathcal{A}^*_n = \mathcal{A} \mid p_0) = 1,$$

$$\lim_{n \to \infty} P^*(\mathcal{A}^*_n = \mathcal{M}_r \mid r) < 1, \quad p_0 < r < p_n,$$

where $\mathcal{M}_r$ is any $r$-dimensional model.

Proof. Note that the adaptive Lasso estimator is a special case of the adaptive elastic-net estimator with $\lambda_{n2} = 0$. Theorem 2.2 holds automatically for $\mathcal{A}^*_n$, from which Corollary 2.4 can be deduced. \qed

Variable selection consistency of the MF procedure can then be deduced from Corollaries 2.1–2.4.

Corollary 2.5. Suppose conditions (A1)–(A2) and (A5) hold. Then the MF procedure is variable selection consistent, e.g.

$$\lim_{n \to \infty} P(M_{r^*} = \mathcal{A}) = 1,$$
where \( M_{\star} \) is the model selected from the MF procedure.

Proof. By definition, \( A^*_{n} \) is an adaptive Lasso estimate of \( A \) using paired or residual bootstrap data. It is easy to see that

\[
E^* \left( \frac{\text{MF}_j}{B} \right) = P^*(A_{n}^* = M_j \mid j), \quad \lim_{B \to \infty} \frac{\text{MF}_j}{B} = P^*(A_{n}^* = M_j \mid j).
\]

Combining with Corollaries 2.1 or 2.4,

\[
\lim_{n \to \infty} P(M_{\text{MF}_p^*} > \text{MF}_r) = 1, \quad p_0 < r < p_n.
\]

Thus the MF procedure for adaptive Lasso can consistently identify the true dimension and true model by selecting the maximum of MFs, \( j \in \{1, \ldots, p - 1\} \), with the highest dimension (if there is a tie). Similarly, Corollary 2.2 and 2.3 imply variable selection consistency of the MF procedure for adaptive elastic-net.

However, the MF procedure has potential issues in application. In Figure 2.1 (middle) excluding the full model case, the maximum occurs at dimension 1 instead of 3 although their MFs are both close to 1. In next section, I propose a WMF procedure to tackle the underfitting issue.

\section{2.4 The WMF Procedure}

\subsection{2.4.1 Methods}

The underfitting issue in MF procedure can be deduced from Corollaries 2.1–2.4. Take \( A^*_{n} \) for an example. Although it was shown that \( \lim_{n \to \infty} P^*(A_{n}^* = A \mid p_0) = \)
1, the conditional probability at some underfit dimensions can also reach one, e.g. \( \lim_{n \to \infty} P^*(A_n^* = M_r \mid r) = 1, 0 < r < p_0 \). Note that the tuning parameter leading to an underfit \( r \)-dimensional estimator, denoted as \( \lambda'_n \), fulfills \( \lambda'_n > \lambda_n \). Hence, the convergence rate of \( P^*(A_n^* = M_r \mid r) \) at some underfit dimensions can exceed the one at the true dimension. Therefore, the MF procedure would select an underfit model even with a sufficiently large \( n \).

In order to fix things, I introduce a weight to the MF procedure. An effective weight should be able to down-weight underfit MFs asymptotically, i.e. the weight is able to identify underfit dimensions and its effects does not vanish as \( n \to \infty \), without significantly up-weighting the overfit MFs.

Shao (1993) showed that the overall unconditional (on \( y \)) expected squared prediction error for the OLS estimator of \( \beta_0 \) under model \( \alpha \) is

\[
T_{\alpha,n} = \sigma^2 + n^{-1}p_\alpha \sigma^2 + \Delta_{\alpha,n},
\]

where \( p_\alpha \) indicates the size of \( \alpha \), \( \Delta_{\alpha,n} = \beta_0^T X^T (I - P_\alpha) X \beta_0 / n \), \( P_\alpha = X_\alpha (X_\alpha^T X_\alpha)^{-1} X_\alpha^T \), \( X_\alpha \) is a sub-matrix of \( X \) whose columns are indexed by the components of \( \alpha \) and \( I \) is an identity matrix.

When \( \alpha \) is a true or overfit model, it has \( X \beta_0 = X_\alpha \beta_\alpha \) and thus

\[
\Delta_{\alpha,n} = 0.
\]

However, if \( \alpha \) is an underfit model, then \( \Delta_{\alpha,n} > 0 \) for any fixed \( n \). He further
assumed that

$$\lim \inf_{n \to \infty} \Delta_{\alpha,n} > 0,$$

(2.13)

which is argued in the paper to be a minimal type of asymptotic model identifiability condition. Under assumption (2.13) and by (2.11)–(2.12),

$$\lim_{n \to \infty} \frac{T_{\nu,n}}{T_{\kappa,n}} > 1,$$

(2.14)

where $\nu$ is an underfit model and $\kappa$ is a true or overfit model. By (2.14) a formula inversely proportional to $T_{\alpha,n}$ will be an ideal choice for the weight.

Rao and Tibshirani (1997) proposed such a formula for estimating the posterior probability of the model size given the data

$$\hat{P}(j \mid y) = \frac{\exp[-\hat{T}_n(j)/c\sigma^2]}{\sum_{j=1}^{p} \exp[-\hat{T}_n(j)/c\sigma^2]},$$

(2.15)

where $\hat{T}_n(j)$ is an estimate of $T_{\alpha,n}$ using a $j$-dimensional model and $c, 1 \leq c \leq 2,$ is a constant. I use the multi-fold CV for $\hat{T}_n(j)$ and define

$$\text{WMF}_j = \hat{P}(j \mid y) \times \text{MF}_j.$$  

(2.16)

Figure 2.1 (bottom) shows the effects of weights in Example 1, which heavily punish underfit MFs and have little effect on true and overfit MFs. The WMF procedure then selects the dimension $r^*$ and model $M_{r^*}$ s.t.

$$r^* = \arg \max_{1 \leq j \leq p-1} \text{WMF}_j.$$
Recall that $\text{MF}_j/B$ is a bootstrap version estimate of the posterior probability of model $M_j$ given the data and dimension, i.e. $P(M_j \mid y, j)$, along with (2.16) it has

$$\text{WMF}_j = \hat{P}(j \mid y) \times \hat{P}(M_j \mid y, j) = \hat{P}(M_j \mid y).$$

Note that BIC is a Laplace approximation to $P(M_j \mid y)$ under a flat prior assumption and is variable selection consistent for adaptive Lasso (Wang and Leng, 2007; Wang et al., 2009), but no convergence rate has been studied. Simulation studies in Section 2.5 show that BIC has a much slower empirical convergence rate than the WMF procedure.

### 2.4.2 Asymptotic properties

Next I show properties of the multi-fold CV using adaptive Lasso or adaptive elastic-net estimators. Then variable selection consistency of the WMF procedure can be established. Let $K$ be a fixed integer and suppose $n = Kt$. In multi-fold CV, one randomly divides a sample of $n$ observations into $K$ mutually exclusive subgroups $s_1, \ldots, s_K$ with each subgroup containing $t$ observations, and selects the model by minimizing the following sum of squared errors

$$\text{MCV}_M = \frac{1}{n} \sum_{i=1}^{K} \| y_{s_i} - X_{s_i, M} \hat{\beta}_{s_i, M} \|^2,$$

where $\hat{\beta}_{s_i, M}$ is an adaptive Lasso or adaptive elastic-net estimator under model $M$ using samples not in $s_i$. Let $\alpha$ and $\alpha'$ be the true or overfit models and $\nu$ be an underfit model. I assume following condition for asymptotic studies of the multi-fold CV procedure.
(A6) \( \sup_{t \to \infty} \sup_{s_i} \| t^{-1} X^T_{s_i,M} X_{s_i,M} - V_M \| = o(1) \), where \( V_M \) is a positive definite matrix.

**Theorem 2.3.** Suppose conditions (A1)–(A2) and (A5)–(A6) hold, then

1. the multi-fold CV for adaptive Lasso or adaptive elastic-net satisfies

\[
\lim_{n \to \infty} |MCV_\alpha - MCV_{\alpha'}| = \lim_{n \to \infty} \left| O_p\left( \frac{p_\alpha - p_{\alpha'}}{n} \right) \right| = 0,
\]

\[
\lim_{n \to \infty} MCV_{\nu} - MCV_\alpha \geq \frac{d \| \beta_{0\nu} \|^2}{2} + O_p \left( \| \beta_{0\nu} \| \sqrt{\frac{p_n}{n}} \right) - O_p \left( \frac{p_\alpha}{n} \right) > 0,
\]

2. model \( M_{r^*} \) selected from the WMF procedure fulfills

\[
\lim_{n \to \infty} P(M_{r^*} = A) = 1.
\]

Proofs of Theorem 2.3 are included in Appendix 2.8.1. Denote \( r' \) an underfit dimension. The ratio of \( \frac{\text{WMF}_{r'}}{\text{WMF}_{r}} \) is exponentially proportional to the bias term, \( \frac{d \| \beta_{0r'} \|^2}{2} \), which is larger than 0 and does not fade as \( n \to \infty \). This guarantees a good finite sample performance of the WMF procedure and a fast vanishing rate of underfitting, which will be confirmed in simulation studies in Section 2.5.

### 2.4.3 Computations

In adaptive elastic-net, \( \lambda_{n2} \) takes the same value in elastic-net for calculating the weights \( \omega_j \)'s, where tuning parameters are chosen by minimizing a two-dimensional BIC (Zou and Hastie, 2005). The rest computational costs then remain the same for adaptive Lasso and adaptive elastic-net, which are to compute a full solution
path against $\lambda_n$’s or $\lambda^+_{n1}$’s. Computational complexity of creating an entire adaptive Lasso solution path is of order $O(np^2_n)$ (Zou, 2006). It is of order $O(np^2_n + p^3_n)$ for adaptive elastic-net (Zou and Hastie, 2005). Since the optimal value often occurs at an early stage, we could stop the algorithms after $m, m < p_n$, steps. In this case, the computational cost reduces to $O(nm^2)$ for adaptive Lasso and $O(m^3 + nm^2)$ for adaptive elastic-net.

Computational cost of a WMF procedure is then $B$ times the cost of computing an adaptive Lasso or adaptive elastic-net solution path.

### 2.5 Empirical Studies

I now investigate empirical performances of the WMF procedure and show it outperforms the BIC, EBIC, GIC, SS, Cp, and 1se-CV (which is often recommended for variable selection) in a wide range of situations for both adaptive Lasso and adaptive elastic-net. The Cp did very poor in all scenarios, thus is excluded in the presentation.

In all simulations, data were generated from

$$y_i = x_i^T \beta + \sigma \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.17}$$

where $x_i \overset{iid}{\sim} N_{p_n}(0, \Sigma)$ and $\varepsilon_i \overset{iid}{\sim} N(0, 1)$. Let $p_n = O(n^\kappa)$ for some constant $\kappa$, $0 \leq \kappa < 1$, $n = 100, 300, 500$. Results were averaged over 100 times of replications.

#### 2.5.1 Simulations of the adaptive Lasso WMF procedure

Three scenarios were designed for the adaptive Lasso WMF procedure. In each scenario, $\Sigma(i, j) = 0.3^{|i-j|}$ and $\sigma = 3$. 
Scenario 1: *Fixed low dimension and moderate proportion of true covariates.* More specifically, set $p_n = 10$ and $\beta = (3, 1.5, 0, 0, 2, 0, \ldots)^T_{10}$. Then the proportion of true covariates is 0.3, and the signal to noise ratios (SNR) are respectively 2.03, 2 and 1.98 for various $n$.

Scenario 2: *Low dimension, moderate proportion of true covariates and weak signals for some true covariates.* Specifically, set $p_n = O(\sqrt{n})$, then $p_n$ equals to 10, 17, 22 accordingly. Let $p_0$ grow with $n$ as follows. Initially $p_0 = 3$ and $\beta = (3, 1.5, 0, 0, 2, 0, \ldots)^T$. Afterwards, $p_0$ increases by 1 for every 40-unit increment in $n$ and the new element equals to 1. As a result, the proportions of true covariates are respectively 0.3, 0.47, and 0.59, and the SNRs are 2, 2.85 and 3.69.

Scenario 3: *High dimension, sparse proportion of true covariates and relatively large signals for all true covariates.* In detail, set $p_n = O(n^{3/4})$, then $p_n$ equals to 32,
Figure 2.3: Results of scenario 2: (a) proportion of correctly specified models; (b) number of false non-zeros; (c) number of false zeros; (d) estimated model size.

72, 106 accordingly. Let $p_0$ grow in the same manner as in scenario 2, but the new elements equal to 2. Accordingly, the proportions of true covariates are 0.09, 0.11 and 0.12, and the SNRs are 2, 5.07, and 8.5.

*Paired bootstrapping* was used in the adaptive Lasso WMF procedure. Simulation results are summarized in Figures 2.2–2.4. In all scenarios, the proposed method has the highest degree of accuracy in identifying the true model and also enjoys a much faster convergence rate than other compared methods. The WMF procedure has an underfitting issue which vanishes quickly as $n$ increases. Other methods (except for the SS) however suffer from an overfitting issue. The sparser the model is, the more serious the issue tends to be. Performance of the SS relies on particular specifications of several unknown parameters. Although I have followed instructions in Meinshausen
Figure 2.4: Results of scenario 3: (a) proportion of correctly specified models; (b) number of false non-zeros; (c) number of false zeros; (d) estimated model size.

and Bühlmann (2010) for setting those parameters throughout the simulations, its performance remains erratic and unsatisfactory.

Simulations for using residual bootstrapping in the adaptive Lasso WMF procedure were also conducted. Results are presented in Section 2.8.2, which are similar to those in above paired bootstrapping simulations.

2.5.2 Simulations of the adaptive elastic-net WMF procedure

I also designed three scenarios for the adaptive elastic-net WMF procedure, each of which mimics a typical structure in applications. Since the adaptive elastic-net fits data with grouping effects, in following simulations true covariates will be added in
blocks with size 3. The SS is excluded due to its poor performance.

Scenario 4: Low dimension, moderate proportion of true covariates, weak signals for some true covariates and moderate correlations between covariates. More specifically, let $\Sigma(i, j) = 0.5|i-j|$, $\sigma = 3$, and $p_n = O(\sqrt{n})$. Initially I have one block of true covariates, then $p_0 = 3$. Elements of $\beta$ in the block equal to 2, the rest are 0. Afterwards, I add 1 block of true covariates for every 200-unit increment in $n$ and the new elements equal to 1. Respectively, the proportions of true covariates are 0.3, 0.35 and 0.41, and the SNRs are 2.45, 3.72 and 3.67.

Scenario 5: High dimension, sparse proportion of true covariates, relatively large signals for all true covariates and moderate correlations between covariates. In detail, let $\Sigma(i, j) = 0.5|i-j|$, $\sigma = 5$, and $p_n = O(n^{3/4})$. Initially set $p_0 = 6$. Then true
Figure 2.6: Results of scenario 5: (a) proportion of correctly specified models; (b) number of false non-zeros; (c) number of false zeros; (d) estimated model size.

covariates follow the same adding scheme as in scenario 4. All non-zero elements in $\beta$ equal to 2. Respectively, the proportions of true covariates are 0.19, 0.13 and 0.11, and the SNRs are 1.79, 3.09 and 3.51.

Scenario 6: High dimension, sparse proportion of true covariates, relatively large signals for all true covariates and high correlations between grouped covariates. Specifically, let $\sigma = 5$ and $p_n = O(n^{3/4})$. True covariates follow the same adding scheme as in scenario 5, all non-zero elements in $\beta$ equal to 2. Moreover, true covariates within each block have correlations almost 1, while true covariates between the blocks have correlation 0. All noise covariates are i.i.d from $N(0, 1)$. Respectively, the proportions of true covariates are 0.19, 0.13 and 0.11, and the SNRs are 2.84, 4.35 and 5.75.

Residual bootstrapping was used in the adaptive elastic-net WMF procedure. Sim-
Figure 2.7: Results of scenario 6: (a) proportion of correctly specified models; (b) number of false non-zeros; (c) number of false zeros; (d) estimated model size.

Simulation results are summarized in Figures 2.5–2.7. In scenarios 4 and 5, the proposed method has the best performance over other compared methods: on average the highest degree of accuracy in identifying the true model; a faster convergence rate; underfitting vanishes quickly. On the other hand, other methods suffer from an overfitting issue. The sparser the model is, the more serious the issue tends to be. In scenario 6, all methods do equally well because the adaptive elastic-net fits the data well with highly grouped effects.

Simulation results for using paired bootstrapping in the adaptive elastic-net WMF procedure are presented in Section 2.8.2, which are similar to those in above residual bootstrapping simulations.
2.5.3 Leukemia data analysis

I now demonstrate the WMF procedure in a real data application. The leukemia data (Golub et al., 1999) contained $p_n = 7129$ genes and $n = 72$ samples. It had 38 out of the 72 samples from the training dataset with 27 ALL’s (acute lymphoblastic leukemia) and 11 AML’s (acute myeloid leukemia). The remaining 34 samples were from the test dataset with 20 ALL’s and 14 AML’s. The goal of this analysis is to identify a subset of genes that can accurately predict the type of leukemia for future data. Similar to Zou and Hastie (2005), I coded the type of leukemia as a binary response variable, denoted as $y$, and defined the classification function as $I(\hat{y} > 0.5)$, where $I(\cdot)$ is the indicator function.

To improve computational efficiency, I selected 1000 candidate genes as the predictors using the sure independence screening (SIS) procedure (Fan and Lv, 2008). The adaptive Lasso and adaptive elastic-net were then applied to explore the data. The screening and variable selection were carried out on the training dataset, while classification errors were examined on the test dataset. Both the LARS and LARS-EN algorithms were stopped after 200 steps of estimation to further reduce the computational costs. Note that since the optimal steps selected by various types of methods were much smaller than the stopping step, this strategy did affect the variable selection.

Classification results were summarized in Tables 2.1–2.2. For adaptive Lasso, although the Cp, CV and BIC had the minimal classification errors for both training and test datasets, the WMF had classification errors close to the minimum using the least number of genes. For adaptive elastic-net, the WMF had the minimal classification errors for both training and test datasets using the least number of
Table 2.1: The leukemia classification using adaptive Lasso

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Ten-fold CV error</th>
<th>Test error</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMF</td>
<td>0/38</td>
<td>5/34</td>
<td>5</td>
</tr>
<tr>
<td>CV</td>
<td>0/38</td>
<td>4/34</td>
<td>13</td>
</tr>
<tr>
<td>Cp</td>
<td>0/38</td>
<td>4/34</td>
<td>18</td>
</tr>
<tr>
<td>BIC</td>
<td>0/38</td>
<td>4/34</td>
<td>18</td>
</tr>
<tr>
<td>EBIC</td>
<td>1/38</td>
<td>6/34</td>
<td>5</td>
</tr>
<tr>
<td>GIC</td>
<td>1/18</td>
<td>6/34</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 2.2: The leukemia classification using adaptive elastic-net

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Ten-fold CV error</th>
<th>Test error</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMF</td>
<td>0/38</td>
<td>4/34</td>
<td>10</td>
</tr>
<tr>
<td>CV</td>
<td>1/38</td>
<td>6/34</td>
<td>42</td>
</tr>
<tr>
<td>Cp</td>
<td>1/38</td>
<td>6/34</td>
<td>36</td>
</tr>
<tr>
<td>BIC</td>
<td>1/38</td>
<td>7/34</td>
<td>34</td>
</tr>
<tr>
<td>EBIC</td>
<td>1/38</td>
<td>7/34</td>
<td>34</td>
</tr>
<tr>
<td>GIC</td>
<td>1/38</td>
<td>7/34</td>
<td>21</td>
</tr>
</tbody>
</table>

genes. Thus we conclude that the WMF procedure is able to identify the set of “important” genes that largely improve the prediction accuracy.

2.6 Extensions

2.6.1 Generalized linear models

In this section, I investigate extensions of the WMF procedure to GLMs, which has the following generic density function (McCullagh and Nelder, 1989)

\[ f(y \mid x, \beta) = h(y) \exp(yx^T \beta - \phi(x^T \beta)). \]
Zou (2006) had extended the adaptive Lasso to GLMs. Its estimator, $\hat{\beta}_a$, is obtained by maximizing the penalized log-likelihood,

$$
\hat{\beta}_a = \arg \min_\beta \sum_{i=1}^{n} (-y_i x_i^T \beta + \phi(x_i^T \beta)) + \lambda_n \sum_{j=1}^{p} \hat{w}_j |\beta_j|,
$$

where $\hat{w}_j = 1/|\tilde{\beta}_j|^{\gamma}$, $\gamma > 0$ and $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$ is the maximum likelihood estimator. Under certain regularity conditions, $\hat{\beta}_a$ was shown to enjoy the oracle properties.

The generalization of Multi-fold CV to GLMs is straightforward (Shao, 1993). Define,

$$
MCV_\alpha = \frac{1}{n} \sum_{i=1}^{k} Q(y_{s_i}, \hat{y}_{s_i; \alpha}),
$$

where $Q(\cdot, \cdot)$ is a loss function, $\hat{y}_{s_i; \alpha}$ is the prediction of $y_{s_i}$ under model $\alpha$ using samples not in $s_i$.

Then I can extend the WMF procedure to GLMs for adaptive Lasso. In this case, I draw $B$ paired bootstrap samples in step 1 of Algorithm 1. Note that the LARS algorithm does not fit for GLMs, but we can use the coordinate descent algorithm (Friedman et al., 2007) instead, which generates a solution path similar to the LARS. Hence in step 2, I use the coordinate descent algorithm to fit each bootstrap data. The rest remain the same. Asymptotic properties of the adaptive Lasso WMF procedure for GLMs can also be established by using some similar techniques for showing Theorem 2.1 in this paper and Theorem 4 in (Zou, 2006).

I demonstrate this extension through one simple example, where binary responses were generated from the logistic regression model

$$
P(y_i | x_i) = \frac{1}{1 + exp(-x_i^T \beta)}, \quad i = 1, \ldots, n,
$$
where $x_i \overset{iid}{\sim} N_{10}(0, \Sigma)$, $\Sigma(i, j) = 0.3|i-j|$, and $\beta = (3, 1.5, 0, 0, 2, 0, \ldots)^T_{10}$. Simulation results were averaged over 100 times of replications and summarized in Figure 2.8. It showed that the WMF procedure was much more accurate in variable selection and enjoyed a faster convergence rate than other compared methods.

Extension of the adaptive elastic-net WMF procedure to GLMs is similar. Define the adaptive elastic-net estimator for GLMs as

$$
\hat{\beta}_{ae} = (1 + \frac{\lambda_{n2}}{n}) \left\{ \arg \min_{\beta} \sum_{i=1}^{n} (-y_i x_i^T \beta + \phi(x_i^T \beta)) + \lambda_{n2} \sum_{j=1}^{p_n} |\beta_j|^2 + \lambda_{n1} \sum_{j=1}^{p_n} w_j |\beta_j| \right\}, \quad (2.18)
$$

where $w_j = |\hat{\beta}_{ej}|^{-\gamma}$, $\gamma > 0$ and $\hat{\beta}_{e} = (\hat{\beta}_{e1}, \ldots, \hat{\beta}_{ep_n})^T$ was defined in (2.18) with $\hat{w}_j = 1$. 

Figure 2.8: Results of the GLM example: (a) proportion of correctly specified models; (b) number of false non-zeros; (c) number of false zeros; (d) estimated model size.
for all $j$’s. The rest follow the same procedures for extension of the adaptive Lasso WMF procedure.

2.6.2 Ultra-high dimensional data

In this section, I discuss applications of the WMF procedure to ultra-high dimensional data in which $p_n > n$. Fan and Lv (2008) proposed the sure independence screening (SIS) method for ultra-high dimensional data to reduce their dimensionality to a moderate scale, $d_n$, s.t. $d_n < n$. Afterwards a lower dimensional estimation method such as the SCAD can be applied to the reduced data. This process was called SIS+SCAD. Under some regularity conditions, they showed that the SIS has an exponentially small probability to omit true features and the SIS+SCAD retains the oracle properties if $d_n = o_p(n^{1/3})$. By replacing the SCAD with adaptive elastic-net, the new procedure was referred to SIS+AEnet (Zou and Zhang, 2009), which holds the oracle properties if $d_n = O_p(n^{\varrho})$, $0 \leq \varrho < 1$. Here I recommend to combine SIS with the WMF procedure when $p_n > n$. I first use the SIS to reduce the dimensionality to $d_n, d_n < n$, then apply the WMF procedure to the reduced data. I call this procedure SIS+WMF.

**Corollary 2.6.** Suppose conditions for Theorem 1 in Fan and Lv (2008) and Theorem 3 in this thesis hold. Let $d_n = n^{\varrho}, 0 \leq \varrho < 1$. Then the SIS+WMF procedure is variable selection consistent.

Note that Corollary 2.6 is a direct conclusion of Theorem 1 in Fan and Lv (2008) and Theorem 3 in this thesis.
2.7 Discussions

I proposed a prediction-weighted maximal frequency procedure to estimate the amount of regularization for adaptive Lasso and adaptive elastic-net. Asymptotic properties were studied with a diverging \( p_n \).

Central idea of the WMF procedure is the importance of conditioning on dimension, which mitigates overfitting. Underfitting can then be handled by using prediction-based weights estimated by multi-fold cross-validation. This simple recipe can also be applied to other regularization methods, say the SCAD and fused Lasso, making the WMF procedure a unified model selection criterion in regularization problems. However, asymptotic properties have yet to be studied, which will be a future topic.

2.8 Appendix

2.8.1 Proofs of Lemma 2.1 and Theorems 2.1-2.3

Proof of Lemma 2.1. Assume \(| \beta_i > | \beta_j | \) and \(| \beta_i | = - | \beta_j | = m \sigma, m > 0 \). We have 4 cases for \( \beta_i, \beta_j \)

\[
\beta_i = \begin{cases} 
\beta_j + m \sigma \text{ or } - \beta_j - m \sigma, & \beta_j \geq 0, \\
- \beta_j + m \sigma \text{ or } \beta_j - m \sigma, & \beta_j < 0.
\end{cases}
\]
Let $Z_i = \beta_i + x_i^T \varepsilon \sim N(\beta_i, \sigma^2)$ and $Z_j = \beta_j + x_j^T \varepsilon \sim N(\beta_j, \sigma^2)$. We have

$$
\begin{align*}
P(| Z_i | \leq z) &= \Phi\left(\frac{z - \beta_i}{\sigma}\right) + \Phi\left(\frac{z + \beta_i}{\sigma}\right) - 1, \quad z \geq 0, \\
P(| Z_j | \leq z) &= \Phi\left(\frac{z - \beta_j}{\sigma}\right) + \Phi\left(\frac{z + \beta_j}{\sigma}\right) - 1, \quad z \geq 0.
\end{align*}
$$

Consider case 1: $\beta_j \geq 0$ and $\beta_i = \beta_j + m \sigma$, $m > 0$.

Let $k$ be a positive constant. The point $\beta_j + k \sigma$ separates the domain of $Z_i$ and $Z_j$ into two parts: $(-\infty, \beta_j + k \sigma]$ and $[\beta_j + k \sigma, \infty]$. The cumulative probabilities of $Z_i$ and $Z_j$ in first part of the domain are respectively

$$
\begin{align*}
P(| Z_i | \leq \beta_j + k \sigma) &= \Phi(k - m) + \Phi(m + k + \frac{2\beta_j}{\sigma}) - 1, \\
P(| Z_j | \leq \beta_j + k \sigma) &= \Phi(k) + \Phi(k + \frac{2\beta_j}{\sigma}) - 1.
\end{align*}
$$

The probability $P(| Z_i | > | Z_j |)$ can then be calculated from

$$
P(| Z_i | > | Z_j |) = 1/2 P(| Z_i | \leq \beta_j + k \sigma, | Z_j | \leq \beta_j + k \sigma) \\
+ 1/2 P(| Z_i | > \beta_j + k \sigma, | Z_j | > \beta_j + k \sigma) \\
+ P(| Z_i | > \beta_j + k \sigma, | Z_j | \leq \beta_j + k \sigma).
$$

After some simple deductions, we get,

$$
P(| Z_i | > | Z_j |) \\
= \frac{1}{2} \left\{ \Phi\left(k + \frac{2\beta_j}{\sigma}\right) + \Phi(k) - \Phi(k - m) - \Phi(k + m + \frac{2\beta_j}{\sigma}) \right\} + \frac{1}{2}, \quad (2.19)
$$
If \( m = 0 \) i.e. \( |\beta_i| = |\beta_j| \), from (2.19) we have

\[
P(\mid Z_i \mid > \mid Z_j \mid) = \frac{1}{2}.
\]

However if \( m > 0 \) i.e. \( |\beta_i| > |\beta_j| \),

\[
P(\mid Z_i \mid > \mid Z_j \mid) = \frac{1}{2} \left\{ \int_{k-m}^{k} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx - \int_{k+2\beta_j/\sigma + m}^{k+2\beta_j/\sigma} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx \right\} + \frac{1}{2}. \tag{2.20}
\]

Since \( m, k, \beta_j, \sigma > 0 \), we have

\[
\max \{ \mid k - m \mid, \mid k \mid \} < \max \{ \mid k + (2\beta_j)/\sigma \mid, \mid k + (2\beta_j)/\sigma + m \mid \}.
\]

Note that two integrals in (2.20) have equal length of integral intervals. Moreover the integral function is an monotonically decreasing function of \( x \) for \( x \geq 0 \), and monotonically increasing for \( x < 0 \). Hence

\[
\int_{k-m}^{k} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx - \int_{k+2\beta_j/\sigma + m}^{k+2\beta_j/\sigma} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx > 0. \tag{2.21}
\]

Combining (2.20) with (2.21), we get

\[
P(\mid Z_i \mid > \mid Z_j \mid) > \frac{1}{2}.
\]

Other three cases can be proved in the same way. I avoid the repetitions here. \( \square \)

**Proof of Theorem 2.1.** By Zou and Zhang (2009), \( \hat{\beta}_a \) enjoys the oracle properties under certain regularity conditions. And \( \hat{\beta}_a^* \) is a paired bootstrap analog of \( \hat{\beta}_a \) by
replacing \((X, y)\) with \((X^*, y^*)\) in estimation. To simplify notations in the proof, I drop the subscript ‘a’ in \(\hat{\beta}_a\) and \(\hat{\beta}_a^*\).

By the KKT regularity conditions, \(\hat{\beta}^*\) is the unique solution of adaptive Lasso given \((X^*, y^*)\) if

\[
\begin{aligned}
\begin{cases}
X_j^T (y^* - X^* \hat{\beta}^*) = \lambda_n \omega_j sgn(\hat{\beta}_j^*), & \hat{\beta}_j^* \neq 0 \\
|X_j^T (y^* - X^* \hat{\beta}^*)| < \lambda_n \omega_j, & \hat{\beta}_j^* = 0
\end{cases}
\end{aligned}
\]

where \(X_j^*\) is the \(j\)th column of \(X^*\) and

\[
sgn(x) = \begin{cases}
1, & x > 0, \\
0, & x = 0, \\
-1, & x < 0.
\end{cases}
\]

Let \(\tilde{s}_A = (\omega_j sgn(\hat{\beta}_j), j \in A)^T\) and \(\hat{\beta}_A^* = (X_A^* X_A^*)^{-1}(X_A^* y^* - \lambda_n \tilde{s}_A)\). I show that \((\hat{\beta}_A^*, 0)\) satisfies (2.22) with probability tending to 1, which is equivalent to prove

\[
\begin{aligned}
\begin{cases}
sgn(\hat{\beta}_j)(\hat{\beta}_j - \hat{\beta}_j^*) < |\hat{\beta}_j|, & j \in A, \\
|X_j^T (y^* - X^* \hat{\beta}_A)| < \lambda_n \omega_j, & j \notin A,
\end{cases}
\end{aligned}
\]

where the first inequation implies \(sgn(\hat{\beta}_A^*) = sgn(\hat{\beta}_A)\).

Note that \(\omega_j = |\hat{\beta}_j|^{-\gamma}\), where \(\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)\) is an OLS or best ridge estimate of \(\beta_0\),

\[
\tilde{\beta}(\lambda_{n2}) = \arg \min_{\beta} \|y - X\beta\|^2 + \lambda_{n2} \sum_{j=1}^p |\beta_j|^2.
\]
By Theorem 3.1 in Zou and Zhang (2009),

\[
E\|\tilde{\beta}(\lambda_{n2}) - \beta_0\|^2 \leq 2\frac{\lambda_{n2}^2\|\beta_0\|^2 + np_nD\sigma^2}{(nd + \lambda_{n2})^2} = O_p\left(\frac{pn}{n}\right)
\]  

(2.24)

under assumption that \(\lim_{n \to \infty} \frac{\lambda_{n2}}{\sqrt{n}} = 0\). It is satisfied automatically for the OLS estimate.

Denote \(x_{iA}^\ast\) the \(i\)th row of \(X_A^\ast\), and \(\otimes\) the element-wise product. We have

\[
\hat{\beta}_A^* - \hat{\beta}_A = (X_A^T X_A)^{-1}(X_A^T y^* - X_A^T X_A \hat{\beta}_A - \lambda \tilde{s}_A)
\]

\[= (X_A^T X_A)^{-1} \left[ \sum_{i=1}^{n} x_{iA}^* (y_i^* - x_{iA}^T \hat{\beta}_A) - \lambda \omega_A \otimes sgn(\hat{\beta}_A) \right] (1 + o_p(1)).
\]

Hence under conditions (A1) and (A5),

\[
E^* \|\hat{\beta}_A^* - \hat{\beta}_A\|^2 \leq \frac{E^* \left\| \sum_{i=1}^{n} x_{iA}^* (y_i^* - x_{iA}^T \hat{\beta}_A) - \frac{\lambda \omega_A}{n} \otimes sgn(\hat{\beta}_A) \right\|^2}{\zeta_{\min}^2(X_A^T X_A)}
\]

\[= \frac{\sum_{i=1}^{n} E^* \left\| x_{iA}^* (y_i^* - x_{iA}^T \hat{\beta}_A) - \frac{\lambda \omega_A}{n} \otimes sgn(\hat{\beta}_A) \right\|^2}{\zeta_{\min}^2(X_A^T X_A)}
\]

\[= \frac{\sum_{i=1}^{n} \left\| x_{iA} (y_i - x_{iA}^T \hat{\beta}_A) - \frac{\lambda \omega_A}{n} \otimes sgn(\hat{\beta}_A) \right\|^2}{\zeta_{\min}^2(X_A^T X_A)}
\]

\[\leq \frac{1}{(nd)^2} \left[ \sum_{i=1}^{n} 2x_{iA}^T x_{iA} (y_i - x_{iA}^T \hat{\beta}_A)^2 + \frac{2\lambda^2 \|\omega_A\|^2}{n} \right]
\]

\[\leq \frac{2p_0 D\sigma^2}{nd^2} + \frac{2\lambda^2 \|\omega_A\|^2}{n^3 d^2}.
\]

Let \(\psi = \min_{j \in A} |\beta_{0j}|\), \(\tilde{\psi} = \min_{j \in A} |\tilde{\beta}_j|\) and \(\hat{\psi} = \min_{j \in A} |\hat{\beta}_j|\). Under conditions
(A1)-(A3) and (A5), the first inequation in (2.23) can be proved by

\[ P^* \left\{ \exists j \in A, \text{sgn}(\hat{\beta}_j)(\hat{\beta}_j - \beta_j^*) \geq |\hat{\beta}_j| \right\} \leq \sum_{j \in A} P^* \left\{ \text{sgn}(\hat{\beta}_j)(\hat{\beta}_j - \beta_j^*) \geq |\hat{\beta}_j|, \bar{\psi} > \psi/2, \hat{\psi} > \psi/2 \right\} + P(\hat{\psi} \leq \psi/2) + P(\hat{\psi} \leq \psi/2) + P(\hat{\psi} \leq \psi/2, \hat{\psi} \leq \psi/2) \leq 4E^* \left( \|\hat{\beta}_A - \beta A\|^2 I(\bar{\psi} > \psi/2) \right) \leq \frac{8}{\psi^2} \left( \frac{p_0 D \sigma^2}{n d^2} + \frac{\lambda_n^2 p_0 (\psi/2)^{-2\gamma}}{n^3 d^2} \right) + c_1 + c_2 + \min\{c_1, c_2\} \rightarrow 0, \]

where

\[ c_1 \leq P(\|\hat{\beta} - \beta_0\| \geq \psi/2) \leq \frac{4E\|\hat{\beta} - \beta_0\|^2}{\psi^2}. \]

By (2.24), it has

\[ c_1 \leq 8\frac{\lambda_n^2 \|\beta_0\|^2 + n p_n D \sigma^2}{\psi^2 (nd + \lambda_n^2)^2} = O_p \left( \frac{p_n}{n \psi^2} \right) \rightarrow 0 \]

Similarly,

\[ c_2 \leq P(\|\hat{\beta}_A - \beta_{0A}\| \geq \psi/2) \leq \frac{4E\|\hat{\beta}_A - \beta_{0A}\|^2 I(\bar{\psi} > \psi/2)}{\psi^2} + c_1. \]
By Theorem 3.1 in Zou and Zhang (2009),

\[
c_2 \leq 16 np_n D \sigma^2 + \frac{\lambda_n^2 p_0 (\psi/2)^{-2\gamma}}{\psi^2 n^2 d^2} + c_1
\]

\[
= O_p \left( \frac{p_n}{n \psi^2} \right) + O_p \left( \left( \frac{\lambda_n}{\sqrt{n \psi \gamma}} \right)^2 \frac{p_0}{n \psi^2} \right)
\]

\[
\to 0.
\]

(2.25)

For proof of the second inequation in (2.23), it suffices to show

\[
P^* \left\{ \exists j \notin A, |X_j^*(y^* - X_A^* \hat{\beta}_A^*)| \geq \lambda_n \omega_j \right\} \to 0.
\]

Since

\[
|X_j^*(y^* - X_A^* \hat{\beta}_A^*)| \leq |X_j^* (y^* - X_A^* \hat{\beta}_A)| + |X_j^* X_A^*(\hat{\beta}_A - \beta_A^*)|
\]

it follows that

\[
P^* \left\{ \exists j \notin A, |X_j^*(y^* - X_A^* \hat{\beta}_A^*)| \geq \lambda_n \omega_j \right\}
\]

\[
\leq \sum_{j \notin A} P^* \left\{ |X_j^* (y^* - X_A^* \hat{\beta}_A)| \geq (1 - \kappa) \lambda_n \omega_j \right\}
\]

\[
+ \sum_{j \notin A} P^* \left\{ |X_j^* X_A^*(\hat{\beta}_A - \beta_A^*)| \geq \kappa \lambda_n \omega_j \right\}
\]

\[
= B_1 + B_2,
\]

where \(\kappa, 0 < \kappa < 1\), is a constant.
For $B_1$, 

$$
\sum_{j \notin \mathcal{A}} E^* |X_j^*(y^* - X_{j\mathcal{A}}^* \hat{\beta}_\mathcal{A})|^2 = \sum_{j \notin \mathcal{A}} E^* \left| \sum_{i=1}^n x_{ij}^*(y_i^* - X_{i\mathcal{A}}^* \hat{\beta}_\mathcal{A}) \right|^2
$$

$$
= \sum_{j \notin \mathcal{A}} E^* \left[ \sum_{i=1}^n x_{ij}^2(y_i^* - X_{i\mathcal{A}}^* \hat{\beta}_\mathcal{A})^2 + \sum_{i \neq k} x_{ij}^*(y_i^* - X_{i\mathcal{A}}^* \hat{\beta}_\mathcal{A}) x_{kj}^*(y_k^* - X_{k\mathcal{A}}^* \hat{\beta}_\mathcal{A}) \right]
$$

$$
= \sum_{j \notin \mathcal{A}} \left\{ \sum_{i=1}^n x_{ij}^2(y_i - X_{i\mathcal{A}}\hat{\beta}_\mathcal{A})^2 + n(n-1) \left[ \frac{1}{n} \sum_{i=1}^n x_{ij}(y_i - X_{i\mathcal{A}}\hat{\beta}_\mathcal{A}) \right]^2 \right\}
$$

$$
= np_{\mathcal{A}^c}\sigma^2 + \frac{n-1}{n} \|X_{\mathcal{A}^c}^T(y - X_{\mathcal{A}}\hat{\beta}_\mathcal{A})\|^2
$$

$$
\le np_{\mathcal{A}^c}\sigma^2 + (n-1)p_{\mathcal{A}^c}D\sigma^2,
$$

where $p_{\mathcal{A}^c}$ indicates the size of $\mathcal{A}^c$. By (2.24), $\forall j \in \mathcal{A}^c$, $E|\tilde{\beta}_j|^2 \le E\|\tilde{\beta} - \beta_0\|^2 = O_p\left(\frac{p_{\mathcal{A}^c}}{n}\right)$, which indicates $|\tilde{\beta}_j| \le O_p\left(\frac{p_{\mathcal{A}^c}}{n}\right)^{1/2}$. Then under condition (A3), $B_1$ fulfills

$$
B_1 \le \sum_{j \notin \mathcal{A}} \frac{E^* |X_j^* (y^* - X_{j\mathcal{A}}^* \hat{\beta}_\mathcal{A})|^2}{(1 - \kappa)^2 \lambda_n^2 \omega_j^2}
$$

$$
\le \frac{np_{\mathcal{A}^c}\sigma^2 + (n-1)p_{\mathcal{A}^c}D\sigma^2}{(1 - \kappa)^2 \lambda_n^2 O_p\left(\frac{p_{\mathcal{A}^c}}{n}\right)^{-\gamma}}
$$

$$
= O_p\left(\frac{n}{\lambda_n^2 n^{1-\phi(1+\gamma)-1}}\right)
$$

$$
\to 0.
$$
Also since
\[
\sum_{j \notin A} E^* \left( |X_j^T X^*_A (\hat{\beta}_A - \hat{\beta}_A^*)|^2 I(\tilde{\psi} > \psi/2) \right)
= E^* \left( |X_A^T X^*_A (\hat{\beta}_A - \hat{\beta}_A^*)|^2 I(\tilde{\psi} > \psi/2) \right)
\leq (nD)^2 E^* \left( ||\hat{\beta}_A - \hat{\beta}_A^*||^2 I(\tilde{\psi} > \psi/2) \right) (1 + o_p(1))
\leq \left( \frac{2np_0 D^3 \sigma^2}{d^2} + \frac{2\lambda_0^2 p_0 (\psi/2)^{-2\gamma} D^2}{nd^2} \right) (1 + o_p(1)),
\]
we have for \( B_2 \),
\[
B_2 \leq \sum_{j \notin A} E^* \left( |X_j^T X^*_A (\hat{\beta}_A - \hat{\beta}_A^*)|^2 I(\tilde{\psi} > \psi/2) \right) \frac{\kappa^2 \lambda_0^2 \omega_j^2}{\lambda_n^2 O_p \left( \frac{p_n}{n} \right)^{-\gamma} \kappa^2 d^2}
\leq \left( \frac{2np_0 D^3 \sigma^2}{\lambda_n^2 O_p \left( \frac{p_n}{n} \right)^{-\gamma} \kappa^2 d^2} + \frac{2\lambda_0^2 p_0 (\psi/2)^{-2\gamma} D^2}{n \lambda_n^2 O_p \left( \frac{p_n}{n} \right)^{-\gamma} \kappa^2 d^2} \right) (1 + o_p(1)) + O_p \left( \frac{p_n}{n\psi^2} \right)
\leq O_p \left( \frac{n}{\lambda_n^2 n (1-\phi)(1+\gamma)-1} \right) + O_p \left( \frac{p_0}{n} \left( \frac{p_n}{n\psi^2} \right)^\gamma \right) + O_p \left( \frac{p_n}{n\psi^2} \right)
\to 0.
\]

Hence (2.23) is proved. I have shown that \( \hat{\beta}^* = (\hat{\beta}_A^*, 0) \) and \( \text{sgn}(\hat{\beta}_A^*) = \text{sgn}(\hat{\beta}_A) \) with probability tending to 1, where \( \hat{\beta}^* \) is the adaptive Lasso estimate using paired bootstrap data. Also it can be deduced from (2.25) that
\[
P(\min_{j \in A} |\hat{\beta}_j| > 0) \to 1.
\]
To sum up, we get \( \lim_{n \to \infty} P^*(\mathcal{A}_n^* = \mathcal{A} | \lambda_n) = 1 \).

I now prove \( \lim_{n \to \infty} P^*(\mathcal{A}_n^* = \mathcal{M}_r | \lambda_n') < 1 \), where \( \mathcal{M}_r \) is any \( r \)-dimensional model, \( p_0 < r < p_n \), and \( \lambda_n' \) is a tuning parameter such that the adaptive Lasso estimator under \( \lambda_n' \) is of dimension \( r \). Then \( \lambda_n' < \lambda_n \), hence \( \lambda_n'/\sqrt{n} \to 0 \). If it also satisfies \( \lim_{n \to \infty} \frac{\lambda_n^2 (1-\phi)(1+\gamma)-1}{n} \to \infty \), we would have \( P^*(\mathcal{A}_n^* = \mathcal{A} | \lambda_n') = 1 \) based on
previous proof, which contradicts with the definition of $\lambda_{n}'$. Therefore,

$$\lim_{n \to \infty} \frac{\lambda_{n}'^2 n^{(1-\phi)(1+\gamma)-1}}{n} < \infty.$$ 

To prove $\lim_{n \to \infty} P^*(A_n^* = M_r | \lambda_{n}') < 1$, by the KKT regularity conditions it suffices to show

$$P^* \left\{ \forall j \notin M_r, |X_j^T (y^* - X^* \hat{\beta}^*)| < \lambda_{n}' \omega_j \right\} < 1,$$

or equivalently

$$P^* \left\{ \exists j \notin M_r, |X_j^T (y^* - X^* \hat{\beta}^*)| \geq \lambda_{n}' \omega_j \right\} > 0. \quad (2.26)$$

Following previous proof, we get

$$P^* \left\{ \exists j \notin M_r, |X_j^T (y^* - X^* \hat{\beta}^*)| \geq \lambda_{n}' \omega_j \right\}$$

$$\leq \sum_{j \notin M_r} P^* \left\{ |X_j^T (y^* - X^* \hat{\beta})| \geq (1 - \kappa) \lambda_{n}' \omega_j \right\}$$

$$+ \sum_{j \notin M_r} P^* \left\{ |X_j^T X^* (\hat{\beta} - \hat{\beta}^*)| \geq \kappa \lambda_{n}' \omega_j \right\}$$

$$= B_1 + B_2.$$ 

However,

$$B_1 \leq \frac{np_{M_r^c} \sigma^2 + (n - 1)p_{M_r^c} D^2 \sigma^2}{(1 - \kappa)^2 \lambda_{n}'^2 n \frac{\sigma^2}{1 - \phi}} = O_p \left( \frac{n}{\lambda_{n}'^2 n^{(1-\phi)(1+\gamma)-1}} \right) \not\rightarrow 0,$$

as $n \to \infty$. Similarly, $\lim_{n \to \infty} B_2 \not\rightarrow 0$. Then (2.26) holds. $\square$
Lemma 2.2. Suppose conditions (A1) and (A5) hold and \( \lim_{n \to \infty} \lambda_n \sqrt{n} = 0 \) in ridge estimates. Then,

\[
E^*[X^T \varepsilon^*] = 0, \quad \lim_{n \to \infty} \text{Var}^*[X^T \varepsilon^*] = X^T X \sigma^2 \text{ with probability 1.}
\]

Proof. Assume \( \hat{\beta} \) is a ridge estimate of \( \beta_0 \),

\[
\hat{\beta} = \arg \min_\beta \|y - X \beta\|^2 + \lambda_n \|\beta\|^2.
\]

By (2.24), \( E\|\hat{\beta} - \beta_0\|^2 \leq O_p\left(\frac{\ln n}{n}\right) \). Calculate centered residuals \( \hat{\varepsilon} \),

\[
\hat{\varepsilon}_0 = y - X \hat{\beta}, \quad \hat{\varepsilon} = \hat{\varepsilon}_0 - \bar{\varepsilon}_0,
\]

where each entry of \( \bar{\varepsilon}_0 \), marked as \( \bar{\varepsilon}_0 \), is the mean of \( \hat{\varepsilon}_0 \). Denote \( \varepsilon^* = (\varepsilon^*_1, \ldots, \varepsilon^*_n)^T \) an i.i.d bootstrap sample from the empirical distribution that puts mass \( n^{-1} \) on each entry of \( \hat{\varepsilon} \).

By definition, we have

\[
E^*[X^T \varepsilon^*] = X^T E^*(\varepsilon^*) = 0,
\]

\[
\text{Var}^*[X^T \varepsilon^*] = X^T X \text{Var}^*(\varepsilon^*_1) = X^T X E^*(\varepsilon^*_1^2),
\]

and

\[
E^*(\varepsilon^*_1^2) = \frac{1}{n} \sum_{i=1}^n (\hat{\varepsilon}_{0i} - \bar{\varepsilon}_0)^2.
\]
In above equation,

\[ \bar{\varepsilon}_0 = \frac{1}{n} \sum_{i=1}^{n} \tilde{\varepsilon}_0 = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} x_i^T (\beta_0 - \hat{\beta}) + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i. \]

Moreover, by the sum of squares inequality,

\[ \left| \frac{1}{n} \sum_{i=1}^{n} x_i^T (\beta_0 - \hat{\beta}) \right| \leq \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[ x_i^T (\beta_0 - \hat{\beta}) \right]^2 \right\}^{1/2} \]
\[ = \left\{ \frac{1}{n} \|X(\beta_0 - \hat{\beta})\|_2^2 \right\}^{1/2} \]
\[ \leq \left\{ \frac{\zeta_{\max}(X^TX)}{n} \|\beta_0 - \hat{\beta}\|_2^2 \right\}^{1/2} \]
\[ = O_p \left( \sqrt{\frac{p_n}{n}} \right). \]

Hence,

\[ \bar{\varepsilon}_0 = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i + O_p \left( \sqrt{\frac{p_n}{n}} \right). \]

Let

\[ s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (\tilde{\varepsilon}_0 - \bar{\varepsilon}_0)^2 \quad \text{and} \quad \sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2, \]

where \( \bar{\varepsilon} = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \). I now prove \( s_n \to \sigma_n \) asymptotically.

Note that

\[ \lim_{n \to \infty} \sigma_n^2 = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 - \left( \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \right)^2 = E(\varepsilon_i^2) - (E(\varepsilon_i))^2 = \sigma^2 \]

with probability 1.

And by the sum of squares inequality,
\[(s_n - \sigma_n)^2 = \left\{ \frac{1}{n} \sum_{i=1}^{n} (\hat{\varepsilon}_{0i} - \bar{\varepsilon}_0)^2 \right\}^{1/2} - \left\{ \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 \right\}^{1/2} \]

\[= \frac{1}{n} \sum_{i=1}^{n} (\hat{\varepsilon}_{0i} - \bar{\varepsilon}_0)^2 + \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 - 2 \left[ \frac{1}{n} \sum_{i=1}^{n} (\hat{\varepsilon}_{0i} - \bar{\varepsilon}_0)^2 \right]^{\frac{1}{2}} \left[ \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 \right]^{\frac{1}{2}}
\]

\[\leq \frac{1}{n} \sum_{i=1}^{n} (\hat{\varepsilon}_{0i} - \bar{\varepsilon}_0)^2 + \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 - \frac{2}{n} \sum_{i=1}^{n} (\hat{\varepsilon}_{0i} - \bar{\varepsilon}_0)(\varepsilon_i - \bar{\varepsilon})\]

\[= \frac{1}{n} \sum_{i=1}^{n} \left[ (\hat{\varepsilon}_{0i} - \bar{\varepsilon}_0) - (\varepsilon_i - \bar{\varepsilon}) \right]^2
\]

\[= \frac{1}{n} \sum_{i=1}^{n} \left[ \hat{\varepsilon}_{0i} - \varepsilon_i - O_p\left(\sqrt{\frac{p_n}{n}}\right) \right]^2
\]

\[\leq \frac{1}{n} \|X(\beta_0 - \hat{\beta})\|^2 + O_p\left(\sqrt{\frac{p_n}{n}}\right) \frac{1}{\sqrt{n}} \|X(\beta_0 - \hat{\beta})\| + O_p\left(\frac{p_n}{n}\right)
\]

\[\leq \frac{\xi_{\max}(X^TX)}{n} \|\beta_0 - \hat{\beta}\|^2 + O_p\left(\sqrt{\frac{p_n}{n}}\right) \sqrt{\frac{\xi_{\max}(X^TX)}{n}} \|\beta_0 - \hat{\beta}\| + O_p\left(\frac{p_n}{n}\right)
\]

\[= O_p\left(\frac{p_n}{n}\right).
\]

Then \(\lim_{n \to \infty} s_n^2 = \sigma^2\) with probability 1. \(\square\)

**Proof of Theorem 2.2.** Let \((X, y^*)\) be a residual bootstrap sample, where \(y^* = X\hat{\beta} + \varepsilon^*\) and \(\hat{\beta}\) is the ridge estimator. Define

\[
\tilde{\beta}^* = (1 + \frac{\lambda n^2}{n}) \left\{ \arg \min_{\beta} \|y^* - X\beta\|^2 + \lambda n^2 \sum_{j=1}^{p_n} |\beta_j|^2 + \lambda_n^+ \sum_{j=1}^{p_n} \omega_j|\beta_j| \right\}, \tag{2.27}
\]

where I dropped the subscript ‘ae’ in \(\tilde{\beta}_{ae}^*\) for simplicity.

Let

\[
\tilde{\beta}_{\mathcal{A}}^* = \arg \min_{\beta} \|y^* - X_{\mathcal{A}}\beta\| + \lambda n^2 \sum_{j \in \mathcal{A}} |\beta_j|^2 + \lambda_n^+ \sum_{j \in \mathcal{A}} \omega_j|\beta_j|,
\]
I prove \((1 + \frac{\lambda_n}{\tilde{n}})^2\) is the solution to (2.27) with probability tending to 1. By the KKT regularity conditions, this suffices to show

\[
P^* \left\{ \forall j \in A, |X_j^T(y^* - X_A\tilde{\beta}_A^*)| < \lambda_{n1}^+\omega_j \right\} \to 1,
\]

or equivalently

\[
P^* \left\{ \exists j \notin A, |X_j^T(y^* - X_A\tilde{\beta}_A^*)| \geq \lambda_{n1}^+\omega_j \right\} \to 0. \tag{2.28}
\]

Note that \(\omega_j = |\hat{\beta}_{ej}|^{-\gamma}\) where \(\hat{\beta}_e = (\hat{\beta}_{e1}, \ldots, \hat{\beta}_{en})^T\) is the elastic-net estimator defined in (2.6). By Theorem 3.1 in Zou and Zhang (2009),

\[
E\|\hat{\beta}_e - \beta_0\|^2 \leq 4\frac{\lambda_n^2\|\hat{\beta}_0\|^2 + np_nD\sigma^2 + \lambda_n^2p_n}{(nd + \lambda_n^2)^2} = O_p\left(\frac{p_n}{n}\right) \tag{2.29}
\]

under condition (A4).

Let \(\psi = \min_{j \in A} |\hat{\beta}_{0j}|\) and \(\tilde{\psi} = \min_{j \in A} |\hat{\beta}_{ej}|\). Then

\[
P^* \left\{ \exists j \notin A, |X_j^T(y^* - X_A\tilde{\beta}_A^*)| \geq \lambda_{n1}^+\omega_j \right\}
\leq P^* \left\{ \exists j \notin A, |X_j^T(y^* - X_A\tilde{\beta}_A^*)| \geq \lambda_{n1}^+\omega_j, \tilde{\psi} > \psi/2 \right\} + P\{\tilde{\psi} \leq \psi/2\}
\leq \sum_{j \notin A} P^* \left\{ |X_j^T(y^* - X_A\tilde{\beta}_A^*)| \geq \lambda_{n1}^+\omega_j, \tilde{\psi} > \psi/2 \right\} + P\{\tilde{\psi} \leq \psi/2\}
= B_1 + B_2.
By (2.29) under condition (A4),

\[ B_2 = P\{ \tilde{\psi} \leq \psi/2 \} \leq P\{ \| \hat{\beta}_e - \beta_0 \| \geq \psi/2 \} \leq \frac{4E\| \hat{\beta}_e - \beta_0 \|^2}{\psi^2} \leq O_p( \frac{p_n}{\psi^2} ) \to 0. \]

Also by (2.29) \( \forall j \in A^c, E|\hat{\beta}_{ej}|^2 \leq E\| \hat{\beta}_e - \beta_0 \|^2 = O_p( \frac{p_n}{n} ) \), which indicates \( |\hat{\beta}_{ej}| \leq O_p( \frac{p_n}{n} )^{1/2} \). Hence

\[ B_1 \leq \frac{O_p( \frac{p_n}{n} )}{\lambda_n^2} E^*\left\{ \sum_{j \notin A} |X_j^T(y^* - X_A\hat{\beta}_A)|^2 I(\tilde{\psi} > \psi/2) \right\}. \]

Note that

\[
E^*\left\{ \sum_{j \notin A} |X_j^T(y^* - X^*_{A^c}\beta_A^*)|^2 \right\} \\
= E^*\left\{ \sum_{j \notin A} |X_j^T\left(X_A\hat{\beta}_A + X_{A^c}\hat{\beta}_{A^c} + \varepsilon^* - X_A\beta^*_A\right)|^2 \right\} \\
\leq 3E^*\|X_{A^c}A\hat{\beta}_A - \beta^*_A\|^2 + 3\|X_{A^c}A\hat{\beta}_{A^c}\|^2 + 3E^*\|X_{A^c}A\varepsilon^*\|^2 \\
\leq 3(nD)^2E^*\|\beta_A - \beta^*_A\|^2 + 3(nD)^2\|\hat{\beta}_{A^c}\|^2 + 3E^*\|X_{A^c}A\varepsilon^*\|^2.
\]

By (2.24),

\[ \| \hat{\beta}_{A^c} \|^2 \leq \| \hat{\beta} - \beta_0 \|^2 \leq O_p( \frac{p_n}{n} ). \quad (2.30) \]

I now study \( E^*\| \hat{\beta}_A - \beta^*_A \|^2 \). Let

\[ \beta^*_A(\lambda_n, 0) = \arg\min_{\beta} \| y^* - X_A\beta \| + \lambda_n \sum_{j \in A} \beta_j^2. \]
By using the same arguments for deriving (6.3) in Zou and Zhang (2009), it can easily be shown

\[
\| \hat{\beta}_A^* - \tilde{\beta}_A^*(\lambda_{n2}, 0) \| \leq \frac{\lambda_{n1}^+ \| \omega_A \|}{\zeta_{\text{min}}(X_A^T X_A) + \lambda_{n2}}. \tag{2.31}
\]

On the other hand,

\[
\tilde{\beta}_A^*(\lambda_{n2}, 0) - \hat{\beta}_A = (X_A^T X_A + \lambda_{n2} I)^{-1}(-\lambda_{n2} \hat{\beta}_A + X_A^T X_A \tilde{\beta}_A + X_A^T \epsilon^*),
\]

by Lemma 2.2,

\[
E^* \| \tilde{\beta}_A^*(\lambda_{n2}, 0) - \hat{\beta}_A \|^2 \leq 3 \frac{\lambda_{n2}^2 \| \hat{\beta}_A \|^2 + \| X_A^T X_A \hat{\beta}_A \|^2 + E^* \| X_A^T \epsilon^* \|^2}{(\zeta_{\text{min}}(X_A^T X_A) + \lambda_{n2})^2}
\]

\[
\leq 3 \frac{\lambda_{n2}^2 \| \hat{\beta}_A \|^2 + (nD)^2 \| \tilde{\beta}_A \|^2 + np_0 D \sigma^2}{(nd + \lambda_{n2})^2}. \tag{2.32}
\]

By assembling (2.30)–(2.32), we get

\[
E^* \| \hat{\beta}_A - \tilde{\beta}_A^* \|^2 \leq 2E^* \| \hat{\beta}_A - \tilde{\beta}_A^*(\lambda_{n2}, 0) \|^2 + 2E^* \| \tilde{\beta}_A^*(\lambda_{n2}, 0) - \hat{\beta}_A \|^2
\]

\[
\leq 6 \frac{\lambda_{n1}^+ \| \omega_A \|^2 + \lambda_{n2}^2 \| \hat{\beta}_A \|^2 + O_p(np_n D^2) + np_0 D \sigma^2}{(nd + \lambda_{n2})^2}.
\]

And

\[
E^* \left\{ \sum_{j \notin A} |X_j^T (y^* - X_A \tilde{\beta}_A)|^2 I(\psi > \psi/2) \right\} \leq 3O_p(np_n D^2) + 3np_{A^c} D \sigma^2 + 18n^2 D^2 \frac{\lambda_{n1}^2 \lambda_{n2}^2 p_0 (\psi/2)^{-2\gamma} + \lambda_{n2} \| \hat{\beta}_A \|^2 + O_p(np_n D^2) + np_0 D \sigma^2}{(nd + \lambda_{n2})^2}
\]

\[
= O_p(np_n) + O_p(\psi^{-2\gamma} \lambda_{n1}^2 p_0).
\]
Then under conditions (A1)–(A2) and (A4)–(A5),

\[
B_1 \leq O_p \left( \frac{p_n}{n} \right)^\gamma \left[ O_p \left( np_n \right) + O_p \left( \psi^{-2\gamma} n^{\lambda_{n1}^+} p_0 \right) \right]
\]
\[
\leq O_p \left( \frac{n}{\lambda_{n1}^+ n(1-\phi)(1+\gamma)-1} \right) + O_p \left( \frac{1}{\psi^2 n(1-\phi)(1+\gamma)-1} \right)
\]
\[
\rightarrow 0.
\]

Hence (2.28) is proved. So far I have shown that \( \tilde{\beta}^* = ((1 + \frac{\lambda_n}{n}) \tilde{\beta}_{A}, 0) \) with probability tending to 1, where \( \tilde{\beta}^* \) is the adaptive elastic-net estimate using residual bootstrap data. To prove \( \lim_{n \to \infty} P^* (T_n^* = A \mid \lambda_{n1}^+) = 1 \), I still need to show that \( P(\min_{j \in A} |\tilde{\beta}_{j^*}| > 0) \to 1 \).

Let \( \hat{\psi} = \min_{j \in A} |\hat{\beta}_j| \) and \( \tilde{\psi}^* = \min_{j \in A} |\tilde{\beta}_{j^*}| \). By (2.24),

\[
P(\hat{\psi} \leq \psi/2) \leq P(\|\hat{\beta} - \beta_0\| \geq \psi/2) \leq O_p \left( \frac{p_n}{n\psi^2} \right) \to 0.
\]

Hence \( P(\hat{\psi} > \psi/2) \to 1 \) as \( n \to \infty \) where \( \psi > 0 \). Under condition (A4),

\[
P(\tilde{\psi}^* \leq \hat{\psi}/2) \leq P(\tilde{\psi}^* \leq \hat{\psi}/2, \hat{\psi} > \psi/2) + P(\tilde{\psi} \leq \psi/2)
\]
\[
\leq P(\|\hat{\beta}_A - \tilde{\beta}_A^*\| \geq \hat{\psi}/2, \tilde{\psi} > \psi/2) + B_2
\]
\[
\leq \frac{16}{\psi^2} E^* (\|\hat{\beta}_A - \tilde{\beta}_A^*\|^2 I(\tilde{\psi} > \psi/2)) + B_2
\]
\[
\leq \frac{96 \lambda_{n1}^{+2} p_0 (\psi/2)^{-2\gamma} + \lambda_{n2}^2 \|\tilde{\beta}_A\|^2 + O_p (np_n D^2) + np_n D \sigma^2}{(nd + \lambda_n)^2} + B_2
\]
\[
= O_p \left( \left( \frac{\lambda_{n1}^{+}}{\sqrt{n\psi^\gamma}} \right)^2 \frac{p_0}{n\psi^2} + O_p \left( \frac{p_n}{n\psi^2} \right) \right)
\]
\[
\rightarrow 0.
\]
which indicates \( P(\tilde{\psi}^* > \hat{\psi}/2) \to 1 \) as \( n \to \infty \). To sum up, \( \lim_{n \to \infty} P(\tilde{\psi}^* > \psi/4) = 1 \).

Thus \( \lim_{n \to \infty} P^*(T_n^* = A \mid \lambda_{n1}^+) = 1 \) is proved.

I now prove \( \lim_{n \to \infty} P^*(T_n^* = M_r \mid \lambda'_{n1}) < 1 \), where \( M_r \) is any \( r \)-dimensional model, \( p_0 < r < p_n \), and \( \lambda'_{n1} \) is a tuning parameter such that the adaptive elastic-net estimator under \( \lambda'_{n1} \) is of dimension \( r \). Then \( \lambda'_{n1} < \lambda_{n1}^+ \), hence \( \lambda'_{n1}/\sqrt{n} \to 0 \). If it also satisfies \( \lim_{n \to \infty} \frac{\lambda_{n1}^2 n(1-\psi(1+\gamma)-1)}{n} \to \infty \), we would have \( P^*(T_n^* = A \mid \lambda'_{n1}) = 1 \) based on previous proof, which contradicts with the definition of \( \lambda'_{n1} \). Therefore,

\[
\lim_{n \to \infty} \frac{\lambda_{n1}^2 n(1-\psi(1+\gamma)-1)}{n} < \infty.
\]

To prove \( \lim_{n \to \infty} P^*(T_n^* = M_r \mid \lambda'_{n1}) < 1 \), by the KKT regularity conditions it suffices to show

\[
P^* \left\{ \forall j \notin M_r, |X_j^*^T (y^* - X_{M_r} \hat{\beta}_{M_r}^*)| < \lambda_{n1}^1 \omega_j \right\} < 1,
\]

or equivalently

\[
P^* \left\{ \exists j \notin M_r, |X_j^*^T (y^* - X_{M_r} \hat{\beta}_{M_r}^*)| \geq \lambda_{n1}^1 \omega_j \right\} > 0.
\]
By following the same arguments for showing (2.28), we get

\[ P^* \left\{ \exists j \notin \mathcal{M}_r, |X_j^*^T(y^* - X_{M_r} \hat{\beta}_{M_r}^*)| \geq \lambda'_{n_1}\omega_j \right\} \]

\[ \leq \frac{O_p \left( \frac{p_n}{n} \right)^\gamma}{\lambda'^2_{n_1}} \left\{ 3O_p(np_nD^2) + 3np_{M_r}D\sigma^2 
+ 18n^2D^2\lambda'^2_{n_1}\|\omega_{M_r}\|^2 + \lambda'^2_{n_2}\|\hat{\beta}_{M_r}\|^2 + O_p(np_nD^2) + np_{M_r}D\sigma^2 \right\} 
\]

\[ = O_p \left( \frac{n}{\lambda'^2_{n_1}n(1-\varphi)(1+\gamma)-1} \right) + O_p \left( \|\omega_{M_r}\| \left( \frac{p_n}{n} \right)^\gamma \right) \]

\[ \not\rightarrow 0. \]

**Lemma 2.3.** Suppose conditions (A1), (A5) and (A6) hold. Denote \( \alpha \) an overfit model including the true model, the adaptive elastic-net estimate \( \hat{\beta}_{s_i,\alpha} \) from the multi-fold CV then satisfies

\[ E\|\hat{\beta}_{s_i,\alpha} - \beta_{0\alpha}\|^2 \leq 4\lambda'^2_{n_2}\|\beta_{0\alpha}\|^2 + (n - t)p_\alpha D\sigma^2(1 + o_p(1)) + \lambda'^2_{n_1}E\|\omega_{\alpha}\|^2 
\]

\[ = O_p \left( \frac{p_n}{n} \right), \]

where the adaptive Lasso estimate is a special case with \( \lambda_{n_2} = 0 \).

**Proof.** Here I provide a proof for the adaptive Lasso estimator. The adaptive elastic-net estimator can be proved by using the same arguments for deriving Theorem 3.1 in Zou and Zhang (2009) and the strategies in below.

The adaptive Lasso estimator from the multi-fold CV is

\[ \hat{\beta}_{s_i,\alpha} = \arg\min_{\beta} \|Y_{s_i} - X_{s_i,\alpha}\beta\|^2 + 2\lambda'_{n_1}\sum_{j \in \alpha} \omega_j|\beta_j|, \]
which satisfies

$$\hat{\beta}^*_{s_i,\alpha} - \beta_{0\alpha} = (X^T_{s_i,\alpha} X_{s_i,\alpha})^{-1} \left( X^T_{s_i,\alpha} \varepsilon_{s_i} - \lambda'_{n1} \omega_{\alpha} \otimes \operatorname{sgn}(\hat{\beta}^*_{s_i,\alpha}) \right).$$

Hence,

$$E\|\hat{\beta}^*_{s_i,\alpha} - \beta_{0\alpha}\|^2 \leq \frac{2E\|X^T_{s_i,\alpha} \varepsilon_{s_i}\|^2 + 2\lambda'_{n1} E\|\omega_{\alpha}\|^2}{\zeta^2_{\min}(X^T_{s_i,\alpha} X_{s_i,\alpha})} \leq \frac{2\zeta_{\max}(X^T_{s_i,\alpha} X_{s_i,\alpha}) p\sigma^2 + 2\lambda'_{n1} E\|\omega_{\alpha}\|^2}{\zeta^2_{\min}(X^T_{s_i,\alpha} X_{s_i,\alpha})} \leq \frac{2(n - t)p\sigma^2(1 + o_p(1)) + 2\lambda'_{n1} E\|\omega_{\alpha}\|^2}{(n - t)^2d^2(1 + o_p(1))} = O_p\left(\frac{p\alpha}{n}\right).$$

The last equation holds because $\lambda'_{n1}$ continuously decreases from $\lambda^+_{n1}$ to 0 as $\alpha$ changes from the true model to full model.

\[\square\]

**Proof of Theorem 2.3.** I integrate the proof for adaptive elastic-net and adaptive Lasso. Denote $\alpha$ an overfit model including the true model. The MCV $\alpha$ is

$$\text{MCV}_\alpha = \frac{1}{n} \sum_{i=1}^{K} \|X_{s_i,\alpha} \beta_{0\alpha} + \varepsilon_{s_i} - X_{s_i,\alpha} \hat{\beta}^*_{s_i,\alpha}\|^2$$

$$= \frac{1}{n} \varepsilon^T \varepsilon + \frac{1}{n} \sum_{i=1}^{K} \|X_{s_i,\alpha} (\beta_{0\alpha} - \hat{\beta}^*_{s_i,\alpha})\|^2$$

$$+ \frac{2}{n} \sum_{i=1}^{K} (\beta_{0\alpha} - \hat{\beta}^*_{s_i,\alpha})^T X^T_{s_i,\alpha} \varepsilon_{s_i}.$$ (2.33)
By Lemma 2.3, the second term in (2.33) satisfies

\[ E\|X_{s_i,\alpha}(\beta_{0\alpha} - \hat{\beta}_{s_i,\alpha})\|^2 \leq \zeta_{\max}(X_{s_i,\alpha}^TX_{s_i,\alpha})E\|\beta_{0\alpha} - \hat{\beta}_{s_i,\alpha}\|^2 \]

\[ \leq t DO_p \left( \frac{p_\alpha}{n} \right) = O_p \left( \frac{tp_\alpha}{n} \right), \]

\[ \text{Var}\|X_{s_i,\alpha}(\beta_{0\alpha} - \hat{\beta}_{s_i,\alpha})\|^2 \leq E\|X_{s_i,\alpha}(\beta_{0\alpha} - \hat{\beta}_{s_i,\alpha})\|^4 \]

\[ \leq O_p \left( \frac{tp_\alpha}{n} \right)^2. \]

Hence,

\[ \|X_{s_i,\alpha}(\beta_{0\alpha} - \hat{\beta}_{s_i,\alpha})\|^2 \leq O_p \left( \frac{tp_\alpha}{n} \right). \]

\[ \frac{1}{n} \sum_{i=1}^K \|X_{s_i,\alpha}(\beta_{0\alpha} - \hat{\beta}_{s_i,\alpha})\|^2 \leq \frac{K}{n} O_p \left( \frac{tp_\alpha}{n} \right) = O_p \left( \frac{p_\alpha}{n} \right). \] (2.34)

The third term in (2.33) fulfills

\[ E[(\beta_{0\alpha} - \hat{\beta}_{s_i,\alpha})^TX_{s_i,\alpha}^T\varepsilon_{s_i}] = 0, \]

\[ E[(\beta_{0\alpha} - \hat{\beta}_{s_i,\alpha})^TX_{s_i,\alpha}^T\varepsilon_{s_i}]^2 \leq E\|\beta_{0\alpha} - \hat{\beta}_{s_i,\alpha}\|^2E\|X_{s_i,\alpha}^T\varepsilon_{s_i}\|^2 \]

\[ \leq O_p \left( \frac{p_\alpha}{n} \right) tp_\alpha D\sigma^2 \]

\[ = O_p \left( \frac{tp_\alpha^2}{n} \right). \]

Hence,

\[ (\beta_{0\alpha} - \hat{\beta}_{s_i,\alpha})^TX_{s_i,\alpha}^T\varepsilon_{s_i} \leq O_p \left( \sqrt{\frac{tp_\alpha^2}{n}} \right), \]

\[ \frac{2}{n} \sum_{i=1}^K (\beta_{0\alpha} - \hat{\beta}_{s_i,\alpha})^TX_{s_i,\alpha}^T\varepsilon_{s_i} \leq \frac{2K}{n} O_p \left( \sqrt{\frac{tp_\alpha^2}{n}} \right) = O_p \left( \frac{p_\alpha}{n} \right). \] (2.35)
By substituting (2.34)–(2.35) to (2.33), we have

$$MCV_\alpha = \frac{1}{n} \varepsilon^T \varepsilon + O_p \left( \frac{P_\alpha}{n} \right).$$

Let \( \alpha \) and \( \alpha' \) be two overfit models including the true model, then

$$\lim_{n \to \infty} |MCV_\alpha - MCV_{\alpha'}| = \lim_{n \to \infty} \left| O_p \left( \frac{P_\alpha - P_{\alpha'}}{n} \right) \right| = 0.$$

I now consider an underfit model \( \nu \). The MCV\( _\nu \) is

$$MCV_{\nu} = \frac{1}{n} \sum_{i=1}^{K} \| X_{s_i} \beta_0 + \varepsilon_{s_i} - X_{s_i,\nu} \hat{\beta}_{s_i,\nu} \|^2$$

$$= \frac{1}{n} \varepsilon^T \varepsilon + \frac{1}{n} \sum_{i=1}^{K} \| X_{s_i} \beta_0 - X_{s_i,\nu} \hat{\beta}_{s_i,\nu} \|^2$$

$$+ \frac{2}{n} \sum_{i=1}^{K} (X_{s_i} \beta_0 - X_{s_i,\nu} \hat{\beta}_{s_i,\nu})^T \varepsilon_{s_i}. \quad (2.36)$$

Let \( \hat{\beta}_{\nu} \) be an adaptive Lasso or adaptive elastic-net estimator under \( \nu \). The second term in (2.36) satisfies

$$\frac{1}{n} \sum_{i=1}^{k} \| X_{s_i} \beta_0 - X_{s_i,\nu} \hat{\beta}_{s_i,\nu} \|^2$$

$$\geq \frac{1}{2n} \sum_{i=1}^{k} \| X_{s_i} \beta_0 - (\hat{\beta}_{0,\nu}) \|^2 - \frac{1}{n} \sum_{i=1}^{k} \| X_{s_i,\nu} (\hat{\beta}_{s_i,\nu} - \hat{\beta}_{\nu}) \|^2$$

$$\geq \frac{1}{2n} \sum_{i=1}^{k} \zeta_{min}(X_{s_i}^T X_{s_i}) \| \beta_0 - (\hat{\beta}_{0,\nu}) \|^2 - \frac{1}{n} \sum_{i=1}^{k} \zeta_{max}(X_{s_i,\nu}^T X_{s_i,\nu}) \| \hat{\beta}_{s_i,\nu} - \hat{\beta}_{\nu} \|^2$$

$$\geq \frac{d \| \beta_{0,\nu} \|^2}{2} - o_p(1). \quad (2.37)$$
For the third term in (2.36),

\[
E[(X_s \beta_0 - X_{s, \nu} \hat{\beta}_{s, \nu})^T \varepsilon_{s_i}] = 0,
\]

\[
E|(X_s \beta_0 - X_{s, \nu} \hat{\beta}_{s, \nu})^T \varepsilon_{s_i}|^2 
\leq 2E[|\beta_0 - (\hat{\beta}_{0, \nu})|^T X_s^T \varepsilon_{s_i}]^2 + 2E|(\hat{\beta}_{\nu} - \hat{\beta}_{s, \nu})^T X_{s, \nu}^T \varepsilon_{s_i}|^2 
\leq 2(\|\beta_{0\nu}\|^2 + \alpha_p(1)) E\|X_s^T \varepsilon_{s_i}\|^2 + 2\alpha_p(1) E\|X_{s, \nu}^T \varepsilon_{s_i}\|^2 
\leq 2(\|\beta_{0\nu}\|^2 + \alpha_p(1)) t_{pn} D\sigma^2 + 2t_{\nu} D\sigma^2 \alpha_p(1) 
= O_p(\|\beta_{0\nu}\|^2 t_{\nu}).
\]

Hence,

\[
(X_s \beta_0 - X_{s, \nu} \hat{\beta}_{s, \nu})^T \varepsilon_{s_i} \leq O_p(\|\beta_{0\nu}\| \sqrt{t_{\nu}}) ;
\]

\[
\frac{2}{n} \sum_{i=1}^{K} (X_s \beta_0 - X_{s, \nu} \hat{\beta}_{s, \nu})^T \varepsilon_{s_i} \leq \frac{2K}{n} O_p(\|\beta_{0\nu}\| \sqrt{t_{\nu}}) 
= O_p(\|\beta_{0\nu}\| \sqrt{\frac{p_n}{n}}). \tag{2.38}
\]

By substituting (2.37)–(2.38) to (2.36), we get

\[
MCV_\nu \geq \frac{1}{n} \varepsilon^T \varepsilon + \frac{d\|\beta_{0\nu}\|^2}{2} + O_p(\|\beta_{0\nu}\| \sqrt{\frac{p_n}{n}}).
\]

If \(\alpha\) is an overfit model and \(\nu\) is an underfit model, we have

\[
\lim_{n \to \infty} MCV_\nu - MCV_\alpha 
\geq \frac{d\|\beta_{0\nu}\|^2}{2} + O_p(\|\beta_{0\nu}\| \sqrt{\frac{p_n}{n}}) - O_p\left(\frac{p_\alpha}{n}\right) > 0. \tag{2.39}
\]
So the first part is proved. I then combine it with Corollaries 2.1–2.4. For any \( r \), \( p_0 < r < p_n \),
\[
\lim_{n \to \infty} \frac{\text{WMF}_{p_0}}{\text{WMF}_r} = \lim_{n \to \infty} \frac{P^*(A_n^* = A \mid p_0) \exp[-\text{MCV}_A/c \sigma^2]}{P^*(A_n^* = M_r \mid r) \exp[-\text{MCV}_{M_r}/c \sigma^2]} = \lim_{n \to \infty} \frac{P^*(A_n^* = A \mid p_0)}{P^*(A_n^* = M_r \mid r)} \exp \left( \frac{\text{MCV}_{M_r} - \text{MCV}_A}{c \sigma^2} \right) = \lim_{n \to \infty} \frac{P^*(A_n^* = A \mid p_0)}{P^*(A_n^* = M_r \mid r)} \exp \left( O_p \left( \frac{r - p_0}{n} \right) \right) > 1.
\]
(2.40)

And for any \( r' \), \( 0 < r' < p_0 \),
\[
\lim_{n \to \infty} \frac{\text{WMF}_{p_0}}{\text{WMF}_{r'}} = \lim_{n \to \infty} \frac{P^*(A_n^* = A \mid p_0) \exp[-\text{MCV}_A/c \sigma^2]}{P^*(A_n^* = M_{r'} \mid r') \exp[-\text{MCV}_{M_{r'}}/c \sigma^2]} = \lim_{n \to \infty} \frac{P^*(A_n^* = A \mid p_0)}{P^*(A_n^* = M_{r'} \mid r')} \exp \left( \frac{\text{MCV}_{M_{r'}} - \text{MCV}_A}{c \sigma^2} \right) \geq \lim_{n \to \infty} \frac{P^*(A_n^* = A \mid p_0)}{P^*(A_n^* = M_{r'} \mid r')} \exp \left( \frac{d}{2} \| \beta_{0M_{r'}} \|^2 + O_p \left( \| \beta_{0M_{r'}} \| \sqrt{\frac{p_n}{n}} \right) - O_p \left( \frac{p_n}{n} \right) \right) > 1.
\]
(2.41)

Then model selection consistency of the WMF procedure can be deduced from (2.40)–(2.41).
2.8.2 Additional simulation results

Figure 2.9: Results of scenario 1 using residual bootstrap data: (a) proportion of correctly specified models; (b) number of false non-zeros; (c) number of false zeros; (d) estimated model size.

Figure 2.10: Results of scenario 2 using residual bootstrap data: (a) proportion of correctly specified models; (b) number of false non-zeros; (c) number of false zeros; (d) estimated model size.
Figure 2.11: Results of scenario 3 using residual bootstrap data: (a) proportion of correctly specified models; (b) number of false non-zeros; (c) number of false zeros; (d) estimated model size.

Figure 2.12: Results of scenario 4 using paired bootstrap data: (a) proportion of correctly specified models; (b) number of false non-zeros; (c) number of false zeros; (d) estimated model size.
Figure 2.13: Results of scenario 5 using paired bootstrap data: (a) proportion of correctly specified models; (b) number of false non-zeros; (c) number of false zeros; (d) estimated model size.

Figure 2.14: Results of scenario 6 using paired bootstrap data: (a) proportion of correctly specified models; (b) number of false non-zeros; (c) number of false zeros; (d) estimated model size.
Chapter 3

Precision Therapeutic Biomarker Identification

3.1 Motivation

The use of drugs to selectively target specific genetic alterations in defined patient subpopulations has seen significant successes. One example can be found in the treatment of chronic myeloid leukemia (CML) where the first consistent chromosomal abnormality associated with a human cancer was identified back in the 1960s. Fast forward to the 1980s where the consequence of this abnormality was discovered to be the production of an abnormal gene called BCR-ABL. Intense drug discovery programs were initiated to shut down the activity of BCR-ABL, and in 1992, imatinib (Gleevec) was developed. In 1998, the drug was tested in CML patients who had exhausted standard treatment options and whose life expectancy was limited, with remarkable results in their blood counts returning to normal. In 2001, the FDA approved imatinib. Today, a once commonly fatal cancer now has a five-year survival rate of 95% (Druker et al., 2006).

Achievements like this largely inspire today’s high throughput screening studies
of linking cancer drugs (known or in development) to specific genomic changes which could be used as therapeutic biomarkers. The hope is that such analyses will shed light on biological mechanisms underlying drug sensitivity, tumor resistance and potential drug combination synergies.

Cancer cell lines have frequently been used as a convenient way of conducting such studies. For a systematic search of therapeutic biomarkers to a variety of cancer drugs, the Genomics of Drug Sensitivity (GDSC) (Garnett et al., 2012) screened 639 human tumor cell lines, which represent much of the tissue-type and genetic diversity of human cancers, with 130 drugs. These drugs, including approved drugs, drugs in development as well as experimental tool compounds, cover a wide range of targets and processes involved in cancer biology. A range of 275–507 cell lines were screened for each drug. The effect of a 72h drug treatment on cell viability was examined to derive such measures of drug sensitivity as the half-maximal inhibitory concentration ($IC_{50}$). The cell lines underInt sequencing of 64 known cancer genes, genome-wide analysis of copy number gains and losses, and expression profiling of 14,500 genes.

Given the degree of complexity of this dataset, the multivariate analysis of variance (MANOVA) and the elastic-net regression applied in Garnett et al. (2012) are insufficient for precise knowledge discovery. First, the marginal drug-feature associations discovered in MANOVA rarely reflect true relationships, as it is more likely that sensitivity of cancer cells to drugs depends on a multiplicity of genomic and epigenomic features with potential interactions. Second, the elastic-net regression fails to concern following issues: 1) since the 639 cell lines come from a variety of cancer tissue types, there is likely additional heterogeneity manifested as subpopulations with overlaps in data; 2) note that the 130 drugs (response variables) are hardly independent, one can improve the prediction accuracy by modeling with multiple drugs
Moreover, there are some direct questions of interest from a subject matter perspective that I want to address. These include, i) can cancer-specific therapeutic biomarkers be detected, ii) can drug resistance patterns be identified along with predictive strategies to circumvent resistance using alternate drugs, iii) can biomarkers of combination therapies be identified to help predict synergies in drug activities? To tackle these questions and previously discussed statistical challenges, I propose a multivariate regression model with a latent overlapping cluster indicator variable. Fitting procedures inducing concurrent variable selection are introduced.

In section 3.2, I give a selective overview of existing clustering and overlapping clustering methods. In Section 3.3, a new statistical model is introduced, a generalized mixture of multivariate regression (GMMR) model in connection with the new model and a new EM algorithm for fitting are provided. I also establish a type of consistency optimality for estimation of the GMMR model. Simulation studies are presented in section 3.4. Section 3.5 contains a comprehensive re-analysis of the GDSC data using the proposed method. Discussions are included in section 3.6.

3.2 Literature Review

3.2.1 Clustering and mixture model

Clustering is a process of grouping data elements according to a measure of similarity. It has long been a very popular topic for statistical learning in diverse fields of marketing, biology, machine learning and etc. The variety of techniques for representing data, measuring similarity between data items and grouping data items have yielded
a rich assortment of clustering methods, see Jain and Dubes (1988); Jain et al. (1999) and Estivill-Castro (2002) for a detailed review of data clustering.

Among numerous clustering methods, the (finite) mixture model is most closely related to statistics. The model postulates that a set of data points come from a finite mixture of latent sub-populations with each sub-population following by a particular distribution; the goal is to estimate the parameters of each distribution and corresponding cluster memberships. As early as 1894, a seminal work of Pearson brought the mixture model into prominence, where he utilized a mixture of two univariate Gaussian model to fit a data of Neapolitan crabs supplied by his colleague W.F.R. Weldon. Since then, the mixture model has become a powerful tool in modeling heterogeneous and multi-populational data in a broad range of branches. Due to its popularity, the method has been extensively studied and developed by many researchers, see McLachlan and Basford (1988), chapter 9 of Bishop (2006) and section 14.3 of Hastie et al. (2009b) for comprehensive review.

Let $X \in \Omega \subset \mathbb{R}^p$ be a random variable, and $g_k(x)$ for $k = 1, \ldots, K$ are probability density functions on $\Omega$. Then $X$ from a mixture model with $K$ components has following density function

$$f(x) = \sum_{k=1}^{K} \pi_k g_k(x),$$  \hspace{1cm} (3.1)

where $\pi_k \geq 0$ for each $k$ and $\sum_{k=1}^{K} \pi_k = 1$.

The mixture model can be characterized by a hierarchical structure. Assume a latent random variable $Z$ which satisfies $P(Z = k) = \pi_k$ for each $k$. Further assume
that $X \mid Z = k$ has density function $g_k(x)$, the joint density of $X$ and $Z$ becomes

$$f(x, z) = \prod_{k=1}^{K} \{\pi_k g_k(x)\}^{I(z=k)},$$

(3.2)

where $I(\cdot)$ is an indicator function. Summing (2.2) over $z = 1, \ldots, K$ yields (2.1): the marginal density of $X$.

Note that $Z$ is a cluster membership variable for the observation of $X$. By applying the Bayes’ rule, we obtain the posterior probability of $Z = k$ given $X = x$ as

$$P(Z = k \mid X = x) = \frac{\pi_k g_k(x)}{\sum_{k'=1}^{K} \pi_{k'} g_{k'}(x)}$$

$$= p_k(x).$$

(3.3)

The set of posterior probabilities is considered as a fuzzy clustering of $X$ where each data point has a variable degree of membership in each of the output clusters.

However it is usually desirable to perform a hard-clustering which assigns each point to a single cluster based on some rule. In the mixture model, this is achieved by using the posterior probabilities in (2.3) via the Bayes’ assignment rule

$$R(x) = \arg \max_{k=1, \ldots, K} p_k(x).$$

(3.4)

Rule (2.4) was proved optimal by Theorem 22.6 in Wasserman (2004). Denote $k^*$ the true cluster ID that $x$ belongs to.

**Theorem 3.1** (Optimality of Bayes’s rule). Rule (2.4) minimizes the loss function

$$L(R) = P(R(X) \neq k^*)$$
over the class of all assignment rules.

3.2.2 Overlapping clustering

Traditionally, clustering techniques generate partitions so that each sample belongs to one and only one cluster. It has long been recognized that such an ideal clustering seldom exists in real data (Needham, 1967). It is more likely that clusters overlap in some parts. For example in films classification, a film can have potentially multiple genres resulting in an overlapping clustering. Or in toothpaste market segmentation studies, an individual may desire both decay prevention and fresh breath in the product. Therefore Arabie (1977) devised an additive clustering (ADCLUS) model for representing overlapping structures in sociological data and later applied it in marketing studies involving products/subjects that belong to more than one cluster (Arabie et al., 1981).

Lazzeroni and Owen (2002) putted forward a noteworthy additive model called the plaid model for two-sided overlapping clustering on gene expression data. Unlike ADCLUS model which works with the pairwise similarity matrix for only one-sided overlapping clustering, the plaid model decomposes the gene expression data as a sum of overlapping layers (amount to clusters). Each layer consists of a subset of genes and samples, each gene and sample can participate in multiple layers.

Let \( x_{ij} \) be a single gene expression data entry and \( K \) is the total number of layers. The plaid model is

\[
x_{ij} = \omega_0 + \sum_{k=1}^{K} (\omega_k + \alpha_{ik} + \beta_{jk}) P_{ik} Q_{jk},
\]

(3.5)

where \( \omega_0, \omega_k, \alpha_{ik}, \beta_{jk} \in \mathbb{R} \), \( P_{ik} \) is 1 if gene \( i \) belongs to the \( k \)th gene-block and 0 oth-
erwise, and $Q_{jk}$ is 1 if sample $j$ is in the $k$th sample-block and 0 otherwise. Although the plaid model is very flexible in depicting complex cluster structures of a data set, finding an optimal fit to the model is infeasible in practice. For a $n \times p$ dimensional gene expression data, there are $(2^n - 1)(2^p - 1)$ ways to choose the participating genes and samples for each layer, which are impossible to fully examined even for moderate $n$ and $p$. See also Turner et al. (2005a,b) for improved and extended (to structured data) plaid models and Zhang (2010) for the Bayesian plaid model.

Segal et al. (2002) introduced a probabilistic model to decompose a gene expression data as a sum of overlapping processes (equivalent to clusters). Different from the plaid model, this is a one-sided overlapping clustering problem. Each process of the probabilistic model involves a set of genes for all samples. Each gene can participate in multiple processes and each sample can have different activity levels in different processes. By assuming a conditional Gaussian distribution for each data entry, the model reduces to a matrix decomposition problem where one wants to decompose the gene expression data $X \in \mathbb{R}^{n \times p}$ into a binary membership matrix $M \in \{0, 1\}^{n \times K}$ and a real valued activation matrix $A \in \mathbb{R}^{K \times p}$ for minimal $\|X - MA\|^2$. Banerjee et al. (2005) generalized the probabilistic model to any regular exponential family distribution. However both of the methods suffer from computation complexity in estimating the binary matrix $M$, which is an integer optimization problem. Denote $M_i$ as a row vector of $M$, it is the cluster membership vector of gene $i$. The explicit enumeration method for estimating single $M_i$ involves evaluating $2^K$ permutations, which is infeasible in practice for moderate to large $K$ which is often seen in data decomposition problems. Therefore in both papers, suboptimal reduced searching algorithms for estimating the $M$ were adopted.

An alternative "naive" overlapping clustering method is the mixture model with
a hard threshold $\alpha$ on posterior cluster membership probabilities. This approach assigns a data point to a cluster $k$ if $P(z_i = k \mid X_i = x_i) > \alpha$ and hence enables an observation to belong to multiple clusters. However as pointed out by Banerjee et al. (2005), this method is problematic because: 1) it is hard to choose the value of $\alpha$; and 2) it is not a natural generative model for overlapping clustering since one underlying assumption of the mixture model is that each sample comes from one and only one mixture component. Later Fu and Banerjee (2008) introduced an overlapping clustering approach based on the multiplicative mixture model (Heller and Ghahramani, 2007) which replaces the (latent) random variable $Z$ in the mixture model with a (latent) binary membership vector $z = (z_1, \ldots, z_K)^T \in \{0, 1\}^K$. This model ends up with the same computational issue as in Banerjee et al. (2005). Therefore a reduced searching algorithm had to be used.

### 3.2.3 Mixture regression models

In regression analysis, the mixture of regression models is commonly used to capture unobserved cross-sectional heterogeneity in data (Jedidi et al., 1996). The model postulates that a sample of observations come from a (finite) mixture of latent sub-populations with each sub-population represented by a regression model. The conditional density of $Y$ given $X$ in a mixture of regression models of order $K$ is

$$f(y \mid x, \theta) = \sum_{k=1}^{K} \pi_k f(y \mid h(x^T \beta_k), \phi_k),$$

where $\theta = (\beta_1, \ldots, \beta_K, \phi, \pi)$, $h(\cdot)$ is a link function, $\phi_k$ are dispersion parameters, $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$.

To best of our knowledge, the mixture of regression models was first introduced by
Quandt (1972), where he proposed a two component mixture of univariate Gaussian linear regression model to fit an econometric data and brought forward the idea of using maximum likelihood function to estimate model parameters. Latter invention of the well-known expectation-maximization (EM) algorithm (Dempster et al., 1977) for maximum likelihood estimation (MLE) with incomplete data revolutionized studies of the mixture of regression models. DeSarbo and Cron (1988) provided an EM algorithm based MLE for the model. Khalili and Chen (2012) putted forward a penalized likelihood approach for its variable selection. Here is a simple sketch of the method.

If formulating the model in (2.6) via a hierarchical structure by introducing a latent random variable $Z$ similar to the mixture model in (2.2), I can rewrite the mixture of regression models as

$$f(y, z \mid x, \theta) = \prod_{k=1}^{K} \{ \pi_k f(y \mid h(x^T \beta_k), \phi_k) \}^{I(z=k)}. \quad (3.7)$$

For a sample of $n$ independent observations $(x_1, y_1), \ldots, (x_n, y_n)$ from model (2.7), the complete (conditional) likelihood equals to

$$L_n(\theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \{ \pi_k f(y_i \mid h(x_i^T \beta_k), \phi_k) \}^{I(z_i=k)}. \quad (3.8)$$

The corresponding complete (conditional) log-likelihood is then

$$l_n(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log f(y_i \mid h(x_i^T \beta_k), \phi_k), \quad (3.9)$$
where $z_{ik} = I(z_i = k)$. And a complete (conditional) penalized log-likelihood is

$$
\tilde{l}_n(\theta) = l_n(\theta) - \rho_n(\theta),
$$

(3.10)

with the penalty function

$$
\rho_n(\theta) = \sum_{k=1}^{K} \pi_k \rho_{nk}(|\beta_k|),
$$

(3.11)

where $\rho_{nk}(|\beta_k|)$ are nonnegative and nondecreasing functions on $|\beta_k|$ for $k = 1, \ldots, K$.

Khalili and Chen (2012) adopted the Lasso (Tibshirani, 1996), the HARD (Fan and Li, 2001) and the SCAD (Fan and Li, 2001) penalty functions respectively. A new EM algorithm which iterates between the E- and M- steps was proposed to maximize $\tilde{l}_n(\theta)$.

E step. Given $\theta^m$, estimate $z_{ik}^m$ with its posterior probability by applying the Bayes’ rule to (2.6) and (2.7),

$$
\hat{P}(z_{ik}^m = 1 \mid y_i, x_i, \theta^m) = \frac{\pi_k^m f(y_i \mid x_i, \beta_k^m, (\sigma_k^2)^m)}{\sum_{k'=1}^{K} \pi_{k'}^m f(y_i \mid x_i, \beta_{k'}^m, (\sigma_{k'}^2)^m)},
$$

(3.12)

for $i = 1, \ldots, n$ and $k = 1, \ldots, K$.

M step. Given $Z^m = (z_{ik}^m)$, update the mixing proportions $\pi^m$ by

$$
\pi_k^{m+1} = \frac{1}{n} \sum_{i=1}^{n} z_{ik}^m, \quad k = 1, \ldots, K.
$$

(3.13)

Note that this is obtained via maximizing the leading term of (2.10) with respect to $\pi$ for simplicity. However this simplified updating scheme worked well in simulations (Khalili and Chen, 2012).
Given $\mathbf{Z}^m, \pi^{m+1}$ and $\{(\sigma^2_l)^m\}_{l=1}^K$, update $\{\beta_k^{m+1}\}_{k=1}^K$ independently by solving

$$
\sum_{i=1}^n z_{ik}^m \frac{\partial}{\partial \beta_k} \left\{ \log f(y_i \mid \mathbf{x}_i, \beta_k, (\sigma^2_k)^m) \right\} - \pi_k^{m+1} \left\{ \frac{\partial}{\partial \beta_k} p_{nk}(\beta_k) \right\} = 0,
$$

(3.14)

for $k = 1, \ldots, K$ respectively.

And given $\mathbf{Z}^m, \pi^{m+1}$ and $\{\beta_k^{m+1}\}_{k=1}^K$, update $\{(\sigma^2_l)^{m+1}\}_{l=1}^K$ by respectively solving

$$
\sum_{i=1}^n z_{ik}^m \frac{\partial}{\partial \sigma^2_k} \left\{ \log f(y_i \mid \mathbf{x}_i, \beta_k^{m+1}, (\sigma^2_k)^{m+1}) \right\} = 0,
$$

(3.15)

for $k = 1, \ldots, K$.

This method was later generalized to the case when $p_n$ is diverging (Khalili and Lin, 2013). It was shown in Khalili and Chen (2012) and Khalili and Lin (2013) that the penalized likelihood approach enjoys oracle properties under some regularity conditions.

Let $\beta_k^T = (\beta_k^T, 0)$ to divide the coefficient vector into non-zero and zero subsets. Denote $\theta^0$ the true value of $\theta$, $\theta^0 = (\theta_0^T, 0)$ is the corresponding decomposition such that $\theta_0^T$ contains all zero coefficients $\beta_k$ for $k = 1, \ldots, K$, and $\hat{\theta}_n = (\hat{\theta}_{n1}, \hat{\theta}_{n2})$ is its estimator. Since the length of $\hat{\theta}_{n1}$ increases with $n$ when $p$ is diverging, Khalili and Lin (2013) studied the asymptotic distribution of its finite linear transformation, $B_n \hat{\theta}_{n1}$, where $B_n$ is a constant matrix of dimension $l \times d_{n1}$ with a finite $l$ and $d_{n1}$ is the dimension of $\theta_0^T$, moreover $B_n B_n^T \to \mathbf{B}$ and $\mathbf{B}$ is a positive definite symmetric matrix.

**Theorem 3.2** (Theorem 2 of Khalili and Lin (2013)). Under some regularity conditions on the penalty function, for any $\sqrt{n/p}$-consistent maximum penalized likelihood
estimator \( \hat{\theta}_n \) of \( \theta^0 \), as \( n \to \infty \):

(1) Variable selection consistency:

\[
P(\hat{\beta}_{nk,2} = 0) \to 1, \quad k = 1, \ldots, K.
\]

(2) Asymptotic normality:

\[
\sqrt{n} \mathbf{B}_n \mathbf{I}_1^{-1/2}(\theta^0_1) \left\{ I_1(\theta^0_1) \left( \frac{p''_n(\theta^0_1)}{n} \right) (\hat{\theta}_{n1} - \theta^0_1) + \frac{p'_n(\theta^0_1)}{n} \right\} \to^d N(0, \mathbf{B}),
\]

where \( \mathbf{I}_1(\theta^0_1) \) is the Fisher information matrix under the true model.

### 3.2.4 Mixture multivariate regression models

So far we have seen various developments in formulating and estimating the mixture of regression models with a single response variable. In diverse subjects of econometrics, psychometrics, medicine and etc., one very often generates multiple response variables with a single set of covariates, leading to multivariate regression problems. As a natural generalization to the mixture of regression models, Jones and McLachlan (1992) introduced the mixture of multivariate regression (MMR) models. A similar concept was present for the multivariate simultaneous equation model (Jedidi et al., 1996). Soffritti and Galimberti (2011) applied a multivariate Gaussian mixture model to the error terms of a multivariate linear regression model, the distribution of which often departs from normality. Afterwards, they put forward a multivariate t mixture model for the error terms to make the model even more robust in presence of outliers (Galimberti and Soffritti, 2014). But both the two papers set the coefficient matrix constant across mixture components. More importantly, none of the above
listed methods allow for simultaneous variable selection. In order to do variable selection, Leisch (2004) and Grun and Leisch (2008) introduced an $R$ package \texttt{flexmix} to fit the MMR model. One drawback of the \texttt{flexmix} is that it ignores the correlations between response variables.

### 3.3 The Proposed Method

#### 3.3.1 Statistical modeling

For a sample of $n$ observations (cell lines), denote $y_i = (y_{i1}, \ldots, y_{iq})^T \in \mathbb{R}^q$ a vector of responses ($IC_{50}$ values), $x_i = (x_{i1}, \ldots, x_{ip_n})^T \in \mathbb{R}^{p_n}$ a vector of predictors (genomic features) and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iq})^T$ a vector of random errors for the $i$th observation. Assume $\varepsilon_i \sim \mathcal{N}_q(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{q \times q}$ is an unstructured variance covariance matrix. Our proposed model is

$$y_i = \sum_{k=1}^{K} B_k^T x_i p_{ik} + \varepsilon_i, \quad i = 1, \ldots, n, \quad (3.16)$$

where $K$ is the total number of clusters, $B_k \in \mathbb{R}^{p_n \times q}$ is an unknown coefficient matrix for the $k$th cluster, and $p_{ik} \in \{0, 1\}$ is $1$ if observation $i$ belongs to the $k$th cluster, otherwise $0$. In traditional clustering problem, it assumes that each observation belongs to one and only one cluster, namely $\sum_{k=1}^{K} p_{ik} = 1$ for all $i$. Here I allow $\sum_{k=1}^{K} p_{ik} \geq 1$ so that each observation can belong to multiple clusters. More importantly, the number of coefficients in model (3.16), which is $Kq p_n$, remains the same as in traditional partitioning clustering.

I provide some interpretation for the clusters in model (3.16). A cluster $k$ contains
a subset of observations for whom \( P_{ik} = 1 \). Since not all genomic features are relevant in describing the cluster \( k \), I assume a sparse coefficient matrix \( B_k \). Because the sparse patterns can vary with \( k \), each cluster is represented by a unique set of biomarkers. For observations belong to multiple clusters, their response variables are explained by multiple sets of biomarkers, indicative of involving in several biological processes simultaneously.

### 3.3.2 The generalized mixture of multivariate regression models

Model (3.16) can be characterized by a hierarchical structure which ends up with an MMR alike model, I call it generalized mixture of multivariate regression (GMMR) models.

Suppose model (3.16) has \( K \) objective clusters indexed by 1 to \( K \). These objective clusters can overlap with each other, leading to \( 2^K - 1 \) types of overlapping patterns. Take \( K = 3 \) for an example, overall overlapping patterns are \( S = \{1, 2, 3, (1, 2), (1, 3), (2, 3), (1, 2, 3)\} \). Since each observation \( i \) from model (3.16) belongs to one and only one overlapping pattern, the overlapping patterns are mutually exclusive. I then define \( 2^K - 1 \) hypothetical clusters. Each hypothetical cluster (simply called “cluster” since after) represents an overlapping pattern indexed by an element in \( S \),

\[
S = \bigcup_{s=1}^{K} \{ (a_1 \ldots a_s) : \{a_1, \ldots, a_s\} \subseteq \{1, \ldots, K\} \},
\]

where \( s \) indicates the number of unique elements in \( (a_1 \ldots a_s) \) and \( S \) consists of all subsets in \( \{1, \ldots, K\} \) except for the null set. For example when \( K = 3 \) and \( s = 2 \), \( (a_1 \ldots a_s) \) can be \( (1, 2), (1, 3) \) or \( (2, 3) \). Cluster \( (1, 2) \) indicates its members belonging
to *objective clusters* 1 and 2 in (3.16).

I now introduce a latent cluster membership random variable $z_i$ for observation $i$ from model (3.16) and characterize the model via a hierarchical structure. Given $\sum_{k=1}^{K} P_{ik} \geq 1$, the range of $z_i$ becomes $S$. Further define

$$P(z_i = (a_1 \ldots a_s)) = \pi_{(a_1 \ldots a_s)}. \tag{3.17}$$

By properties of probability,

$$\sum_{(a_1 \ldots a_s) \in S} \pi_{(a_1 \ldots a_s)} = 1, \quad \pi_{(a_1 \ldots a_s)} \geq 0,$$

where $\pi = (\pi_{(a_1 \ldots a_s)} \in T)$ is a vector of unknown parameters.

Let $\boldsymbol{\Theta} = (\mathbf{B}_1, \ldots, \mathbf{B}_K, \Sigma, \pi)$. The vector of responses $y_i$ from model (3.16) satisfies

$$(y_i \mid z_i = (a_1 \ldots a_s), x_i, \Theta) \sim N_q \left( \sum_{k=1}^{s} \mathbf{B}_{ak}^T x_i, \Sigma \right). \tag{3.18}$$

By (3.17) and (3.18), the joint density of $y_i$ and $z_i$ is

$$f(y_i, z_i \mid x_i, \Theta) = \prod_{(a_1 \ldots a_s) \in S} \left\{ \pi_{(a_1 \ldots a_s)} f(y_i \mid z_i = (a_1 \ldots a_s), x_i, \Theta) \right\} I(z_i = (a_1 \ldots a_s)). \tag{3.19}$$

Summarizing (3.19) over $z_i$ leads to the GMMR model

$$f(y_i \mid x_i, \Theta) = \sum_{(a_1 \ldots a_s) \in S} \pi_{(a_1 \ldots a_s)} f(y_i \mid z_i = (a_1 \ldots a_s), x_i, \Theta). \tag{3.20}$$

Note that if $\pi_{(a_1 \ldots a_s)} = 0$ for $s > 1$, (3.20) reduces to traditional MMR model.
The (conditional) log-likelihood function of $\Theta$ for a sample of $n$ observations from model (3.16) is

$$
\bar{l}_n^0(\Theta) = \sum_{i=1}^n \log \left( \sum_{(a_1 \ldots a_s) \in S} \pi(a_1 \ldots a_s) f(y_i \mid z_i = (a_1 \ldots a_s), x_i, \Theta) \right).
$$

Maximizing above log-likelihood function yields non-zero estimates for all regression coefficients. Since not all genomic features are informative, variable selection to obtain a parsimonious model is necessary. I employ a penalized likelihood approach

$$
\bar{l}_n^0(\Theta) = l_n^0(\Theta) - \sum_{k=1}^K \rho_{nk}(B_k).
$$

The $L_1$-penalty in Lasso (Tibshirani, 1996) is used for simultaneous variable selection and estimation. The penalty function is

$$
\rho_{nk}(B_k) = \lambda_k \|B_k\|_1,
$$

where the tuning parameter $\lambda_k \geq 0$. Since $L_1$-penalty is singular at the origin, it can shrink some coefficients to exact 0 for sufficiently large $\lambda_k$ (Fan and Li, 2001).

### 3.3.3 Numerical solutions

The EM algorithm are used to compute parameters in the GMMR model. When using the EM algorithm, $(x_i, y_i)$ is regarded as incomplete data for missing $z_i$, so one works on the complete data $(x_i, y_i, z_i)$ with joint density in (3.19). For a sample of $n$ observations from model (3.16), the complete (conditional) log-likelihood function of
is

\[ l_n(\Theta) = \sum_{i=1}^{n} \sum_{(a_1 \ldots a_s) \in S} z_{i,(a_1 \ldots a_s)} \log \pi_{(a_1 \ldots a_s)} \]

\[ + \sum_{i=1}^{n} \sum_{(a_1 \ldots a_s) \in S} z_{i,(a_1 \ldots a_s)} \log f(y_i | z_i = (a_1 \ldots a_s), x_i, \Theta), \]

where \( z_{i,(a_1 \ldots a_s)} = I(z_i = (a_1 \ldots a_s)) \). Then the penalized complete log-likelihood function is

\[ \tilde{l}_n(\Theta) = l_n(\Theta) - \sum_{k=1}^{K} \rho_{nk}(B_k). \] (3.22)

Due to the overlapping setting, each \( B_k \) involves in multiple clusters. For example when \( K = 3 \), \( B_1 \) shows up in clusters 1, (1,2), (1,3) and (1,2,3). Thus \( B_k \) for \( k = 1, \ldots, K \) can not be optimized independently like in usual EM algorithm. A block decent approach is used instead to update \( B_k \) sequentially. More specifically, our revised EM algorithm iteratively maximizes \( \tilde{l}_n(\Theta) \) in two steps:

E-step: Given \( \Theta^m \) in the \( m \)th iteration, estimate \( z_{i,(a_1 \ldots a_s)}^m \) with its posterior probability by applying the Bayes’ rule to (3.19) and (3.20),

\[ P(z_{i,(a_1 \ldots a_s)}^m = 1 | y_i, x_i, \Theta^m) = \frac{\pi_{(a_1 \ldots a_s)}^m f(y_i | z_i = (a_1 \ldots a_s), x_i, \Theta^m)}{\sum_{(a_1' \ldots a_s') \in S} \pi_{(a_1' \ldots a_s')}^m f(y_i | z_i = (a_1' \ldots a_s'), x_i, \Theta^m)}. \]

M-step: Given \( Z^m = (z_{i,(a_1 \ldots a_s)}^m) \) in the \( m + 1 \)th iteration, update mixing proportions \( \pi^{m+1} \) by

\[ \pi_{(a_1 \ldots a_s)}^{m+1} = \frac{1}{n} \sum_{i=1}^{n} z_{i,(a_1 \ldots a_s)}^m, \quad (a_1 \ldots a_s) \in S. \]
Given $Z^m$ and $\Theta^{m+1} = (B^m_1, \ldots, B^m_K, \Sigma^m, \pi^{m+1})$, sequentially update $B^{m+1}_k$ by

$$
B^{m+1}_k = \arg \max_{B_k} \sum_{i=1}^n \sum_{(a_1 \ldots a_s) : \ k \in \{a_1, \ldots, a_s\}} z^m_{i,(a_1 \ldots a_s)} \log f (y_i \mid z_i = (a_1 \ldots a_s), x_i, \Theta^{m+1} \setminus B^m_k) - \rho_{nk}(B_k),
$$

where $\{(a_1 \ldots a_s) : k \in \{a_1, \ldots, a_s\}\}$ implies the subsets in $S$ which contains element $k$. By (3.18),

$$
B^{m+1}_k = \arg \min_{B_k} \sum_{i=1}^n \sum_{(a_1 \ldots a_s) : \ k \in \{a_1, \ldots, a_s\}} z^m_{i,(a_1 \ldots a_s)} (y^*_i - B^T_k x_i)^T (\Sigma^m)^{-1} (y^*_i - B^T_k x_i) + 2\rho_{nk}(B_k)
$$

$$
= \arg \min_{B_k} \sum_{(a_1 \ldots a_s) : \ k \in \{a_1, \ldots, a_s\}} \text{tr} \left( (Y^* - XB_k)^T z^m_{(a_1 \ldots a_s)} (Y^* - XB_k) (\Sigma^m)^{-1} \right) + 2\rho_{nk}(B_k),
$$

(3.23)

where

$$
y^*_i = y_i - \sum_{t \in \{a_1, \ldots, a_s\} \setminus k} (B^m_t)^T x_i, \quad Y^* = (y^*_1, \ldots, y^*_n)^T, \quad z^m_{(a_1 \ldots a_s)} = \text{diag}(z^m_{1,(a_1 \ldots a_s)}, \ldots, z^m_{n,(a_1 \ldots a_s)}).
$$

Equation (3.23) can be solved by the multivariate regression with covariance estimation (MRCE) algorithm (Rothman et al., 2010). The $L_1$ penalty not only induces sparsity in $\hat{B}_k$, but also it makes $\hat{B}_k$ be a function of $\Sigma$. Without the penalty, estimator in (3.23) is simply

$$
B^{m+1}_k = \sum_{(a_1 \ldots a_s) : \ k \in \{a_1, \ldots, a_s\}} (X^T z^m_{(a_1 \ldots a_s)} X)^{-1} X^T z^m_{(a_1 \ldots a_s)} Y^*;
$$
which amounts to the weighted ordinary least squares estimate for each response variable and does not take use of the covariances between response variables. I refer readers to Rothman et al. (2010) for more details on this topic.

Given $Z^m$ and $\Theta^{m+1} = (B_1^{m+1}, \ldots, B_K^{m+1}, \Sigma^m, \pi^{m+1})$, update $\Sigma^{m+1}$ by

$$
\Sigma^{m+1} = \arg \max_{\Sigma} \sum_{i=1}^n \sum_{(a_1, \ldots, a_s) \in S} z_{a_i(a_1, \ldots, a_s)}^m \log f(y_i \mid z_i = (a_1 \ldots a_s), x_i, \Theta^{m+1} \setminus \Sigma^m)
$$

$$
= \arg \min_{\Sigma} \sum_{i=1}^n \sum_{(a_1, \ldots, a_s) \in S} -z_{a_i(a_1, \ldots, a_s)}^m \log |\Sigma^{-1}|
$$

$$
+ \sum_{(a_1, \ldots, a_s) \in S} tr\left( (Y - X \sum_{k=1}^s B_{a_k}^{m+1})^T z_{a_i(a_1, \ldots, a_s)}^m (Y - X \sum_{k=1}^s B_{a_k}^{m+1}) \Sigma^{-1} \right). \tag{3.24}
$$

By taking the derivative of (3.24) according to $\Sigma^{-1}$, we get

$$
\Sigma^{m+1} = \frac{\sum_{(a_1, \ldots, a_s) \in S} (Y - X \sum_{k=1}^s B_{a_k}^{m+1})^T z_{a_i(a_1, \ldots, a_s)}^m (Y - X \sum_{k=1}^s B_{a_k}^{m+1})}{\sum_{(a_1, \ldots, a_s) \in S} \sum_{i=1}^n z_{a_i(a_1, \ldots, a_s)}^m},
$$

where $\sum_{(a_1, \ldots, a_s) \in S} \sum_{i=1}^n z_{a_i(a_1, \ldots, a_s)}^m = n$.

Commencing with an initial value $\Theta^0$, the algorithm iterates between E- and M-steps until the relative change in log-likelihood, $\left| \left( l_n^{m+1}(\hat{\Theta}) - l_n^m(\hat{\Theta}) \right) / l_n^m(\hat{\Theta}) \right|$, is smaller than some threshold value, taken as $10^{-5}$ in simulation studies and $10^{-3}$ in real data analysis. Additionally, a cluster, whose mixing proportion is smaller than some threshold value taken as 0.01 in the paper, will be removed during iterations to avoid over estimations.
3.3.4 Selections of tuning parameters and the \( K \)

In preceding penalized likelihood approach, one needs to choose the values of component-wise tuning parameters \( \lambda_k \) for \( k = 1, \ldots, K \), which controls the complexity of an estimated model. Here I use a component-wise 10-fold cross-validation (Stone, 1977) method for the tuning parameters selections in (2.36).

Selection of the number of components \( K \) is essential in mixture (regressions) models. In applications, the choice can be based on prior knowledge. With respect to formatted methodologies, information criteria (IC) remain by far the most popular strategy for selection of \( K \). See Claeskens et al. (2008) for general treatments on this topic. The method chooses the \( K \) by minimizing below \( IC_n \) as a function of \( K \),

\[
IC_n(K) = -2\ln(\Theta) + N_K a_n. \tag{3.25}
\]

Above \( N_k \) is the effective number of parameters in the model,

\[
N_K = |\{B_k, k = 1, \ldots, K\}| + |\pi| - 1 + |\Sigma|,
\]

where \( |A| \) calculates the number of nonzero elements in \( A \). In (3.25), \( a_n \) is a positive sequence depending on \( n \). The well known AIC (Akaike, 1974) and BIC (Schwarz et al., 1978) correspond to \( a_n = 2 \) and \( a_n = \log(n) \) respectively. It was shown in Keribin (2000) that under some regularity conditions, information criteria can identify the true order of a mixture model asymptotically. Feasibility of his results to mixture regressions models is unknown yet. I examined the performance of AIC for selection of \( K \) in a GMMR model in simulations. It obtained a high degree of accuracy.
3.3.5 Asymptotic properties

Khalili and Lin (2013) showed that their approach by maximizing a penalized log-likelihood function for estimation of the FMR model is consistent in both estimation and variable selection under certain regularity conditions. Denote $f(w; \Theta)$ the joint density function of data $w = (x, y)$ with $\Theta \in \Omega$. The conditional density function of $y$ given $x$ follows an FMR model in Khalili and Lin (2013) and a generalized FMMR model in (3.20) in our paper. As defined in (3.21), the penalized log-likelihood function is a summation of the $\log(f(w_i; \Theta))$’s minus a penalty function. Because the regularity conditions for asymptotic establishments in Khalili and Lin (2013) are made on $f(w; \Theta)$ and the penalty function directly, their theoretical achievements can be extended to our problem.

Denote $B^j_k$ the $j$th column of $B_k$ for $j = 1, \ldots, q$. Let $B^j_k = (B^j_{k1}, B^j_{k2})$ to divide the coefficient vector into non-zero and zero subsets. Denote $\Theta^0$ the true value of $\Theta$, $\Theta^0 = (\Theta^0_1, \Theta^0_2)$ is the corresponding decomposition such that $\Theta^0_2$ contains all zero coefficients, and $\hat{\Theta}_n = (\hat{\Theta}_{n1}, \hat{\Theta}_{n2})$ is its estimate. Since the dimension of $\hat{\Theta}_{n1}$ increases with $n$, I investigate the asymptotic distribution of its finite linear transformation, $D_n\hat{\Theta}_{n1}$, where $D_n$ is an $l \times d_{n1}$ constant matrix with a finite $l$ and $d_{n1}$ is the dimension of $\Theta^0_1$. Moreover $D_nD_n^T \to D$ and $D$ is a positive definite and symmetric matrix. I assume that $K$ is independent of the sample size $n$ and known beforehand. Its selection is discussed in Section 3.3.4.

**Lemma 3.1.** Suppose the penalty function $\rho_{nk}(B_k)$ satisfies conditions $\mathcal{P}_0 - \mathcal{P}_2$ and the joint density function $f(w; \Theta)$ satisfies conditions $R_1 - R_5$ in Khalili and Lin (2013).

1. If $\frac{\alpha^4}{n} \to 0$, then there exists a local maximizer $\hat{\Theta}_n$ for the penalized log-likelihood
where $q_{2n} = \max_{kij} \left\{ \frac{|\theta^0_{kij}|}{\sqrt{n}} : b^0_{kij} \neq 0 \right\}$, $b^0_{kij}$ is an entry of $B^0_k$.

2. If $\rho_{nk}(B_k)$ also satisfies condition $P_3$ in Khalili and Lin (2013) and $\frac{p_n}{n} \to 0$, for any $\sqrt{n/p_n}$-consistent maximum likelihood estimate $\hat{\Theta}_n$, as $n \to \infty$ it has:

i. Variable selection consistency:

$$P(\hat{B}^j_{k2} = 0) \to 1, \ k = 1, \ldots, K \ and \ j = 1, \ldots, q.$$ 

ii. Asymptotic normality:

$$\sqrt{n}D_nI_{1}^{-1/2}(\Theta^0) \left\{ \left[ I_{1}(\Theta^0) - \frac{\rho''_n(\Theta^0)}{n} \right] (\hat{\Theta}_{n1} - \Theta^0_1) + \frac{\rho''_n(\Theta^0)}{n} \right\} \to^d N(0, D), \ (3.26)$$

where $I_{1}(\Theta^0)$ is the Fisher information matrix under the true subset model.

Proof. Write $\Theta = (\theta_1, \theta_2, \ldots, \theta_{t_n})^T$, where $t_n$ is the total number of parameters in the model. Then Lemma 3.1 is a direct extension from Khalili and Lin (2013). □

By Lemma 3.1, $\hat{\Theta}_n$ under the Lasso or Elastic Net penalty has a convergence rate $\sqrt{p_n/n}$ via appropriate choice of the tuning parameters. However consistent estimation does not necessarily guarantee consistent variable selection. By the Lemma, the Lasso or Elastic Net penalty does not lead to consistent variable selection, because on one hand $\lambda_k$ must be large enough to achieve sparsity, on the other hand the bias term $q_{2n}$ is proportional to the $\lambda_k/\sqrt{n}$. This problem can be solved by using adaptive Lasso or adaptive Elastic Net penalty instead, which leads to concurrent estimation
and variable selection consistency.

### 3.4 Simulation Studies

#### 3.4.1 Design

Two scenarios were simulated. In each scenario, data were generated from model (3.16). Predictors $x_i$ were IID from $N_p(0, \Sigma_1)$ and random errors $\varepsilon_i$ IID from $N_q(0, \Sigma_2)$ with $\Sigma_1(i, j) = 0.4|i-j|$ and $\Sigma_2(i, j) = 0.7|i-j|$.

Scenario 1: Let $p_n = 100$, $K = 3$, $q = 3$ and $n = 450,900$. The 3 clusters are unoverlapping, each cluster contains $n/3$ observations.

Scenario 2: Let $p_n = 100$, $K = 3$, $q = 3$ and $n = 450,900$. 72% of the observations involve in single clusters, 18% of the observations in two clusters and 10% of the observations in three clusters.

The sparse coefficient matrixes $B_k, k = 1, 2, 3$ were generated as

$$B_k = W \otimes T \otimes S,$$  \hspace{1cm} (3.27)

where $\otimes$ indicates the element-wise product. Each entry of $W$ was drawn independently from $N(3,1)$ for scenarios 1-2 and $N(5,1)$ for scenario 3, each entry of $T$ was randomly sampled from $(-1,1)$ and each entry of $S$ was drawn independently from the Bernoulli distribution $B(1,0.2)$ for scenarios 1-2 and $B(1,0.08)$ for scenario 3. Thus I expected 20 non-zero coefficients for each response variable in scenarios 1-2 and 40 in scenario 3. Simulations were repeated 100 times. In each repetition, a new sequence of $B_k, k = 1, 2, 3$ was generated. The $K$ was selected by minimizing the AIC.
3.4.2 Cluster recovery quality measures

Quality measures in Baeza-Yates et al. (1999) were used to evaluate the clustering performance of the proposed method. Denote $N_S$ the number of elements in $S$. The quality measures were introduced to compare a retrieved cluster $\hat{A}$ with a target cluster $A$,

\[
\text{“specificity”} = \frac{N_{A \cap \hat{A}}}{N_A}, \quad \text{“sensitivity”} = \frac{N_{A \cap \hat{A}}}{N_{\hat{A}}},
\]

\[
\text{“F}_1\text{ measure”} = \frac{2N_{A \cap \hat{A}}}{N_A + N_{\hat{A}}},
\]

The $F_1$ measure, taken as the harmonic mean of specificity and sensitivity, gives an overall measure of the clustering.

Moreover, I used a one-to-one correspondence match approach (Turner et al., 2005a) to evaluate a sequence of retrieved clusters. It includes three steps. First make the number of retrieved clusters to be the same as the number of target clusters by adding null clusters to retrieved clusters or dropping addition poorly retrieved clusters. Retrieved clusters are then matched to target clusters in pair. Finally, calculate the mean quality measures for the sequence of paired clusters.

3.4.3 Results

The proposed method (gmmr-mrce) was compared with two other methods. First method was the ‘flexmix’ R package to fit MMR models where multivariate responses were treated as independent. Since it does not fit overlapping clusters, a naive overlapping clustering method was used where a data point was assigned to a cluster $k$ if $P(z_{ik} = 1 \mid x_i, y_i) > \frac{1}{K+1}$. Second method (gmmr-sepLasso) was the proposed
GMMR model, but in the model each column of $B_k$ was estimated separately by the Lasso.

Clustering performances were evaluated via the three quality measures introduced in Section 3.4.2. Results were summarized in Figure 3.1–3.2. In scenario 1 when there was no overlap, the three methods did equally well demonstrating that the GMMR model retains the ability to recover non-overlapping clusters. In scenario 2 when there was 28% overlap, specificity of the flexmix dropped down dramatically, resulting in a poor $F_1$ measure. The GMMR model (gmmr-sepLasso and gmmr-mrce methods) did much better than the flexmix by estimating overlapping clusters. The gmmr-mrce slightly outperformed gmmr-sepLasso when $n = 450$. The two methods converged very fast as $n$ increased, both achieving a median 100% clustering accuracy when
Figure 3.2: Simulation results for scenario 2. The two box plots under each method correspond to $n = 450, 900$ respectively.

$n = 900$. When $n = 900$, AIC identified the true $K$ for almost 100% of the times in the GMMR model, numerically showing that AIC is consistent in selection of $K$ in the GMMR model.

### 3.5 The GDSC Data Analysis

I now turn our attention back to analysis of the GDSC high throughput drug sensitivity dataset for cancer. An updated version of the drug response data, which are continuous log($IC_{50}$) values, (http://www.cancerrxgene.org/downloads/) assayed 707 human tumor cell lines (samples) with 140 drugs (response variables). The feature data had 624 cell lines with 13831 genomic features including 13 cancer types (binary), mutation status of 71 cancer genes (binary), continuous copy numbers of
426 genes causally implicated in cancer and continuous genome-wide transcriptional profiles. Continuous features were standardized to have mean 0 and variance 1. After removing the cell lines for which less than 50% of the drugs were tested in response data, there are in total 591 cell lines in both data sets (n = 591). The remaining drug response data has 16.5% missing values, which were imputed by the random forest imputation algorithm (Ishwaran et al., 2008).

**Direct Q1: Can cancer-specific therapeutic biomarkers be detected?**

The new method was applied to identify patterns of cancer-specific therapeutic biomarkers. Note that by “cancer-specific” I do not assume separate patterns for each cancer type but rather clusters that may be driven by one or a small number of cancer types which were identified by the GMMR model. Throughout the analysis, AIC was used for selection of the number of clusters by letting $K = 1, 2, 3, 4$. Note that the GMMR model automatically reduces to the multivariate regression model when $K = 1$. Therefore the use of GMMR model over the multivariate regression model were justified.

Although simulation studies show that it is beneficial to model with multiple drugs, it is unreasonable to simply fit all 140 drugs in one GMMR model. Because by doing so, one assumes that cell line clustering is the same for all 140 drugs, which can hardly be true.

Therefore, I first investigated each drug by fitting a generalized mixture of standard regression models. To reduce computational costs, for each drug 2000 continuous features which were most correlated with the drug and mutation status of 71 cancer genes were selected as input covariates. The model produced drug-specific cell line
clustering and cluster-wise coefficient estimates, $\hat{\beta}_k^c$ for $k = 1, \ldots, K$. Covariates with non-zero coefficient estimates were considered as biomarkers for a particular drug in a specific cell line cluster, where the cell lines came from one or some cancer types. Highly significant biomarkers were defined as those which had coefficients $\pm 1$ s.d. from the mean.

Bottom panel of Figure 3.3 shows highly significant biomarkers for each of the 9 out of 140 drugs (one small plot box) in a particular cell line cluster (one column in a small plot box). Red indicates drug-sensitive biomarkers and blue indicates drug-resistant biomarkers. The upper panel shows cancer type frequencies in each cell line cluster. The scale of frequencies were represented by the size of colored bubbles. It’s very clear that clusters being driven by different cancer types have very different therapeutic biomarkers and yet no clusters are homogeneous in a particular cancer
Secondly, 5 drug groups were identified and each was fitted in a GMMR model. To identify drug groups, pair-wise correlations among 140 drugs were calculated. Drugs were grouped only if their pair-wise correlations were $\geq 0.75$, this results in 5 groups. Drugs in each group target approximately the same biological process and belong to almost one target family, see Table 3.1 for informations of the 5 drug groups. Hence it is reasonable to assume that drugs in each drug group have similar cell line clusterings, making it appropriate to fit each drug group in a GMMR model.

Estimation results were summarized in Figure 3.4. Its explanations are similar to those made in Figure 3.3 except for that each column in a small plot box displays highly significant biomarkers (bottom) and cancer type frequencies (frequencies) for

---

### Table 3.1: Informations of the 5 drug groups fitted in the GMMR model

<table>
<thead>
<tr>
<th>Drug name</th>
<th>Drug targets</th>
<th>Target family</th>
<th>Effector pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paclitaxel</td>
<td>Microtubules</td>
<td>Chemotherapy</td>
<td>Cytoskeleton, Mitosis</td>
</tr>
<tr>
<td>S.Trityl.L.cysteine</td>
<td>KIF1</td>
<td>Other</td>
<td>Mitosis</td>
</tr>
<tr>
<td>BI.2536</td>
<td>PLK1/2/3</td>
<td>S/T Kinase</td>
<td>Mitosis</td>
</tr>
<tr>
<td>GW843682X</td>
<td>PLK1</td>
<td>S/T Kinase</td>
<td>Mitosis</td>
</tr>
<tr>
<td>JW.7.52.1</td>
<td>MTOR</td>
<td>S/T Kinase</td>
<td>Metabolism, PI3K/MTOR</td>
</tr>
<tr>
<td>A.443654</td>
<td>AKT1/2/3</td>
<td>S/T Kinase</td>
<td>Metabolism, PI3K/MTOR, Apoptosis</td>
</tr>
<tr>
<td>Dasatinib</td>
<td>ABL, SRC,</td>
<td>CTK, RTK</td>
<td>Cytoskeleton, Adhesion,</td>
</tr>
<tr>
<td></td>
<td>KIT, PDGFR</td>
<td></td>
<td>ERK Signalling, PI3K/MTOR</td>
</tr>
<tr>
<td>WH.4.023</td>
<td>SRC family, ABL</td>
<td>CTK</td>
<td>Cytoskeleton, ERK Signalling</td>
</tr>
<tr>
<td>Doxorubicin</td>
<td>TOP2</td>
<td>Chemotherapy</td>
<td>Replication</td>
</tr>
<tr>
<td>Etoposide</td>
<td>DNA replication</td>
<td>Chemotherapy</td>
<td>Replication</td>
</tr>
<tr>
<td>RDEA119</td>
<td>MEK1/2</td>
<td>S/T Kinase</td>
<td>ERK Signalling</td>
</tr>
<tr>
<td>CI.1040</td>
<td>MTOR</td>
<td>S/T Kinase</td>
<td>Metabolism, PI3K/MTOR</td>
</tr>
<tr>
<td>PD.0325901</td>
<td>MEK1/2</td>
<td>S/T Kinase</td>
<td>ERK Signalling</td>
</tr>
</tbody>
</table>
biomarkers were very different between different cell line clusters. On the other hand, however, the biomarkers have a lot of overlaps for drugs within each drug group. This makes sense because drugs have ≥ 0.75 pair-wise correlations in each drug group. One can gain new insights into the functions for similar drugs in each drug group by comparing their overlapping and non-overlapping biomarkers in Figure 3.4.
drug resistance based on mutation data. It is also true that tumors can develop resistance to first sequence known cancer genes in tumor genomes and then design therapies accordingly (Bailey et al., 2014). I now show how our methodology can be used to identify predictive strategies for circumventing drug resistance based on mutation data.

**Direct Q2:** Can drug resistance patterns be identified along with predictive strategies to circumvent resistance using alternative drugs?

I discussed in the Introduction of this chapter the success story of Gleevec which was used for the treatment of CML based on the known specificity of targeting the BCR-ABL gene. It is now becoming more and more common in the treatment of cancer to first sequence known cancer genes in tumor genomes and then design therapies accordingly (Bailey et al., 2014). It is also true that tumors can develop resistance to first line therapies by accumulating mutations which confer resistance.

| Gene       | CDKN2a(p14) MUT | SMARCA4 MUT | BCR_ABL MUT | FAM123B MUT | MAP2K4 MUT | CCND3 MUT | MYCL1 MUT | FGFR2 MUT | SETD2 MUT | MLLT3 MUT | HRAS MUT | PTEN MUT | CYLD MUT | CDK6 MUT | TP53 MUT | JAK2 MUT |
|------------|-----------------|-------------|-------------|-------------|------------|-----------|-----------|-----------|-----------|-----------|----------|----------|----------|----------|----------|----------|----------|
| Breast     | ●               | ●           | ●           | ●           | ●          | ●         | ●         | ●         | ●         | ●         | ●        | ●        | ●        | ●        | ●        | ●        |
| Blood      | ●               | ●           | ●           | ●           | ●          | ●         | ●         | ●         | ●         | ●         | ●        | ●        | ●        | ●        | ●        | ●        |
| Bone       | ●               | ●           | ●           | ●           | ●          | ●         | ●         | ●         | ●         | ●         | ●        | ●        | ●        | ●        | ●        | ●        |
| Kidney     | ●               | ●           | ●           | ●           | ●          | ●         | ●         | ●         | ●         | ●         | ●        | ●        | ●        | ●        | ●        | ●        |
| Lung       | ●               | ●           | ●           | ●           | ●          | ●         | ●         | ●         | ●         | ●         | ●        | ●        | ●        | ●        | ●        | ●        |
| Skin       | ●               | ●           | ●           | ●           | ●          | ●         | ●         | ●         | ●         | ●         | ●        | ●        | ●        | ●        | ●        | ●        |
| Digestive system | ●     | ●           | ●           | ●           | ●          | ●         | ●         | ●         | ●         | ●         | ●        | ●        | ●        | ●        | ●        | ●        |
| Nervous system | ●    | ●           | ●           | ●           | ●          | ●         | ●         | ●         | ●         | ●         | ●        | ●        | ●        | ●        | ●        | ●        |
| Soft tissue | ●              | ●           | ●           | ●           | ●          | ●         | ●         | ●         | ●         | ●         | ●        | ●        | ●        | ●        | ●        | ●        |

**Figure 3.5:** Estimation results from the GMR model for 7/140 drugs with covariates of mutation status of 71 cancer genes.
which contain a high percentage of blood cancer cell lines, along with cluster-wise resistant tumors. Take blood cancers for an example. First, extracted those clusters shown in Figure 3.5.

Mutation status of 71 known cancer genes were used as input covariates. Estimated cant biomarkers (bottom panel) for selected drugs.

Estimation results can be used for drug repurposing (Martins et al., 2015) for each drug was fitted in a generalized mixture of standard regression models. Mutation status of 71 known cancer genes were used as input covariates. Estimated highly significant biomarkers and cell line clusterings for 7 out of 140 drugs were shown in Figure 3.5.

Estimation results can be used for drug repurposing (Martins et al., 2015) for resistant tumors. Take blood cancers for an example. First, extracted those clusters which contain a high percentage of blood cancer cell lines, along with cluster-wise highly significant biomarkers. Extracted results were presented in Figure 3.6. Its upper panel shows $IC_{50}$ values of blood cancer cell lines and bottom panel shows

Figure 3.6: Extracted $IC_{50}$ values of blood cancers (upper panel) and highly significant biomarkers (bottom panel) for selected drugs.
highly significant biomarkers in extracted clusters. Drugs in Figure 3.6 were further filtered by only keeping those that have low $IC_{50}$ values without resistance biomarkers and that have high $IC_{50}$ values with resistance biomarkers, which leads to Figure 3.7. In this figure, the former set of drugs can be used as alternatives for the later set of drugs which progress resistance to blood cancers.

3.6 Discussions

In this chapter, I proposed a new statistical model for identifying therapeutic biomarkers for cancer which can answer specific questions regarding sensitivity and resistance.
I used a penalized likelihood approach for the mixture of multivariate regression models which enforced sparsity in genomic features. To enable overlapping clustering, the mixture of multivariate regression models was generalized and a new EM algorithm derived for estimation of model parameters.

Some improvements can be made for future developments on this work. First, other useful penalty functions for multivariate regression estimations can be adopted, such as the MAP (MAster Predictor) penalty (Peng et al., 2010), the $L_2$SVS (Similä and Tikka, 2007) and the $L_\infty$SVS (Turlach et al., 2005). Second, the assumption of multivariate normal distribution on $Y_i$ can be extended to other flexible families of multivariate distributions, such as the multivariate skew-normal distribution (Azzalini, 2005) and the multivariate skew-$t$ distribution (Chen et al., 2014), for frequent presence of skewness and kurtosis in real data.

Missing values in GDSC were imputed by the random forest imputation algorithm. It would be of interest to explore the impact of this step more thoroughly. An alternative would be to generalize to this problem, the recently developed E-MS algorithm Jiang et al. (2015) for model selection with incomplete data.
Chapter 4

Discordancy Partitioning for Validating Potentially Inconsistent Pharmacogenomic Studies

4.1 Motivation

It’s now widely believed that cancers are far more heterogeneous than once thought - that in fact, they represent a myriad of different diseases with varying biological determinants rather than a single entity whose effective treatment will rely on some over-arching theoretical construct of our understanding of the disease. To this end, high throughput pharmacogenomic screening of small molecules and other compounds has the potential to implicate new drug leads (or drug combinations) that can be used for more personalized treatments.

A typical pharmacogenomic workflow involves characterizing interesting compounds for dose-response effects on cancer cell lines, and then doing functional genomic characterization in additional screens. Of interest is the elucidation of therapeutic biomarkers whose patterns might be predictive of a compound’s activity against a particular cancer cell line.
Given the complexity of such assays, the necessity of proper validation is paramount. Given a fresh set of data generated from the same workflow, one should be able to demonstrate both accurate predictions of drug activity as well as reproducibility of therapeutic genomic signatures.

The opportunity to study validation of models arose with the generation of two major pharmacogenomic datasets - the Genome Drug Sensitivity in Cancer (GDSC) project (Yang et al., 2013; Garnett et al., 2012) and the Cancer Cell Line Encyclopedia (CCLE) project (Barretina et al., 2012). Both represent large-scale efforts in which experimental and FDA approved drug compounds were screened against panels of molecularly characterized cancer cell lines.

A number of recent studies looked at the 15 drugs and 706 cell lines which were in common between the two studies (Consortium et al., 2015; Haibe-Kains et al., 2013; Papillon-Cavanagh et al., 2013). One study acted as a training dataset and the other the test dataset. Successful validation of training set models would be evidenced by low test set prediction error of drug response and reproducibility of therapeutic genomic signatures on the test dataset. It should be noted however, that in general developing accurate signatures does not always imply low prediction errors. Overfit models are also known to predict well - however the test set error differences between overfit and accurate more sparse models is usually not large (Ishwaran and Rao, 2014).

Curiously, validating pharmacogenomic models has proven elusive. A lack of concordance of drug response behaviors across a number of drugs makes validation difficult. Interestingly, the exact causes of these discordances has still not been elucidated (Haibe-Kains et al., 2013; Safikhani et al., 2016a). A number of re-analyses of these data using additional suggestions for analysis methods did not result in significantly
improved validation findings (Safikhani et al., 2016a). Test set prediction errors were not as low as expected apriori, and very few genomic markers were re-detected on the test dataset. Even honing in on markers with known drug interactions, many were not reproduced. A consequence of these findings is that they put into doubt the potential to use pharmacogenomics to discover new therapeutic biomarkers. Some additional studies have been carried out by adding a third new drug screening dataset to the mix to see if improved concordance could be achieved (Greshock et al., 2010; Mpindi et al., 2016; Haverty et al., 2016). In all cases, the GDSC-CCLE discordancies could not be completely resolved. For instance, a third independent dataset called the Genentech Cell Line Screening Initiative (gSCI) (Haverty et al., 2016) was generated. Improved agreement appears to be present between gSCI and CCLE was found but concordance with GDSC was weaker. However, careful reading of this paper shows that there are still issues with the improved agreement that was found. The proportion of shared genomic features found is low, association values are not compared (meaning the actual magnitudes and signs of the effects are not displayed), and the elastic net model they used was only fit once and all non-zero estimates were regarded as candidate biomarkers. Typically these models need setting of tuning parameters which are found by methods like cross-validation. Cross-validation in itself can be unstable and thus it’s more robust to repeat the process many times and look at a more rigorous definition of a candidate biomarker in terms of repeated detection over many iterations of cross-validation. A third independent dataset from the Institute for Molecular Medicine Finland (FIMM) (Mpindi et al., 2016) was generated. This had a significantly higher level of consistency with CCLE than with GDSC. However, only 26 cell lines were shared, in contrast CCLE and GDSC shared 268 cell lines. So the improved results may be due to some sort of selection bias, but even with
this, large inconsistencies in therapeutic signatures were still present. Whatever the underlying reasons, it appears that discordancies between pharmacogenomic datasets may be persistent.

In this chapter, I present an approach that may resolve the discordancy issues and provide a direct path with which to directly assess whether models validate (i.e. both types of validation mentioned above). I develop a discordancy partitioning approach using the data sharing strategy (Gross and Tibshirani, 2016; Chen et al., 2015) across datasets which can easily be extended to more than two datasets. In doing so, I approach the validation problem quite differently. I formally acknowledge within the analysis strategy that discordancies in drug responses may exist and in doing so, partition the effects of genomic predictors into two pieces - the true underlying effect and the portion due to discordancy. The result is that I find significantly more accurate training set models in terms of test set prediction accuracy and discover reproducible therapeutic genomic signatures that appear much more comprehensive than ones previously found.

4.2 Methods

Figure 4.1 presents a schematic of our analysis strategy. I use all of the training dataset (say CCLE), and a portion of the other dataset (GDSC). The remainder of the GDSC dataset will be withheld from modeling and used purely for evaluating test set prediction errors. In order to have proper representation of all tissue types, I ensure that the portion of GDSC used for modeling contains cell lines sampled at random from all tissue types.

The goal of the analysis is to identify therapeutic genomic signatures for each
Figure 4.1: Schematic of data sharing for the combined GDSC-CCLE analysis.

drug (and potentially further by cancer type). The response variable $y$ comes from
the drug response assay and has typically been taken to be the $\log(IC50)$ value or the
drug-specific area under the curve (AUC) value from the dose-response curve. The
predictors $\mathbf{x}$, come from the molecular assays and include, mutation status of known
cancer genes, gene expression values, copy number values and cancer tissue types. In
order to relate $y$ to $\mathbf{x}$, I consider two different model formulations:

Formulation 1:

$$y_i = \beta^T \mathbf{x}_i + \varepsilon_i, i \in G$$
$$y_i = (\beta + \delta)^T \mathbf{x}_i + \varepsilon_i, i \in C,$$  \hspace{1cm} (4.1)

where $\in G$ or $\in C$ represents cell line membership in GDSC or CCLE datasets
respectively. The model dimensions are $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^p$, $\beta \in \mathbb{R}^p$, $\delta \in \mathbb{R}^p$. The
errors $\varepsilon_i \in \mathbb{R}$ are identically and independent distributed with zero mean and finite
A more general version of the model is to allow multiple drug responses being modeled simultaneously where \( y_i \in \mathbb{R}^q \) and then fit a generalized finite mixture of regressions model. This can be used to identify and mitigate drug resistance or identify drug synergies. I will report on this multivariate model elsewhere.

Model (4.1) formally partitions the common genomic effects across datasets (\( \beta \)) from the potential dataset discordancies (\( \delta \)). A version of this was called data shared Lasso (Gross and Tibshirani, 2016) and used Lasso estimation (Tibshirani, 1996) to permit sparsity in parameter estimation. An earlier version of this approach traces to the data enriched linear regression (Chen et al., 2015). I use a different constraint on the model parameters by using a version of the elastic net (enet) (Zou and Hastie, 2005). Specifically, I add an additional \( L_2 \) constraint on the model parameters which allows finding sparse grouped signal which represent correlated genomic features. The following penalized optimization function is used,

\[
(\hat{\beta}, \hat{\delta}) = \arg \min_{\beta, \delta} \frac{1}{2} \sum_{i \in G} (y_i - x_i^T \beta)^2 + \sum_{i \in C} (y_i - x_i^T (\beta + \delta))^2 + \lambda_1 \left( \sum_{j=1}^p \omega_j |\beta_j| + r \sum_{j=1}^p \psi_j |\delta_j| \right) + \lambda_2 (||\beta||_2^2 + r ||\delta||_2^2),
\] (4.2)

where \( \lambda_1, \lambda_2 > 0 \) are shrinkage tuning parameters. The weights \( \omega_j \) and \( \psi_j \) are estimated from the elastic net estimation, e.g. \( \omega_j = (|\hat{\beta}_j^{enet}|)^{-1} \), and similarly for \( \psi_j \). I let \( r = 1/\sqrt{2} \) as recommended (Gross and Tibshirani, 2016). Ten-fold cross validation is used to optimize \( \lambda_1 \) and \( \lambda_2 \) from a pre-specified set of values.

To avoid a lack of a potential invariance due to the choice of the reference group in (4.1), an alternate formulation of the model can be made as represented in Formulation 2:
Formulation 2:

\[ y_i = (\beta + \delta_G)^T x_i + \varepsilon_i, i \in G \]

\[ y_i = (\beta + \delta_C)^T x_i + \varepsilon_i, i \in C, \]

where \( \delta_G \in \mathbb{R}^{p \times q} \) are discrepancy parameters of \( G \), \( \delta_C \in \mathbb{R}^{p \times q} \) are discrepancy parameters of \( C \). Here \( \beta \) represents a baseline effect that is shared across datasets. I will restrict most of our attention to Formulation 1 since \( \varepsilon \) parameters need to be estimated and fitted models are generally more stable as a result.

**Simulation design.** In order to study the performance of our proposed approach and compare it against alternative strategies, I will first conduct a simulation study where I know the true level of dataset discordancy. Specifically, I will consider the following simulation design.

Data were generated from model (4.1) where \( x_i \) are i.i.d from \( \mathcal{N}_p(0, \Sigma) \) with \( p = 1000 \) and \( \Sigma(i, j) = 0.75^{|i-j|} \), and \( \varepsilon_i \) i.i.d from \( \mathcal{N}(0, 1) \). To mimic the group effect of genomic data, non-zero elements of \( \beta \) are distributed in blocks with block size 3 and 16 non-zero blocks in total with each non-zero value drawn randomly from \( \mathcal{N}(1.5, 1) \). Twenty indices of non-zero entries and 10 indices of zero entries of \( \beta \) were randomly selected which constituted the indices of the non-zero elements of \( \delta \) with non-zero values drawn randomly from \( \mathcal{N}(0, 0.5) \).

Let the sample size of the reference dataset \( G \) increase from 100 to 400 and the sample size of the alternative dataset \( C \) be fixed at 300. I also drew a test dataset with sample size 100 from the upper equation of (4.1), i.e. from the same model as \( G \). I applied model (4.1) to \( G \) and \( C \) and calculated the mean prediction error on test data. As shown in (4.1), coefficients associated to data \( G \) and \( C \) in shared model are
respectively $\hat{\beta}$ and $\hat{\beta} + \hat{\delta}$, of which the models with regard to are called shared G and shared C. The mean prediction error using either set of coefficients was calculated. Prediction performance of the shared model was compared to other candidate models: individual model built on G, individual model built on C and individual model built on aggregated data of G and C. The simulation was repeated 200 times and results averaged.

Analyses of GDSC and CCLE Datasets. Data Preprocessing: While 706 cell lines were both present in the updated versions of GDSC and CCLE data sets and had associated genomic data, only a subset of those had drug screening data in both data sets. After filtered the 8 cell lines with different SNP identity, potentially due to mislabeling or contamination of cell lines (Safikhani et al., 2016b), and cell lines of which the number of missing genomic data was greater than 10000, only a range of 77-274 cell lines per compound had shared drug response data (median = 88 cell lines; mean = 162). Mutation status of known cancer genes, gene expression profiles and tissue type were used as input variables, among which those with missing rate larger than 0.3 were excluded, resulting in $p = 16223$ features in the model. The remaining feature data had $< 0.1\%$ missing data and was imputed by each feature mean. Each covariate vector and drug response vector was standardized to have mean 0 and variance 1. Sure independence screening (SIS) (Fan and Lv, 2008) was used to reduce the number of covariates to a moderate size. More specifically, 2000 covariates most correlated with the response variable were selected in each GDSC and CCLE dataset. This strategy makes sense because dataset discordancy may infiltrate into the screening procedure as well. If our goal is validation, then I must acknowledge this. Thus screening both datasets is an attempt to deal with this. Selected features were then combined as input variables in model.
Data Analysis Strategy: Formulation 1 model was fitted to the GDSC and CCLE data sets using CCLE as the training dataset. I fit a model with AUC as a response variable. For the data sharing strategy, I used all of the CCLE data and a random sample of 55% of the GDSC data as shared training data. This shared training data was used for the sure independence screening and model estimation. This process was repeated 200 times. Test set prediction errors were estimated on the 45% left out GDSC portions and averaged over the 200 runs. I compared prediction errors against the following methods: i) using the 55% GDSC sample alone (this represents a gold standard subject to sample size limitations when discordancies exist between the datasets), ii) using only the CCLE sample, iii) pooling the CCLE and the 55% GDSC sample as one common training set and iv) using CCLE only based on parameter estimates from the shared model. In order to see the effects of increased sharing, I increased the training set sample size to include 85% of the GDSC leaving only 15% out for testing purposes.

For signature validation, I ran a shared analysis using all cell lines in GDSC and CCLE. Here I included the copy number data into the feature set of the model because of the increased sample size available for analysis and followed previous procedures for data processing. I examined the estimated $\beta$ and $\delta_C$ values in order to determine how many genomic markers reliably reproduced their effects across datasets, and how many effects were washed away because of dataset discordancies. Variables with non-zero estimates are determined to be factors associated with the drug response. This procedure was repeated 200 times for each drug to assess the stability of variables when applying the ten-fold cross validation method. A variable list was built for each drug. It consists of all variables that appeared in any of the 200 runs along with the frequency the feature appears and average coefficient given to that feature over the
Figure 4.2: Test set error boxplot pairs by analysis method. Left hand member of the pair represents a training set sample size of 100 and the right hand member, a training set sample size of 400.

runs it appears. The average coefficient was used to assess the effect size of a feature in drug response. The most significant predictors associated with the drug response are defined as those with an effect size $\pm 2$ s.d. from the mean and a frequency rate $\geq 80\%$. I also fitted Formulation 2 and plotted the most significant estimates of $\beta + \delta_G$ and $\beta + \delta_C$.

### 4.3 Results

*Simulation*: Prediction results were summarized in Figure 4.2. When sample sizes are small, the shared G model outperformed all other methods and individual G model has much larger test errors than other methods do (which exceeds the range of the
which the result \( \hat{\delta} \) and calculated its correlation with the true \( \delta \) model and individual aggregated model never beat the shared G model in all cases.

\( \beta \) (as the test data were from this model), and more importantly the shared G model performs exactly as good as this gold standard. Since there are non-negligible disparities between data G and C due to the \( \delta \) in our simulations, the shared/individual C model and individual aggregated model never beat the shared G model in all cases.

Correlations of \( \hat{\delta} \) from model (4.1) with the true \( \delta \) in 200 simulations were calculated and present in boxplot of Figure 4.3. For comparison, I subtracted \( \hat{\beta}_G \) (estimate of \( \beta \) from individual G model) from \( \hat{\beta}_C \) (estimate of \( \beta \) from individual C model) of which the result \( \hat{\beta}_C - \hat{\beta}_G \) can be used to estimate \( \delta \) in absence of the shared model, and calculated its correlation with the true \( \delta \). Not surprisingly, the shared method

Figure 4.3: Top left panel: correlations between \( \hat{\delta} \) and the true \( \delta \) across the 200 runs of the simulation. The rest panels show the scatterplots of \( \hat{\delta} \) versus true values at the median correlations from top left boxplots.
compared for each of the 15 drugs in common to the two datasets. Here, AUC was calculated for estimating group discrepancies produced much higher correlations than the other method did, see Figure 4.3. I also calculated correlations of $\hat{\beta}$ from model (4.1) with the true $\beta$ in 200 simulations and compared them to the correlations of $\hat{\beta}_G$ with the true $\beta$. Results were summarized in Figure 4.4. Again, shared model produced much higher correlations than individual R model did when sample size is small. As sample size largely increases, individual G model becomes the gold standard and the shared model is as good as this gold standard.

**GDSC and CCLE Datasets:** Figure 4.5 shows side-by-side boxplots (over the 200 random splits of the GDSC dataset) of test set error rates for all methods being compared for each of the 15 drugs in common to the two datasets. Here, AUC was
Figure 4.5: Test error rate boxplots by drug where 55% of GDSC data were used in training set.

used as the response variable in the modeling. Highlighted in yellow, are the names of drugs where other groups had found reasonable concordance (Pearson correlation $> 0.45$) between the two drug response data sets (Consortium et al., 2015).

Some striking conclusions are evident. First, for drugs with reasonable concordance, the new shared methodology does not provide much in way of test set error reductions. This is exactly what should happen. However, for the other 11 drugs, the data sharing methodology clearly produces the most accurate models in terms of test set error rates - and this pattern is observed pretty much uniformly.
Figure 4.6: Test error rate boxplots by drug where 85% of GDSC data were used in training set.
By using the 55% GDSC training data alone, the GDSC-alone method had much larger test error rates than the GDSC-shared method, this indicates a clear deleterious effect by small sample size. To further prove this point, I increased the training set size to 85% and left 15% out for testing purposes. Figure 4.6 shows these results. Clearly, the increased training set size has helped the GDSC-alone method reduce its test error rates. However the GDSC-alone model is not the scenario that has caused trouble but rather serves as a sort of gold standard with which to compare methods against because the test data set come from it. Typical modeling would use the CCLE dataset alone (the dark blue boxplots). This clearly suffers from not being able to deal with the discordancies across datasets. When pooling the 55% GDSC and full CCLE datasets as one big overall training dataset, one cannot unravel the discordancy effect from the true effect resulting in poor predictions. So in conclusion, clearly the new data sharing methodology is capable of producing highly accurate predictive models when confronted with fresh test data which may be discordant in drug responses.

Figure 4.7 shows a very interesting plot of signature validation for all of the 15 drugs using AUC as the response variable. The y-axis shows results from the fit on the GDSC data (lowercase g) and the CCLE data (lowercase c). The x-axis depicts a subset of genomic effects with at least one highly significant non-zero effect across the 15 drugs. The body of the plot shows the estimated marker effect sizes. For g-drugs, the $\hat{\beta}$ values were plotted. For c-drugs, the $(\hat{\beta} + \hat{\delta})$ values are plotted. Blue colors indicate markers which predict drug sensitivity, red the opposite - the darker the color, the more intense the effect. Other genomic predictors not plotted showed no highly significant non-zero estimated effects for any of the 15 drugs. Shown in color are the known gene-drug associations in Table 1 (Safikhani et al., 2016b). All
Figure 4.7: Signature validation plot by fitting Formulation 1 model to GDSC (g) and CCLE (c) datasets. The x-axis shows the markers which had significant effects. 

of the associations in the table except for Nutlin-3–MDM2 expression were recovered. The association pattern of Crizotinib–HGF expression was recovered. It however was not identified as highly significant based on previously described rule, hence was not present in Figure 4.7. What is more interesting is how many new reproducible markers the data sharing strategy finds for each drug. What’s also noticeable is that the bands do not entirely overlap between g and c. These areas without overlap are markers where the dataset discordancies were large enough to wash away true effects (|δ_j| ≫ |β_j|) such that they were not detected as reproducible. Closer examination of particular drugs reveals that the washing out effect is happening in those drugs where discordancies were previously established and much less so where concordance was found. This is exactly what the theory would have predicted.

One can dig further into the signature validation plot. For instance, population stratifications can be made to better identify only those patients who may provide a
higher likelihood of a favorable drug responses. Take Nilotnib as an example. Here, having increased gene expression in C15orf26, SLC4A1, MPO, IGLL1, and APOL4 (i.e. the reproducible markers shaded dark blue) would be predictive of a favorable drug response. In addition, this was accentuated for thyroid cell lines as compared to the other cancer cell lines. On the other hand, for PLX4720, having the BRAF mutation and TRIM51 expression present, would predict a favorable response as long as no NRAS mutation was present. Other interesting patterns like this can be gleaned for each drug.

The horizontal banding is clearly what is most visible initially. However, one can also look vertically at specific genomic markers and find interesting information. For example, with the KRAS mutation, the only drug which shows strong reproducible predicted sensitivity to its presence is PD-0325901 - a MEK inhibitor. Many other drugs are predicted to encounter a resistance effect as indicated by the red c-g pairs.
along the KRAS column. In fact, recent evidence suggests that PD-0325901 used in conjunction with dacomitinib may be of use when KRAS mutations are present (Hamidi et al., 2014). Contrast this to the BRCA2 mutation. It does not show up at all on Figure 4.7. Thus one may conclude that no detectable signal was associated with this marker for any drug. This can be confirmed by examining Figure 4.8 which shows the signature validation plot using Formulation 2. Here BRCA2 does show and does indicate a smaller reproducible effect for Nutlin-3. This may seem puzzling since what is being plotted under Formulation 2 are estimates of $\delta_G$ and $\delta_C$ rather than estimated value of $\beta$ alone. However the coloring is identical indicating that the estimated values of $\delta_C$ and $\delta_g$ are near zero. The difference in the conclusion between the two model Formulations is then due to the stringency of the inclusion rules that are used.

4.4 Discussions

I have demonstrated that reproducible signal can in fact be partitioned from signal due to dataset discordancy using data sharing strategies. The most exciting thing may in fact be that I found new reproducible biomarkers for every drug in the analysis. Furthermore, patient populations can be partitioned based on their reproducible biomarker profiles to more precisely predict a favorable response to a drug. Additionally, I have shown that test set prediction error rates are markedly lower when using discordancy partitioning models and very much follow established theory. Thus while experimental challenges may still exist in order to better standardize protocols, it’s likely datasets will never be completely concordant. Discordancy partitioning approaches like what I have presented can adapt easily to varying degrees of discordancy
to produce more accurate assessments of validation.

As for the methodology itself, a few more remarks are in order. With regards to estimation, it should be noted that I am not saying that the combined elastic net/Lasso estimation technique is the only one that could be used here. In fact I developed our own modeling strategy which can be used to find interesting new patterns of therapeutic biomarkers (Liu and Rao, 2017). However, even our own modeling strategy can be embedded within a data sharing strategy for validation purposes. The reason I chose to illustrate results with the elastic net/Lasso approach is because it’s a more widely known approach for estimating sparse genomic models.
Bibliography


Efron, B. and Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*, volume 38. SIAM.


