PENALIZED ESTIMATION METHODS AND THEIR APPLICATIONS IN GENOMICS AND BEYOND

Ting-Huei Chen

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2014

Approved by:

Wei Sun

Jason P. Fine

Yun Li

William Valdar

Fei Zou

ABSTRACT

TING-HUEI CHEN: Penalized Estimation Methods and Their Applications in Genomics and Beyond
(Under the direction of Wei Sun and Jason P. Fine)

Various forms of penalty functions have been developed for regularized estimation. The tuning parameter(s) of a penalty function play a key role in penalizing all the noise to be zero and obtaining unbiased estimation of the true signals. For penalty functions with more than one tuning parameters, previous studies have not emphasized on the joint effect of all the tuning parameters. In the first topic, we conduct a theoretical analysis to relate the ranges of tuning parameters of penalty functions with the dimensionality of the problem and the minimum effect size. We exemplify our theoretical results in several well-known penalty functions. The results suggest that a class of penalty functions that bridges L_0 and L_1 penalties require less restrictive conditions for variable selection consistency. The simulation analysis and real data analysis support these theoretical results.

For the second topic, we consider the problem of identifying genomic features to predict cancer drug sensitivity. Several drugs that share a molecular target may also have some common predictive features. Therefore, it is desirable to analyze these drugs as a group to identify the associated genomic features. Motivated by this problem, we develop a new method for high-dimensional feature selection using a group of responses that may share a common set of predictors in addition to their individual predictors. Simulation results show that our method has better performances than existing methods. Between-study validation in real data shows that the genomic fea-

tures selected for a drug target can form good predictors for other drugs designed for the same target.

For the third topic, we address an estimation problem where certain parameter values such as 0 would cause an identifiability issue. In the maximum likelihood estimation framework, due to the issue of the unidentifiable parameter, the maximum likelihood estimator have regular properties only if the likelihood function is specified correctly with respect to the parameter values. We propose a penalized estimation procedure using the adaptive Lasso penalty to address the potential identifiability issue. We study the asymptotic property of the proposed estimator and evaluate our method in extensive simulations and real data analysis.

I would like to dedicate my dissertation work to my beloved grandparents and parents.

ACKNOWLEDGEMENTS

First and foremost I offer my deepest gratitude to my advisors, Drs. Wei Sun and Jason P. Fine. Their guidance helped me in all the time of research. I would also like to thank them for encouraging and helping me to shape my interest and ideas. I could not have imagined having better advisors and mentors for my graduate studies.

Besides my advisors, I would like to thank the rest of my committee: Drs. Yun Li, William Valdar and Fei Zou, for their encouragement, insightful comments and questions.

I would like to express my gratitude to Drs. Jianwen Cai, Haibo Zhou, Hongtu Zhu and Donglin Zeng for their suggestions and comments on the questions during my Ph.D. study.

My greatest appreciation and friendship goes to my many friends and schoolmates. They have helped me stay positive through these years.

TABLE OF CONTENTS

LIST OF TABLES . . . . . **ix**

LIST OF FIGURES . . . . . **x**

1 Introduction **1**

1.1 The role of tuning parameters of penalty functions 1

1.2 Prediction of cancer drugs' sensitivities 5

1.3 Models that are subject to unidentifiable parameters 10

2 The role of tuning parameters **13**

2.1 Introduction 13

2.2 Theoretical results 16

2.2.1 Notations and problem setup 16

2.2.2 The role of the tuning parameters 17

2.3 Algorithm and tuning parameter selection 24

2.4 Simulation 28

2.4.1 Linear model 29

2.4.2 Simulation for logistic model 31

2.5 Real data analysis 33

2.6 Asymptotic results 34

3 Prediction of cancer drug sensitivity **47**

3.1 Introduction 47

3.2 Method 50

3.2.1 Objective function 50
3.2.2 Computation 52
3.2.3 A Bayesian interpretation of BipLog 54
3.2.4 A penalized maximum likelihood estimation perspective 55
3.2.5 Tuning parameter selection 56
3.3 Simulation Studies 57
3.3.1 Simulation setup 57
3.4 Genomic signatures of cancer drug sensitivity 59
3.4.1 Evaluation of prediction model using training/testing data . . 61
3.4.2 Construction of prediction model 63
3.4.3 Validation of the prediction model 67

4 Models that are subject to unidentifiable parameters. 71
4.1 Introduction. 71
4.2 Asymptotic results. 73
4.2.1 Notations 73
4.2.2 The estimation procedure 74
4.2.3 Model of ($\beta = 0$; unidentifiable ζ) 75
4.2.4 Model of ($\beta \neq 0$; identifiable (θ, ζ)) 76
4.3 Simulation studies 79
4.4 Real data analysis 81
4.4.1 Stagnant band height data example 82
4.4.2 Metabolic pathways data example 84
4.4.3 Drug sensitivity data example 85
4.5 Additional conditions and asymptotic results 86

5 Conclusion 98

Bibliography 100

LIST OF TABLES

2.1 Simulation results for penalized linear regression 30

2.2 Simulation results for penalized logistic regression 32

2.3 Running time rounded to minutes per simulation 32

3.1 Comparisons of three bi-level selection method 60

3.2 Prediction R-squares 70

4.1 Empirical studies of the penalized estimation I 81

4.2 Empirical studies of the penalized estimation II 82

4.3 Empirical studies of the penalized estimation III 83

4.4 Empirical studies of the penalized estimation IV 84

LIST OF FIGURES

2.1 Marginal association p-values for 645,316 SNPs on chromosome 1 . . 14

2.2 GWA marginal p-values 34

3.1 The Log and Lasso penalty functions 51

3.2 Summary of the results of the within-study analysis 64

3.3 Distribution of the number of selected features 64

3.4 Genomic features associated with four groups of drugs 65

3.5 Evaluation of the predictive model 68

3.6 Pairwise box-plots of $\log IC_{50}$ for the 12 drugs 69

4.1 Stagnant band height data 85

4.2 17-AAG and the gene expression 87

CHAPTER 1: Introduction

1.1 The role of tuning parameters of penalty functions

Variable selection has been well studied in the classical setting of fixed dimensional covariates, with numerous penalization methods shown to yield sparse oracle estimation. Asymptotically, such procedures guarantee that the zero coefficients are estimated to be zero exactly and the non-zero coefficients are efficiently estimated with variance equal to that with known zero coefficients. Extending such methods to high dimensional covariates is technically challenging. Valid estimation is only possible if the regression model is sufficiently sparse, that is, a high percentage of covariates have no effect, with the number of non-zero effects growing at some rate that depends on the sample size.

Several penalty functions have been proposed for regularized estimation in such high dimensional setting. One of the most popular penalty functions is the Lasso penalty [Tibshirani, 1996]. Lasso is a convex penalty, so that including this penalty in the objective function (e.g., adding it to residual sum squares or subtracting it from log likelihood) does not change the convexity (e.g., residual sum squares) or concavity (e.g., log likelihood of generalized linear model) of the objective function. Therefore, it is computationally efficient to solve the penalization problem because finding the global minimum/maximum is equivalent to finding the local minimum/maximum.

Recently, several groups have studied the theoretical properties of Lasso for fixed

p [Zou, 2006] or for high-dimensional regression problems [Zhao and Yu, 2006; Meinshausen and Bühlmann, 2006; Zhang and Huang, 2008]. One important finding of these studies is that the variable selection consistency of Lasso requires the irrepresentable condition on the design matrix [Zhao and Yu, 2006], or equivalently, the neighborhood stability condition [Meinshausen and Bühlmann, 2006]. Intuitively, this condition requires that the covariates not in the true model (which are referred to as *"unimportant covariates"* hereafter) cannot be represented by the covariates belonging to the true model (which are referred to as *"important covariates"* hereafter). This condition is often not satisfied at high dimensionality such as NP dimensionality, i.e. the dimensionality of nonpolynomial (NP) order of sample size. For example, in GWAS, an important covariant, which is a SNP associated with the disease status, is often correlated with several nearby SNPs that are unimportant. In other words, the SNP-to-SNP correlations are totally due to linkage disequilibrium and have nothing to do with disease association.

In a pioneering work, Fan and Li [2001] have build a theoretical framework for nonconcave penalized likelihood for variable selection, and advertised a folded-concave penalty, the Smoothly Clipped Absolute Deviation Penalty (SCAD) proposed by Fan [1997], which is defined by

$$p'_{\texttt{SCAD}}(|\beta_j|) = \{\lambda I(|\beta_j| \leq \lambda) + [(a\lambda - |\beta_j|)/(a-1)]I(\lambda < |\beta_j| < a\lambda)\}\,,$$

where $\lambda > 0$ and $a > 2$ are two regularization parameters. SCAD employs Lasso penalty for signals smaller than a threshold λ, then reduces the penalty increase rate for stronger signals, and finally the penalty becomes a constant for signals larger than $a\lambda$. This reduction of penalty for stronger signals effectively removes the bias of Lasso for strong signals.

Another penalty, the Minimax Concave Penalty (MCP) [Zhang, 2010], is defined by

$$p'_{\texttt{MCP}}(|\beta_j|) = I(|\beta_j| < a\lambda)(a\lambda - |\beta_j|)/a,$$

where $\lambda > 0$ and $a > 0$ are two regularization parameters. MCP increases with a rate of λ from effect size zero, i.e., $\lim_{|\beta_j|\to 0^+} p'_{\texttt{MCP}}(|\beta_j|) \to \lambda$. Then it immediately reduces the penalty increase rate. The penalty becomes constant for effect size larger than $a\lambda$. MCP converges to L_0 penalty when $a \to 0$, and it converges to L_1 penalty when $a \to \infty$.

Another folded-concave penalty, the Smooth Integration of Counting and Absolute deviation penalty (SICA) [Lv and Fan, 2009] is a linear combination of L_0 and L_1 penalties:

$$p_{\texttt{SICA}}(|\beta_j|) = \lambda \left[\frac{|\beta_j|}{|\beta_j| + \tau} I(|\beta_j| \neq 0) + \frac{\tau}{|\beta_j| + \tau} |\beta_j| \right],$$

where $\lambda > 0$ and $\tau > 0$ are two regularization parameters. A more general class of linear combination of L_0 and L_1 penalties has been studied by Liu and Wu [2007].

The Log penalty [Friedman, 2008; Sun et al., 2010] is defined by

$$p_{\texttt{log}}(|\beta_j|) = \lambda \log(|\beta_j| + \tau),$$

where $\lambda > 0$ and $\tau > 0$ are two tuning parameters. As mentioned by Friedman [2008], Log penalty bridges L_0 and L_1 penalties. Specifically, it converges to L_0 or L_1 penalties if $\tau \to 0$ or $\tau \to \infty$, respectively. Lv and Fan [2009] pointed out the Log penalty is closely related with the SICA penalty. Sun et al. [2010] suggested the Log penalty can be viewed as iterative adaptive Lasso and provides a Bayesian interpretation of the Log penalty.

Another class of folded-concave penalty is the bridge penalty $p_{\texttt{Bridge}}(|\beta_j|) = |\beta_j|^a$, where $0 < a < 1$. Friedman [2008] has shown that the bridge penalty spans a similar spectrum as Log penalty, and the latter has smaller discontinuities, hence more stable coefficient estimates. In addition, $\lim_{|\beta_j|\to 0^+} p'_{\texttt{Bridge}}(|\beta_j|) \to \infty$, which leads to extra computational challenge for implementation. Therefore we do not include the bridge penalty for the latter theoretical studies.

It has been established, in both finite dimensions and diverging dimensions where $p = O(n^a)$ or $p = O(\exp(n^a))$ $(a > 0)$ that penalization methods based on folded-concave penalties provide consistent estimates without requiring the irrepresentability condition [Fan and Lv, 2010]. In addition, Mazumder et al. [2011] have studied properties of Log, SCAD and MCP in the optimization using a coordinate-descent approach.

The performances of the variable selection rely on the proper selection of regularization parameters ϖ. All of these four penalties (SCAD, MCP, SICA and Log) have two regularization parameters. In practice, immediate questions concerning these regularization parameters are whether they both should be tuned, and what is the consequence to tune only one of them in order to improve computational efficiency? Previous works have provided recommendations regarding to the choice of tuning parameters, but there is no systematic asymptotic studies on the roles of multiple tuning parameters. This motivates us to asymptotically study the relation of the choice of tuning parameters with the difficulty of the variable selection problem, namely the minimum effect size and the dimensions (both the number of important and unimportant covariates).

We will study the role of tuning parameters of penalty functions by evaluating if they could satisfy the conditions of weak oracle properties. Weak oracle property of penalized likelihood method in NP dimensionality (i.e., the dimensionality of nonpolynomial order of sample size) was introduced by Lv and Fan [2009] for penalized least squares, and was extended to generalized linear regression by [Fan and Lv, 2011]. An estimator $\hat{\beta} = (\hat{\beta}_1^{\mathsf{T}}, \hat{\beta}_2^{\mathsf{T}})^{\mathsf{T}}$ is considered to have weak oracle property if $\hat{\beta}_2 = \mathbf{0}$ with probability tending to 1 as $n \to \infty$, and consistency for $\hat{\beta}_1$ under L_∞ loss.

However, the conditions of weak oracle properties in Fan and Lv [2011] are mainly imposed on a single tuning parameter of penalty functions, and it is unclear the role of multiple tuning parameters. Therefore, we propose to generalize the theorems of weak oracle properties in Fan and Lv [2011]. This modification is necessary to allow more penalties to be studied for their tuning parameters.

1.2 Prediction of cancer drugs' sensitivities

Cancer drugs development has shifted from traditional one-size-fits-all cytotoxic chemotherapy to molecularly targeted cancer drug therapy. The cytotoxic chemotherapy drugs target the signaling pathway for cell division. Although cancer cells have out of control of growth pattern, normal cells such as cells in the bone marrow and hair follicles also divide regularly. The chemotherapy drugs cannot distinguish cancer cells from normal ones so that its side-effects can be severe. Unlike chemotherapy drugs, molecularly targeted cancer drugs aim to exploit the specific vulnerability of cancer cells. With advances in biotechnology, the studies of genomics and proteomics have generated huge amount of molecular data for targeted cancer drugs development.

Hoelder et al. [2012] gives a review about the targeted cancer drugs development. For instances, several drugs have been approved by FDA to target mutational activation of BCR-ABL tyrosine kinase in chronic myeloid leukemia, EGFR tyrosine kinase in non-small cell lung cancer, BRAF kinase in melanoma, and HER2 amplification in breast cancer [Yap and Workman, 2012].

Despite the effectiveness of these drugs in many patients, not all the patients who have a targeted mutation response to the corresponding drug, which is partly due to the (genome-wide) genetic heterogeneity among cancer patients. For example, only 30% patients with HER2 amplification and 50% patients with BRAF mutation respond to the corresponding drugs [De Palma and Hanahan, 2012]. Therefore, statistical models that can predict drug sensitivities from patient-specific genomic data will be of great value for cancer treatment. Such genomic data may include DNA alterations, gene expression, and epigenetic marks. Owing to the advance of high-throughput array/sequencing techniques, these genomic data can be collected in routine clinical practice in the near future [Yap and Workman, 2012]. Robust preclinical model systems such as cancer cell lines that reflect the genomic diversity of human cancers can be used to build such predictive model [Caponigro and Sellers, 2011].

Recently, two groups have studied drug sensitivities in a large number of cancer cell lines [Garnett et al., 2012; Barretina et al., 2012]. In a panel of several hundred human cancer cell lines, Garnett et al. [2012] measured the sensitivities of 130 drugs, mutation statuses of 64 commonly mutated cancer genes, and genome-wide copy number alterations and gene expression. In a panel of 479 cancer cell lines, Barretina et al. [2012] have screened 24 anticancer drugs and measured the mutation statuses

of 1600 genes, as well as genome-wide copy number alterations and gene expression. Both studies conducted univariate drug-by-drug analysis to select genomic features associated with drug sensitivity as measured by the half-maximal inhibitory concentration (IC_{50}), i.e., the amount of drugs to kill 50% of the cancer cells. Since drugs can be grouped by their targets such as a gene product or a signaling pathway, jointly analysis of the drugs sharing a target may improve the power to identify common genomic features.

Regarding feature selection for multivariate responses, two types of methods have been applied: group-wise selection and bi-level selection. Group-wise variable selection methods, such as group Lasso [Yuan and Lin, 2006] or group adaptive Lasso [Wang and Leng, 2008], assume all the response variables within a group are associated with the same set of covariates [Huang et al., 2012]. The assumption that all drugs sharing the same target (response variables within a group) have the same associated genomic features is unrealistic. The analysis results of the studies (Garnett et al. [2012] and Barretina et al. [2012]) show that in addition to some shared features, drugs with the same target have their own individual features respectively. In contrast, bi-level selection methods encourage the selection of covariates associated with all the response variables, but also allow some covariates to be associated with one or a few response variables [Breheny and Huang, 2009] are more appropriate for the application. A few methods have been developed for bi-level selection, such as group bridge [Huang et al., 2009] and composite MCP [Breheny and Huang, 2009].

Suppose in a group of n samples, we observe q response variables, denoted by $y_k = (y_{1k}, ..., y_{nk})^T$ $(1 \leq k \leq q)$, and p covariates, denoted by $\mathbf{x}_j = (x_{1j}, ..., x_{nj})^T$ $(1 \leq j \leq p)$. In addition, let β denote the coefficients matrix with each row as

$\mathbf{b}_j = (\beta_{j1}, ..., \beta_{jq})$ and each column as $\mathbf{b}_k = (\beta_{1k}, ..., \beta_{pk})$, and $\|\|_1$ to be the 1-norm.

The objective function of 1-norm group bridge is

$$\frac{1}{2n}\sum_{k=1}^{q}\|y_k - X\mathbf{b}_k\|_2^2 + \lambda\sum_{j=1}^{p} c_j\|\mathbf{b}_j\|_1^{\gamma}, \tag{1.2.1}$$

where $\lambda > 0$ is the tuning parameter, γ is the bridge index, and c_j is constants.

Following [Huang et al., 2012], composite penalties are defined as:

$$\rho_O(\sum_{k=1}^{q}\rho_I(|\beta_{jk})|),$$

where ρ_O is an outer penalty applying to a some of inner penalties ρ_I. Composite MCP is using both ρ_O and ρ_I to be the MCP penalty, which is presented in the previous section.

Although these methods work satisfactorily in many real data analyses, we find that their performances are limited in our preliminary simulation analysis for genomic applications where the genomic features have strong correlations. As shown in our study results on the penalty functions in the previous section, their limited performance may be due to the properties of incorporated penalty functions. These issues motivate us to develop a new method to construct predictive models of cancer drug sensitivities using genomic features.

Based on the study results of the first paper, the Log penalty has its advantages in the high dimension and low sample size problem. In addition, the previous study of penalized estimation with univariate response variable has shown that the method incorporated with Log penalty has better performance than other existing penalty functions including Lasso or elastic net [Sun et al., 2010]. Therefore, we propose to

extend the univariate version of method in [Sun et al., 2010] for multivariate penalized estimation.

The penalized estimation method in [Sun et al., 2010] is built based on the Bayesian hierarchical model, which can be considered as Bayesian shrinkage estimation. The principle is to assign priors with mean 0 on the parameters that are subject for shrinkage. Consider a univariate linear regression problem with both response and covariates being standardized, $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + e_i$, where $\mathbf{e} \sim N(\mathbf{0}_{n\times 1}, \sigma^2 I_{n\times n})$ and p is the number of covariates. In this case, the parameters which are subject for shrinkage are the coefficients β_j. One choice for the prior of β_j is Normal distribution with mean 0 and variance σ_j^2. The assigned priors on σ_j^2 are key to the performance of the Bayesian shrinkage methods. Several priors have been proposed for σ_j^2 such as inverse-Gamma or exponential prior, and the obtained Bayesian shrinkage methods have been suggested as Bayesian Lasso [Yi and Xu, 2008].

The priors in [Sun et al., 2010] are set as

$$p(\beta_j|\kappa_j) = \frac{1}{2\kappa_j}\exp\left(-\frac{|\beta_j|}{\kappa_j}\right), \tag{1.2.2}$$

$$p(\kappa_j|\delta,\tau) = \text{inv-Gamma}(\kappa_j;\delta,\tau) = \frac{\tau^\delta}{\Gamma(\delta)}\kappa_j^{-1-\delta}\exp\left(-\frac{\tau}{\kappa_j}\right), \tag{1.2.3}$$

where $\delta > 0$ and $\tau > 0$ are two hyperparameters. The Bayesian shrinkage method constructed by the above priors has shown its advantages compared to other existing methods [Sun et al., 2010].

1.3 Models that are subject to unidentifiable parameters

The problem of statistical inference in the presence of nuisance parameters that are not identified under the null hypothesis has been studied in several literatures. It is a non-regular testing framework since the nuisance parameter only present under the alternative hypothesis. Therefore, the standard large sample asymptotic theory cannot be directly applicable (Davtes [1977], Davies [1987]). Andrews [1993] considers the tests of structural change with unknown change point, where the unknown change point is not identifiable under the null hypothesis, and provides tests for various nonlinear models applied in econometric applications. Hansen [1996] studies the asymptotic distribution theory for the tests of model that are subject to unidentifiable parameters including the form of additive nonlinearity and allowing for stochastic regressors and weak dependence.

For the estimation problems, there are extensive literatures on estimation of the change point. For instance, Bai [1997] establishes the convergence rate and asymptotic distribution for the least square estimation of a change point in multiple regression. Muggeo [2003] considers the regression models with one or more break-points parameters and utilizes a linearization technique for fitting piecewise terms in the models. He and Severini [2010] studies the theoretical properties of maximum likelihood estimators of the parameters of a multiple change-point model. They establish the consistency, the convergence rate, and the asymptotic distribution of the maximum likelihood estimators.

In the maximum likelihood estimation framework, due to the unidentifiable parameter (ζ) issue under null hypothesis ($\beta = 0$), the maximum likelihood estimator

(MLE) have regular properties only if the likelihood function is specified correctly with respect to the parameter value of β. Specifically, when $\beta = 0$, the parameters ζ and β should be both absent from the likelihood function; then the MLE for the rest parameters are regular. On the contrary, when $\beta \neq 0$, the parameters ζ and β are both present in the likelihood function; the MLE do not have identifiability issues. Take the change point model estimation as an example. The parameter ζ is the change point parameter, and it exists only when $\beta \neq 0$. Instead of estimating the change points like the above methods, we are interested in designing an estimation procedure that can automatically take care of the specification of correct likelihood function with respect to the values of β without assuming the existence of change points.

Since whether β equals to 0 plays a key role in determining the form of likelihood function, we utilize the idea of penalization estimation procedure and apply adaptive Lasso penalty to β. The adaptive Lasso penalty incorporated to β has the form: $\lambda|\beta|w$, where λ is a tuning parameter, and w stands for the adaptive weight associated to β. As shown in [Zou, 2006], given a proper chosen w, adaptive lasso performs as well as if the true underlying likelihood were given in advance.

To choose a proper weight for β, we propose to apply the idea of constructing a test statistics in (Davtes [1977]. They have established the weak asymptotic optimality properties against local alternatives for their proposed test statistics, and its form of critical region is:

$$\{\sup_{L\leq\zeta\leq U} T(\zeta) > c\}, \tag{1.3.1}$$

where $T(\zeta)$ is assumed to be an appropriate test statistic and the range of $[L, U]$ is the possible values for ζ. For large values of $\sup_{L\leq\zeta\leq U} T(\zeta)$, the null hypothesis will

be rejected. Similarly, we take the supremum of profile likelihood estimates of $\hat{\beta}(\zeta)$ over a range of possible values of ζ to be the weight for β to construct our estimation procedure.

CHAPTER 2: The role of tuning parameters

2.1 Introduction

In genome-wide association (GWA) studies, the goal is to identify the genetic factors such as single nucleotide polymorphisms (SNPs) that are associated with diseases. With the availability of a dense map of SNPs, it is statistically very challenging to select the important SNPs from millions of SNPs using only a couple of thousand samples. Regularized estimation procedures can be applied for simultaneous selection of important variables (SNPs) and estimation of their effects for high dimensional data in GWA studies. The objective function of the regularized estimation is composed of a model fitting metric (e.g., likelihood function) and a penalty function for the parameters subject to regularization. Prior to the usage of regularized estimation, screening can be applied to reduce the number of SNPs to be considered for penalized estimation. However, due to the high correlation of neighboring SNPs, the number of SNPs that pass a reasonable screening criterion is often larger than or much larger than the sample size.

We use the real SNP genotype data from a recent study [Wright et al., 2014] to illustrate the correlation structure of genotype data. We take the genotypes of 645,316 SNPs in chromosome 1 from 1,198 samples, and randomly pick 30 SNPs as important variables to simulate the response under the linear model assumption. The effect size is simulated as 0.7 and the residual errors are standard normal variables. Figure 4.1 shows a Manhattan plot of the marginal association p-values. The 30 important SNPs are labeled by grey vertical lines. It is obvious that the high correlation among

nearby SNPs leads to small p-values for those SNPs that are close to the 30 important SNPs. If we apply screening using the p-value cut-off 10^{-4}, 3,087 SNPs will be selected which include 20 of the 30 important SNPs. Alternatively, if the p-value cut-off is 10^{-8}, 991 SNPs will be selected, which include only 13 of the 30 important SNPs. Thus screening method can be helpful to certain extend, and screening with stringent threshold would lead to many false negatives. This conclusion is consistent with the extensive empirical study by Bühlmann and Mandozzi [2012]. Therefore, the penalty function itself is still the key for high dimensional data analysis, and it is desirable to identify penalty functions that can tolerate higher dimension.

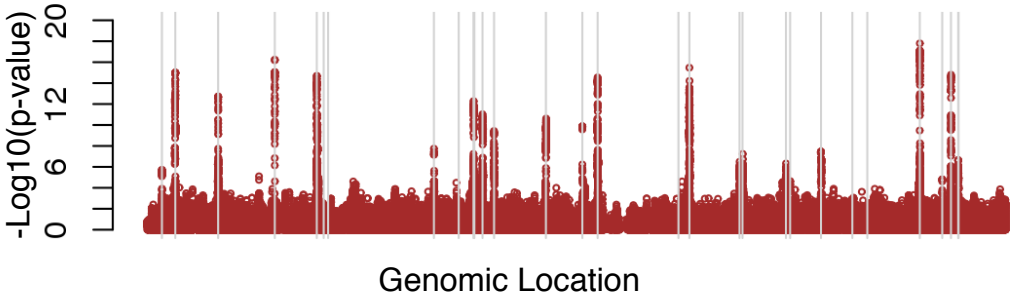

Figure 2.1: Marginal association p-values for 645,316 SNPs on chromosome 1. The grey vertical lines denote the positions of 30 important SNPs. The genomic location spans 248,484,829 base-pairs. Note that a SNP is at a single base-pair location.

Several penalty functions have been proposed for high dimensional data analysis. One of the most popular penalty functions is the Lasso penalty [Tibshirani, 1996]. The variable selection consistency of the Lasso requires the irrepresentable condition [Zhao and Yu, 2006] that there is no strong correlation between the "*important covariates*" that have non-zero effects and the "*unimportant covariates*" that have zero effects. This condition may not be satisfied in some applications, such as GWAS studies. Recent studies have shown that a class of folded concave penalties can achieve variable selection consistency without requiring such an irrepresentable condition [Fan and Lv, 2010]. These folded concave penalties include, but are not

limited to SCAD (Smoothly Clipped Absolute Deviation) [Fan, 1997; Fan and Li, 2001], MCP (Minimax Concave Penalty) [Zhang, 2010], SICA (Smooth Integration of Counting and Absolute deviation) [Lv and Fan, 2009], and a Log penalty (Friedman [2008], Sun et al. [2010]).

A common concern in real data applications of penalized estimation is to tune the regularization parameters to achieve the two fundamental goals of penalized estimation: to penalize all the noise to be zero and to obtain an unbiased estimation of the true signals. However, it may not be clear whether such "optimal" tuning is possible, and this is the focus of our study. Moreover, all the aforementioned folded-concave penalties have two tuning parameters, and thus in practice, the immediate questions concern whether they both should be tuned, and what is the consequence of tuning only one of them in order to improve computational efficiency. Previous work has provided recommendations regarding the choice of tuning parameters, but there is no systematic asymptotic study on the roles of multiple tuning parameters. To address those issues, we will relate the choice of tuning parameters to the difficulty of the variable selection problem, namely the minimum effect size and the dimensions, i.e., the number of important and unimportant covariates.

The results suggest that a class of penalty functions that bridges L_0 and L_1 penalties such as Log and SICA requires less restrictive conditions on dimensionality and minimum effect sizes, while achieving the two fundamental goals of penalized estimation. For the tuning of the regularization parameters, our study shows that both SICA and Log penalties have very limited performance if only one of the two regularization parameters is tuned, while tuning both regularization parameters can significantly improve their performances, although at the price of heavier computa-

tional burden. Our results are also insightful for designing other penalty functions. For example, our results imply that two tuning parameters are sufficient to achieve the two fundamental goals. Therefore, penalties with more than two regularization parameters may not be needed due to the substantial increase of computational cost.

We conducted empirical analyses of the penalty functions using both simulated data and real data in GWA settings. Those empirical results support the idea that the class of penalty functions that bridges L_0 and L_1 hold promise for genomic studies.

2.2 Theoretical results

2.2.1 Notations and problem setup

Let $p_{\varpi}(\beta)$ be a penalty function of β, where ϖ are regularization parameters with arbitrary dimension. $p_{\varpi}(\beta)$ is referred to as a folded concave penalty if it satisfies the following condition:

Condition 1. $p_{\varpi}(\beta)$ is concave in $\beta \in [0, \infty)$, with continuous derivative $p'_{\varpi}(\beta) \geq 0$, and $p'_{\varpi}(0+) > 0$.

We formulate the effects of the covariates via a generalized linear regression model, permitting continuous and discrete outcome variables. Consider a sample of n responses, $y = (y_1, ..., y_n)^{\mathsf{T}}$, where each $y_i, i = 1, ..., n$, is independently generated from an exponential family distribution with a density: $p(y_i|\theta_i) = \exp\{[y_i\theta_i - b(\theta_i)]/\phi + c(y_i, \phi)\}$, where θ_i is the canonical parameter and $\phi \in (0, \infty)$ is the dispersion parameter. Let x_{ij} be the value of the j-th covariate in the i-th sample, and let $X = (x_{ij})$ be a

$n \times p$ matrix of the covariates' values. We assume that X has been normalized such that $\sum_{i=1}^{n} x_{ij}^2 = n$, for $j = 1, ..., p$. Under the assumed generalized linear model, $\theta_i = \sum_{j=1}^{p} x_{ij}\beta_j$, where β_j's are regression coefficients. Let $E(y) = \mu(\theta) = (\partial_{\theta_1} b(\theta_1), ..., \partial_{\theta_n} b(\theta_n))^\mathsf{T}$ and $\Sigma(\theta) = \text{diag}\left\{\partial_{\theta_1}^2 b(\theta_1), ..., \partial_{\theta_n}^2 b(\theta_n)\right\}$. We maximize the penalized likelihood $Q_n(\beta) = l_n(\beta) - \sum_{j=1}^{p} p_\varpi(|\beta_j|)$, where $l_n(\beta) = n^{-1}\left[y^\mathsf{T}\boldsymbol{\theta} - \mathbf{1}^\mathsf{T} b(\boldsymbol{\theta})\right]$ is an affine transformation of the log-likelihood.

Without loss of generality, we assume that the first s covariates of X are important (i.e., having non-zero effect on the response variable) and denote them collectively by X_1, and then denote the remaining $p - s$ unimportant covariates by X_2, such that $X = (X_1, X_2)$. Similarly, we partition β and $\boldsymbol{\theta} = X\beta$ for the important and unimportant covariates such that $\beta = (\beta_1^\mathsf{T}, \beta_2^\mathsf{T})^\mathsf{T}$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\mathsf{T}, \boldsymbol{\theta}_2^\mathsf{T})^\mathsf{T}$. Let $\beta_0 = (\beta_{01}^\mathsf{T}, \beta_{02}^\mathsf{T})^\mathsf{T} = (\beta_{01}, ..., \beta_{0p})^\mathsf{T}$ be the true coefficients, such that $\beta_{02} = 0$. Let $\boldsymbol{\theta}_0$ be the true values of θ such that $\boldsymbol{\theta}_0 = X\beta_0$.

It is difficult to analytically study the global maximizer of the penalized likelihood. Following previous works [Fan and Lv, 2011], we study the local maximizer of the penalized likelihood that satisfies the set of sufficient and almost necessary conditions specified in Theorem 1 (see Appendix).

2.2.2 The role of the tuning parameters

The dimension of the regression problem and the minimum effect size are assumed to satisfy the following conditions:

Condition 2.1. $\log p = O(n^{\alpha})$ and $s = O(n^{\nu})$, respectively, with $0 \leq \alpha < 1$ and $0 \leq \nu < 1/2$.

Condition 2.2. $d_n \equiv 2^{-1} \min_{1 \leq j \leq s}\{|\beta_{j0}|\} = O(n^{-\gamma_0}(\log n)^{1/2})$ for some $\gamma_0 \in (\nu, 1/2)$.

The restriction of $\gamma_0 > \nu$ (which is equivalent to $s < n^{\gamma_0}$) in Condition 2.2 can be understood as an identifiability condition so that $d_n s = O(n^{\nu-\gamma_0}(\log n)^{1/2})$ can be bounded by a constant. Otherwise the response variable is unbounded, with non-trivial probability.

A maximizer of the penalized likelihood, $\hat{\beta} = (\hat{\beta}_1^{\mathsf{T}}, \hat{\beta}_2^{\mathsf{T}})^{\mathsf{T}}$, is considered to have weak oracle property if $\hat{\beta}_2 = 0$ with probability tending to 1 as $n \to \infty$, and $\hat{\beta}_1$ is consistent under L_∞ loss [Lv and Fan, 2009]. We will study the role of tuning parameters by studying the conditions for the weak oracle property. To this end, we generalize the conditions for the weak oracle property in Fan and Lv [2011] to impose constraints on the penalty function rather than particular tuning parameters, which gives the following conditions 3.1-3.3. This generalization is necessary because the original conditions are too stringent for any penalty function whose $p'_{\varpi}(0+)$ involves more than one tuning parameter. For example, the Log penalty cannot satisfy the original conditions for the weak oracle property. After generalizing the conditions, we can show that the Log penalty can indeed fulfills the conditions of the weak oracle property.

Condition 3.1. $p'_{\varpi}(d_n) \ll b_s^{-1} d_n$, where $b_s \equiv O(n^{\gamma_s}) = O(n\|[X_1^{\mathsf{T}}\Sigma(\boldsymbol{\theta}_0)X_1]^{-1}\|_\infty)$ with $\gamma_s \geq 0$. A corollary of condition 3.1 is $p'_{\varpi}(d_n) \ll d_n$.

Condition 3.2. $\left\|X_2^{\mathsf{T}}\Sigma(\boldsymbol{\theta}_0)X_1[X_1^{\mathsf{T}}\Sigma(\boldsymbol{\theta}_0)X_1]^{-1}\right\|_\infty \leq \min\left\{Kp'_\varpi(0+)/p'_\varpi(d_n), O(n^\nu)\right\}$ for $K \in (0,1)$.

Condition 3.3. $p'_\varpi(0+) \gg \max(n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}(\log n)^{1/2})$ and $p'_\varpi(0+) > \eta_p\sigma^{-1/2}(1-K)^{-1}$, where K is defined in condition 3.2, σ is a constant that is defined based on the range of the response variable y (see proposition A1 in the Supplementary Materials for details), and $\eta_p = n^{-1/2+\alpha/2}(\log n)^{1/2}$.

Condition 3.1 requires the derivative of the penalty function (i.e., the increase of penalization as the regression coefficient increases) for important covariates to be small enough. Condition 3.2 says that the ratio of the penalties' derivatives for unimportant covariates and for important ones ($p'_\varpi(0+)/p'_\varpi(d_n)$) should be large enough relative to the maximum correlation between important and unimportant covariates, which is a generalization of the irrepresentable condition for Lasso [Zhao and Yu, 2006]. Condition 3.3 requires the derivative of the penalty function for unimportant covariates to be large enough. In contrast to the conditions for the weak oracle property in Fan and Lv [2011], a critical modification is that we restrict the size of $p'_\varpi(0+)$ in condition 3.3, which replaces the condition $\lambda_n \gg n^{-\alpha}(\log n)^2$ stated in equation (18) of Fan and Lv [2011]. For SCAD and MCP, $p'_\varpi(0+) = \lambda_n$, and thus constraints on λ_n or $p'_\varpi(0+)$ are equivalent. However, for Log and SICA, $p'_\varpi(0+) = O(\lambda_n/\tau_n)$. Therefore, the generalized condition only requires the ratio of the two regularization parameters to be large enough instead of imposing a constraint on λ_n itself. Given conditions 2.1-2.2, conditions 3.1-3.3, and conditions 4.1-4.4 (presented in the Appendix), which are for the design matrix X, we have the weak oracle property (Theorem 2 in the Appendix).

One immediate conclusion from conditions 3.1-3.3 is that the constraints on the penalty function $p_{\varpi}(\beta)$ are applied on the two quantities $p'_{\varpi}(0+)$ and $p'_{\varpi}(d_n)$. With the appropriate design, two tuning parameters can give enough degrees of freedom on these two quantities so that conditions 3.1-3.3 are satisfied.

Next we discuss the implications of conditions 3.1-3.3 for the four folded concave penalties: SCAD, MCP, Log, and SICA. It is more convenient to define SCAD and MCP by their derivatives.

$$p'_{\texttt{SCAD}}(|\beta_j|;\lambda,a) = \{\lambda I(|\beta_j| \le \lambda) + [(a\lambda - |\beta_j|)/(a-1)]I(\lambda < |\beta_j| < a\lambda)\},$$

where $\lambda > 0$ and $a > 2$ are two regularization parameters.

$$p'_{\texttt{MCP}}(|\beta_j|;\lambda,a) = I(|\beta_j| < a\lambda)(a\lambda - |\beta_j|)/a,$$

where $\lambda > 0$ and $a > 0$ are two regularization parameters. The Log and SICA penalties are defined as

$$p_{\texttt{log};\lambda,\tau}(|\beta_j|) = \lambda \log(|\beta_j| + \tau), \text{ and}$$

$$p_{\texttt{SICA}}(|\beta_j|;\lambda,\tau) = \lambda\{I(|\beta_j| \neq 0)|\beta_j|/(|\beta_j| + \tau) + \tau|\beta_j|/(|\beta_j| + \tau)\},$$

respectively, where $\lambda > 0$ and $\tau > 0$ are two regularization parameters. In the following discussions, the tuning parameters employed by a penalty are indicated by subscripts. For example, the SCAD penalty with one tuning parameter λ_n (the other regularization parameter a being set as constant) is denoted by SCAD_{λ_n} and the SCAD penalty with two tuning parameters λ_n and a_n is denoted by $\text{SCAD}_{\lambda_n,a_n}$.

Let $\eta_p = n^{-1/2+\alpha/2}(\log n)^{1/2}$, which is a monotone transformation of dimension $\log(p) = O(n^{\alpha})$. Let $\eta_d = \min(n^{\gamma_0/2}(\log n)^{-1/4}, n^{-\gamma_0+1/2})$, which, by condition 2.2, is

a function of the minimum effect size: $d_n \equiv \min_{1\leq j\leq s}\{|\beta_{j0}|\} = O(n^{-\gamma_0}(\log n)^{1/2})$. In the following propositions, we will discuss the properties of different penalties with respect to s (the number of non-zero coefficients), d_n, η_d, and η_p.

Proposition 1. [SCAD_{λ_n}, $\text{SCAD}_{\lambda_n,a_n}$, or MCP_{λ_n}] If $d_n \gg \eta_p$ and $s \ll \eta_d$, there exist λ_n such that $d_n \gg \lambda_n > \eta_p$ to satisfy conditions 3.1-3.3 for the weak oracle property. However, there is no such tuning parameter if $d_n \ll \eta_p$.

Proposition 2. [$\text{MCP}_{\lambda_n,a_n}$] There are tuning parameters that satisfy conditions 3.1-3.3 for the weak oracle property without further constraints other than $s \ll n^{\gamma_0}$, as is specified in condition 2.2.

Proposition 3. [SICA_{λ_n} or Log_{λ_n}] There are tuning parameters that satisfy conditions 3.1-3.3 for the weak oracle property if $d_n \gg \eta_p$, $s \ll \eta_d$, and

$$\left\|X_2^\mathsf{T}\Sigma(\boldsymbol{\theta}_0)X_1(X_1^\mathsf{T}\Sigma(\boldsymbol{\theta}_0)X_1)^{-1}\right\|_\infty \leq K\left(d_n/\tau+1\right)^2,$$

where $K \in (0,1)$ was defined in condition 3.3. There is no such tuning parameter if $d_n \ll \eta_p$.

Proposition 4. [$\text{SICA}_{\lambda_n,\tau_n}$ or $\text{Log}_{\lambda_n,\tau_n}$] There are tuning parameters that satisfy conditions 3.1-3.3 for the weak oracle property without further constraints other than $s \ll n^{\gamma_0}$, as is specified in condition 2.2.

Corollary 1. [Restriction on tuning parameter if $d_n \ll \eta_p$] To satisfy condition 3.1-3.3 requires $a_n \to 0+$ for $\text{MCP}_{\lambda_n,a_n}$, and $\tau_n \to 0+$ for $\text{SICA}_{\lambda_n,\tau_n}$ and $\text{Log}_{\lambda_n,\tau_n}$.

The proofs of Propositions 1-4 and Corollary 1 are presented in the Supplementary Materials.

By Proposition 1, if $d_n \gg \eta_p$ or $d_n \ll \eta_p$, SCAD has similar theoretical properties when one or two tuning parameters are used. This conclusion is consistent with many previous works where SCAD has satisfactory performance when the regularization parameter a is set to be a constant, e.g., 3.7. Using two tuning parameters (λ_n and a_n) does have some advantage over one tuning parameter (λ_n) when $d_n = O(\eta_p)$. However, since the situation of $d_n = O(\eta_p)$ only covers a negligible part of the space for d_n, we do not discuss it further here. Proposition 1 also states that if $d_n \ll \eta_p$, in other words, if the effect size is not large enough relative to the dimension, then there is no tuning parameter of SCAD to satisfy conditions 3.1-3.3. Specifically, condition 3.1 requires $p'_\varpi(d_n) \ll d_n$, and condition 3.3 requires $p'_\varpi(0+) > c\eta_p$, where c is a constant. These two conditions cannot both be satisfied if $d_n \ll \eta_p$. Specifically, if SCAD satisfies condition 3.3, then $p'_\varpi(0+) = \lambda_n > c\eta_p$. Given $d_n \ll \eta_p$ and $\eta_p < \lambda_n/c$, we have $d_n \ll \lambda_n$, and then we can show that $p'_\varpi(d_n) = \lambda_n$, which contradicts condition 3.1. In addition, we can see that in this situation, both $p'_\varpi(0+)$ and $p'_\varpi(d_n)$ are functions of λ_n so that a plays no role in fulfilling conditions 3.1 and 3.3. On the other hand, tuning only one regularization parameter is a computational advantage of SCAD.

By Propositions 1 and 2, tuning both λ_n and a_n significantly improves the performance of MCP if $d_n \ll \eta_p$. Specifically, if MCP satisfies condition 3.3, then $p'_\varpi(0+) = \lambda_n > c\eta_p$. Then given $d_n \ll \eta_p$, we have $d_n \ll \lambda_n$. However, given a properly tuned $a_n = o(1)$ such that $d_n \geq a_n\lambda_n$, we have $p'_\varpi(d_n) = 0$, which allows MCP to satisfy condition 3.1.

By Proposition 3, if we set $\tau = O(1)$ and only tune the regularization parameter λ, then SICA_{λ_n} and Log_{λ_n} require the following condition to achieve the weak oracle property:

$$\left\| X_2^{\mathsf{T}} \Sigma(\boldsymbol{\theta}_0) X_1 (X_1^{\mathsf{T}} \Sigma(\boldsymbol{\theta}_0) X_1)^{-1} \right\|_\infty \leq K \left(d_n/\tau + 1 \right).$$

This condition is similar to the irrepresentable condition of Lasso because when $\tau = O(1)$, $d_n/\tau + 1 \to 1$. Therefore, asymptotically SICA_{λ_n} and Log_{λ_n} would perform in a way similar to Lasso. If $d_n \ll \eta_p$, then SICA_{λ_n} and Log_{λ_n} cannot simultaneously satisfy conditions 3.1 and 3.3, even if the irrepresentable condition is satisfied.

By Proposition 4, tuning both λ_n and τ_n significantly improves the performance of SICA and Log. Specifically, SICA and Log can have satisfactory variable selection performances even if the minimum effect size is much smaller with respect to the dimension of the problem: $d_n \ll \eta_p$. This can be justified by the following arguments. For Log penalty, $p'_\varpi(d_n) = p'_\varpi(0+)/(d_n/\tau_n + 1)$. Even condition 3.3 requires a large value of $p'_\varpi(0+)$; a small enough τ_n can help $p'_\varpi(d_n)$ to satisfy condition 3.1. SICA has similar properties since it has $p'_\varpi(d_n) = p'_\varpi(0+)/(d_n/\tau_n + 1)^2$. Therefore, the implications of Proposition 3 and Proposition 4 for the practical use of SICA and Log penalties would be that we should not treat τ as a constant.

Corollary 1 shows that for a difficult variable selection problem where $d_n \ll \eta_p$, the tuning parameter a_n of MCP or τ_n of SICA or Log should be on the scale of $o(1)$. Zhang [2010] suggests that a larger tuning parameter a in MCP leads to a bigger bias and less accurate variable selection, $a = 1$ leads to a singularity problem, and $a < 1$ leads to a dramatic increase in computational cost. Similarly, Lv and Fan [2009] suggest that for penalized estimates using SICA, the bias decreases to 0 as τ_n goes to 0+, but the computational difficulty increases because the maximum concavity

goes to infinity. Similar conclusions apply to the Log penalty. Although $\text{MCP}_{\lambda_n,a_n}$, $\text{SICA}_{\lambda_n,\tau_n}$, and $\text{Log}_{\lambda_n,\tau_n}$ have similar theoretical properties by Propositions 2 and 4, the following numerical studies show that the computation cost for SICA and Log is more affordable than that of MCP.

2.3 Algorithm and tuning parameter selection

We obtain the penalized estimates using SCAD or MCP by the coordinate descent algorithms implemented in the R package NCVREG [Breheny and Huang, 2011]. We implement the penalized estimation using SICA and Log penalties by a combination of the coordinate descent algorithm and Local Linear Approximation (LLA) [Zou and Li, 2008]. Specifically, we update the estimate of each regression coefficient sequentially (which is the coordinate decent part), and the solution of each coefficient is obtained after applying a local linear approximation:

$$p_{\varpi}\left(|\beta_j|\right) \approx p_{\varpi}\left(|\hat{\beta}_j^{(k)}|\right) + p'_{\varpi}\left(|\hat{\beta}_j^{(k)}|\right)\left(|\beta_j| - |\hat{\beta}_j^{(k)}|\right),$$

where $\hat{\beta}_j^{(k)}$ is the estimate of regression coefficient β_j at the k-th iteration.

We present the computational algorithms for linear and logistic regression separately. The objective function for linear regression is:

$$Q_n(\beta) = -\frac{1}{2n}\left(y - X\beta\right)^{\mathsf{T}}\left(y - X\beta\right) - \sum_{j=1}^{p} p_{\varpi}(|\beta_j|).$$

After applying LLA for the penalty function, the objective function to be maximized at the $(k+1)$-th step, while solving for β_j, is

$$Q_n^{(k+1)}(\beta_j) = -\frac{1}{2n}\|y - X_{-j}\hat{\beta}_{-j}^{(k)} - x_j\beta_j\|^2 + \sum_{j=1}^{p} p'_{\varpi}\left(|\hat{\beta}_j^{(k)}|\right)|\beta_j|,$$

where X_{-j} is the matrix X without the jth column, and $\hat{\beta}_{-j}^{(k)}$ is $\hat{\beta}^{(k)}$ without the jth element. By letting $\partial Q_n^{(k+1)}(\beta_j)/\partial\beta_j = 0$, we can obtain the solution for β_j

$$\begin{cases} \hat{\beta}_j^{(k+1)} = 0 & \text{if } |z_j^{(k)}| \leq v_j^{-1} p'_\varpi \left(|\hat{\beta}_j^{(k)}|\right) \\ \hat{\beta}_j^{(k+1)} = \text{sgn}(\hat{\beta}_j^{(k)}) \left[|z_j^{(k)}| - v_j^{-1} p'_\varpi \left(|\hat{\beta}_j^{(k)}|\right)\right] & \text{if } |z_j^{(k)}| > v_j^{-1} p'_\varpi \left(|\hat{\beta}_j^{(k)}|\right) \end{cases},$$

where $z_j^{(k)} = x_j(y - X_{-j}\beta_{-j}^{(k)})/v_j$, and $v_j = x_j^{\mathsf{T}} x_j$.

The penalized likelihood for logistic regression is

$$Q_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left\{ y_i \log\frac{\pi_i}{1-\pi_i} + \log(1-\pi_i)\right\} - \sum_{j=1}^{p} p_\varpi(|\beta_j|),$$

where $\pi_i = \Pr(y_i = 1)$. By applying the iteratively reweighted least squares algorithm [McCullagh and Nelder, 1989] and the LLA of the penalty function, the objective function to be maximized at the $(k+1)$-th step, while solving for β_j, is

$$\begin{aligned} Q_n^{(k+1)}(\beta_j) \;\approx\; & -\frac{1}{2n}\left(\tilde{y}^{(k)} - X_{-j}\hat{\beta}_{-j}^{(k)} - x_j\beta_j\right)^{\mathsf{T}} W^{(k)} \left(\tilde{y}^{(k)} - X_{-j}\hat{\beta}_{-j}^{(k)} - x_j\beta_j\right) \\ & + \sum_{j=1}^{p} p'_\varpi \left(|\hat{\beta}_j^{(k)}|\right)|\beta_j|, \end{aligned}$$

where $\tilde{y}^{(k)} = X\hat{\beta}^{(k)} + \left(W^{(k)}\right)^{-1}(y - \boldsymbol{\pi}^{(k)})$, $W^{(k)}$ is a diagonal matrix with the i-th diagonal element $w_i^{(k)} = \pi_i^{(k)}(1-\pi_i^{(k)})$, and $\pi_i^{(k)} = \exp\left(X\hat{\beta}^{(k)}\right) / \left[1 + \exp\left(X\hat{\beta}^{(k)}\right)\right]$. Letting $\partial Q_n^{(k+1)}(\beta_j)/\partial\beta_j = 0$, the estimate of β_j is

$$\begin{cases} \hat{\beta}_j^{(k+1)} = 0 & \text{if } |z_j^{(k)}| \leq v_j^{-1} p'_\varpi \left(|\hat{\beta}_j^{(k)}|\right) \\ \hat{\beta}_j^{(k+1)} = \text{sgn}(\hat{\beta}_j^{(k)}) \left[|z_j^{(k)}| - v_j^{-1} p'_\varpi \left(|\hat{\beta}_j^{(k)}|\right)\right] & \text{if } |z_j^{(k)}| > v_j^{-1} p'_\varpi \left(|\hat{\beta}_j^{(k)}|\right) \end{cases},$$

where $z_j^{(k)} = x_j^{\mathsf{T}} W^{(k)}(\tilde{y}^{(k)} - X_{-j}\beta_{-j}^{(k)})$ and $v_j = x_j^{\mathsf{T}} W^{(k)} x_j$.

The iterative estimation process ends if the maximum difference of the estimates of β between consecutive iterations is less than 10^{-5}.

We follow a strategy similar to the ones in Breheny and Huang [2011] to obtain an initial set of tuning parameter combinations. For SCAD and MCP, the tuning parameter a is given as a constant or a vector of legitimate values such as $a > 2$ for SCAD and $a > 1$ for MCP (the implementation of MCP in the R package NCVREG requires $a > 1$). The λ's for SCAD and MCP are given as N numbers equally spaced on a log scale, with the largest one corresponding to the largest marginal effect size and the smallest one being a fraction of the largest one. In our experience, the fraction is set as 1/10 from the linear model, and 1/100 for the logistic model.

For SICA and Log, the tuning parameter τ is set as a constant or a vector of legitimate values such as $\tau > 0$. The theoretical results in previous sections suggest that τ should be much smaller than the minimum effect size. In practice, because we do not know which set of variables is important, we use the largest marginal effect size as the upper bound for τ. Neither λ nor τ alone determines the penalization strength. Instead, their combination in the form of the threshold $v_j^{-1} p'_\varpi(|\hat{\beta}_j^{(k)}|)$ specifies the penalization strength. Without loss of generality, we assume x_j $(j = 1, ..., p)$ is standardized with mean 0 and $v_j = \sum_{i=1}^n x_{ij}^\mathsf{T} x_{ij} = n$. It follows that the thresholds for SICA and the Log penalties are $p'_\varpi(0)/v_j = p'_\varpi(0)/n$. The largest threshold corresponds to the largest marginal coefficient estimates (by absolute value), denoted by $\hat{\beta}_M$, a predefined number of τ's uniformly distributed on a log scale from 10^{-6} to $\hat{\beta}_M$, and the smallest threshold is 1/10 of the largest one, i.e., $\hat{\beta}_M/10$ for the linear model, and 1/100 for the logistic model respectively:

$$\{\texttt{Threshold}_1, ..., \texttt{Threshold}_N\} = \left\{ \hat{\beta}_M, ..., \frac{\hat{\beta}_M}{100} \right\}.$$

For example, for Log penalty, the threshold in the first iteration is $\lambda/(n\tau)$. Then given a specific value of τ and a set of thresholds, N λ's can be generated based on

the equation:

$$\{\lambda_1/n\tau, ..., \lambda_N/n\tau\} = \{\texttt{Threshold}_1, ..., \texttt{Threshold}_N\}.$$

A similar strategy is used to determine the initial set of tuning parameters for SICA.

We select a particular combination of tuning parameters from the initial tuning parameter pool using the extended BIC [Chen and Chen, 2008, 2012]. As discussed in Chen and Chen [2008], if $\log p/\log n > 0.5$, the conventional BIC [Schwarz, 1978] is not consistent. In all the scenarios considered in this paper, $\log p/\log n > 1$. Our empirical studies confirm that in these scenarios the conventional BIC tends to be too liberal, and the extended BIC performs satisfactorily. The extended BIC for the linear model m is:

$$\texttt{BIC}_\varrho(m) = -2\log l_n\{\hat{\theta}(m)\} + df_m \log n + 2\varrho \log \varsigma(S_{df_m}),$$

where df_m is the degrees of freedom for model m and $\varsigma(S_{df_m})$ is the number of the models containing df_m covariates. We take the number of the nonzero coefficient estimates in the model m as df_m and set $\varsigma(S_{df_m}) = \binom{p}{df_m}$, the number of combinations of df_m covariates chosen from p covariates. In addition, we set $\varrho \simeq 1 - 1/(2\log p/\log n)$ while $\varrho > 1 - 1/(2\log p/\log n)$ is suggested in Chen and Chen [2008]. The extended BIC for a generalized linear model m is:

$$\texttt{BIC}_\varrho(m) = -2\log l_n\{\hat{\theta}(m)\} + df_m \log n + 2df_m \varrho \log p,$$

where df_m is the number of nonzero coefficient estimates, and similar to the above $\varrho \simeq 1 - 1/(2\log p/\log n)$, as suggested in Chen and Chen [2012].

2.4 Simulation

We evaluated those four penalties using a set of simulated data for multiple loci mapping problems. Specifically, the response variable is either a continuous trait (linear regression) or the case/control status (logistic regression), and the covariates are the genotypes of the SNPs. One particular challenge in a multiple loci mapping problem is that nearby SNPs often have correlated genotypes due to linkage disequilibrium, and such correlations may violate the irrepresentable condition, which is needed for the consistency of Lasso. To faithfully reproduce such correlation structure, we directly used genotype data of European Ancestry (EA) samples from a GWAS study of schizophrenia [Shi et al., 2009]. The dataset was obtained from NCBI dbGaP, which includes GAIN (Genetic Association Information Network) samples (2,686/2,656: cases/controls, dbGaP Accession: phs000021.v3.p2) and non-GAIN samples (1,217/1,442: cases/controls, dbGaP Accession: phs000167.v1.p1) genotyped by Affymetrix 6.0 SNP arrays with ∼900,000 SNPs.

To compare the performances of those penalty functions, we use two criteria to select the tuning parameters. One is the extended BIC as introduced earlier, and the other is an oracle criterion that uses the knowledge of the true model to select the tuning parameters. Certainly the oracle criterion is not applicable in practice when the true model is unknown. However, in simulation studies, the oracle criterion permits us to evaluate the performance of a penalty function rather than the combined outcome of a penalty function and a tuning parameter selection method. The oracle criterion is defined as follows. Let $\mathtt{D}$ be the number of discoveries, i.e., the covariates with non-zero regression coefficient estimates. $\mathtt{D} = \mathtt{TD} + \mathtt{FD}$, where $\mathtt{TD}$ and $\mathtt{FD}$ are the number of true discoveries and false discoveries, respectively. Our oracle criterion evaluates a model based on the three measures, the false discovery rate $\mathtt{FD/D}$, power

$\texttt{TD}/s$, and the sum of squared error of regression coefficient estimates $\sum_{j=1}^{p} |\hat{\beta}_j - \beta_{0j}|^2$, where β_{0j} is the true value of β_j. The model with the minimum of $\texttt{wt}(\texttt{FD/D} - \texttt{TD}/s) + \sum_{j=1}^{p} |\hat{\beta}_j - \beta_{0j}|^2$ is selected, where `wt` is a weight to balance the number of true/false discoveries and bias. Models selected with larger `wt` tend to have more true discoveries and fewer false discoveries, but have a larger bias in their regression coefficient estimates.

2.4.1 Linear model

For computational efficiency when there are a large number of simulations, we randomly selected $n = 222$ samples and 12,656 SNPs with no missing values, and with a minor allele frequency greater than 5% on chromosome 20. The response variables y were simulated by $y = X\beta + \epsilon$, where $\epsilon \sim N(\mathbf{0}, I_{n\times n})$. We considered 3 situations involving different combinations of p and s: p = 12,656 and s = 12, 16, or 20. Let $u_1^{\mathsf{T}} = (0.5, -0.5, 0.4, -0.4)$. When s = 12, 16, and 20, β_0 are set by repeating u_1 three, four, and five times, respectively. In addition, we considered null situations with $s = 0$ and p = 12,656.

The tuning parameter grids were chosen as follows: a = (2.1, 2.5, 3.0, 3.7, 4.5, 6.0) for SCAD, a = (1.1, 2.0, 3.0, 4.0, 5.0, 6.0) for MCP, and 6 τ's for Log and SICA as described in the section 3. We also applied Lasso implemented in R/GLMNET. For each of these five penalties, 100 λ's uniformly distributed on a log scale were generated as described in section 3.

We used the extended BIC and oracle criteria $10(\texttt{FD/D} - \texttt{TD}/s) + \sum_{j=1}^{p} |\hat{\beta}_j - \beta_{j0}|^2$ to select the tuning parameters. We give the term $(\texttt{FD/D} - \texttt{TD}/s)$ a larger weight

of 10 so that the oracle criterion selects the model with the smaller false discovery rate FD/D, greater power TD/s first, and use the sum of squared error of regression coefficient estimates $\sum_{j=1}^{p} |\hat{\beta}_j - \beta_{j0}|^2$ as a secondary criterion.

For null simulation situations, all penalties have at most 1 or 2 false discoveries by the extended BIC tuning parameter selection criterion. Table 2.1 summarizes the simulation results in non-null situations with 12, 16, or 20 important covariates. The folded concave penalties perform better than the Lasso penalty. Among the four folded concave penalties, SICA, Log and MCP have comparable performance, and are better than SCAD when the tuning parameters are selected by the oracle criterion. When the tuning parameters are selected by the extended BIC, SICA and Log have comparable performance, and are better than SCAD and MCP. In additional simulation studies (results not shown), SCAD and MCP with one tuning parameter (λ) have slightly worse performance than the situations with two tuning parameters. In contrast, Log and SICA with one tuning parameter (λ) have much worse performance than the situations with two tuning parameters. Therefore, the extra tuning parameter (a or τ) gives SCAD and MCP limited additional advantage, but significantly improves the performances of Log and SICA.

Table 2.1: Simulation results for penalized linear regression with (n=222, p = 12,656). The headers indicate the tuning parameter selection criterion (Oracle or the extended BIC) and the numbers in parentheses are the number of important covariates. For each penalty, we present the median of the number of true discoveries, false discoveries (in parentheses), and average bias of the true discoveries (in brackets) across 100 simulations.

	Oracle (12)	Ext BIC (12)	Oracle (16)	Ext BIC (16)	Oracle (20)	Ext BIC (20)
Lasso	11 (8) [0.33]	0 (0) [−]	7 (3) [0.39]	0 (0) [−]	14 (112) [0.34]	0 (0) [−]
SCAD	11 (3) [0.28]	0 (0) [−]	15 (25) [0.13]	0 (0) [−]	19 (27) [0.12]	0 (0) [−]
MCP	11 (1) [0.08]	10 (20) [0.08]	14 (2) [0.07]	11 (39) [0.10]	17 (3) [0.08]	5 (39) [0.11]
Log	11 (1) [0.07]	10 (3) [0.07]	14 (3) [0.07]	11 (7) [0.07]	17 (3) [0.08]	8 (10) [0.08]
SICA	11 (1) [0.06]	10 (3) [0.06]	14 (2) [0.06]	11 (6) [0.07]	17 (4) [0.07]	5 (7) [0.08]

2.4.2 Simulation for logistic model

For penalized logistic regression, a larger sample size is needed for simulations with reasonable effect sizes. We randomly selected 10,156 SNPs (with a minor allele frequency larger than 5%) from chromosomes 1 to 22 and X and 750 samples (with a missing values percent smaller than 3%). We simulated the individual SNP effect so that the disease odds ratios are 2.0, corresponding to regression coefficients of 0.7. The binary response variable y was simulated based on the logistic regression model: $\log\{\Pr(y=1)/\Pr(y=0)\} = X\beta$, where $s = 4$, 8, or 12. In addition, the null model where $s = 0$ was simulated. The intercept was set as -2, corresponding to a disease prevalence of 0.12. The initial pool of tuning parameters were generated in the same way as linear regression, and then a particular combination of tuning parameters was selected to minimize the extended BIC, or an oracle criterion $10(\texttt{FD/D} - \texttt{TD}/s) + \sum_{j=1}^{p} |\hat{\beta}_j - \beta_{j0}|^2$.

For the simulation of null models, all penalties have at most 1 or 2 false discoveries by the extended BIC tuning parameter selection criterion. The simulation results of non-null models are shown in Table 2.2. In general, the results of logistic model simulation have a trend similar to that of linear model simulation. When the oracle criterion is used, all penalties have satisfactory variable selection performances, although SICA and Log have a smaller bias on effect size estimation. It can be observed that the models chosen by the oracle criterion are different from those selected by the extended BIC for SCAD and MCP. This is because the models chosen by the oracle criterion tend to have larger biases, which reduces the likelihood, and thus increases the realized value of the extended BIC. On the other hand, for Log and SICA, the models chosen by the oracle criterion are similar to those chosen by the extended BIC. Additional simulations (results not shown) confirm that SCAD with

one or two tuning parameters have similar performance, and an additional tuning parameter improves MCP's performance. Moreover, the additional tuning parameter significantly improves the performance of the SICA and Log penalties.

Finally, Table 2.3 presents the comparison of the computational burden for MCP, Log and SICA across various values of a and τ, respectively. It can be observed that the computation time of Log and SICA is much less than that of MCP.

In summary, Log and SICA have a smaller bias for the coefficient estimates of important covariates, and therefore, more accurate estimates of the likelihood function. In addition, they have lower computational burden compared to MCP. As a consequence, Log and SICA penalties have advantages in empirical usage.

Table 2.2: Simulation results for penalized logistic regression (n=750, p = 10,156). The headers indicate the tuning parameter selection criterion (Oracle or the extended BIC) and the numbers in parentheses are the number of important covariates. For each penalty, we present the median of the number of true discoveries, the number of false discoveries (in parentheses), and the average bias of true discoveries (in brackets) across 100 simulations.

	Oracle (4)	Ext BIC (4)	Oracle (8)	Ext BIC (8)	Oracle (12)	Ext BIC (12)
Lasso	4(0) [0.49]	4 (0) [0.47]	7(0) [0.55]	6 (0) [0.53]	11(2) [0.59]	0 (0) [–]
SCAD	4 (0) [0.48]	4 (0) [0.39]	7 (0) [0.53]	6 (0) [0.43]	11(2) [0.58]	0 (0) [–]
MCP	4 (0) [0.093]	4 (0) [0.097]	7 (0) [0.25]	6 (1) [0.14]	11(1) [0.32]	11 (7) [0.25]
Log	4 (0) [0.085]	4 (0) [0.096]	7 (0) [0.085]	7 (1) [0.09]	11(1) [0.10]	11 (1) [0.10]
SICA	4 (0) [0.084]	4 (0) [0.094]	7 (0) [0.095]	7 (1) [0.099]	11(1) [0.12]	11 (1) [0.096]

Table 2.3: Running time rounded to minutes per simulation (n=750, s = 12, p = 10,156) for 100 λ's and a fixed a of MCP or τ of Log and SICA.

MCP	21.1 ($a = 1.1$)	5.2 ($a = 2.0$)	7.1 ($a = 3.0$)	6.3 ($a = 4.0$)	9.7 ($a = 5.0$)
Log	2.1 ($\tau = 10^{-6}$)	1.9 ($\tau = 10^{-5}$)	1.9 ($\tau = 10^{-4}$)	1.9 ($\tau = 10^{-3}$)	1.8 ($\tau = 0.6$)
SICA	2.0 ($\tau = 10^{-6}$)	2.1 ($\tau = 10^{-5}$)	1.9 ($\tau = 10^{-4}$)	1.8 ($\tau = 10^{-3}$)	1.8 ($\tau = 0.6$)

2.5 Real data analysis

We analyzed the data of GWA studies of schizophrenia on European-ancestry samples (2,195 cases vs. 2,617 controls). The missing genotypic data were imputed using BEAGLE software [Browning and Browning, 2007], and 677,163 autosome SNPs with minor allele frequency no less than 5% were selected for the analysis. We included 23 principle components (PCs) of genotype data in the model to account for possible population stratification. First, a univariate logistic regression is conducted on the case-control status for each of the 677,163 SNPs, conditioning on the covariates: age, gender and 23 PCs. Using the resulting 677,163 p-values, we calculated a genomic control factor of 1.0445 [Devlin and Roeder, 1999], implying that there is no strong population stratification not accounted for in our model. The 7,984 SNPs with p-values smaller than 0.01 were selected for the following variable selection. We applied the penalized logistic regression on the 7,984 SNPs and 4,812 samples with the four folded-concave penalties, while accounting for the effects of age, gender and 23 PCs, by including them as unpenalized covariates.

We applied SCAD with $a = 3.7$ and MCP with $a = 3$, the default value of R package NCVREG, and chose to use two tuning parameters for SICA and Log. Using the extended BIC for tuning parameter selection, the penalized logistic regressions with Log and SICA selected 38 and 22 SNPs, respectively (Supplementary Table 1-2). However, penalized logistic regressions with both MCP and SCAD selected the null model since the null model has the lowest value of the extended BIC.

A joint model was fitted by a logistic regression using the 38 SNPs identified by the Log penalty together with age, gender, and 23 PCs to obtain the p-values for the 38 SNPs. The results are illustrated in Figure 3.5, together with the marginal p-values for the 677,163 SNPs. There are 43 genes within 10kb distance of these 38 SNPs,

and among them 21 are in the Database for Annotation, Visualization and Integrated Discovery (DAVID) [Huang et al., 2008]. By functional category enrichment analysis at the DAVID website, 16 of the 21 genes are bound by transcription factor FOXO1, with significant enrichment p-value after a Benjamini correction. Recent studies have shown that FOXO1 regulates neuroblastoma differentiation [Mei et al., 2012], which is relevant to schizophrenia. In contrast, we also did the functional category analysis for those genes within 10 kb of the 38 SNPs with the smallest marginal p-values, but no functional category was significantly over-represented.

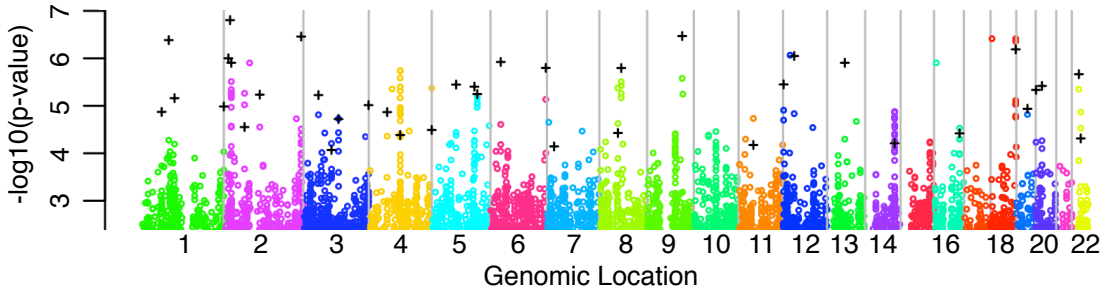

Figure 2.2: GWA marginal p-values (colored circles) and the 38 SNPs (black crosses) identified by penalized logistic regression using Log penalty.

2.6 Asymptotic results

We present the following Theorem 1 of Fan and Lv [2011] for the self-completeness of this paper. This Theorem gives a set of sufficient and almost necessary conditions of a local maximizer of the penalized likelihood.

Theorem 1. (Characterization of PMLE): $\hat{\beta} \in R^p$ is a strict local maximizer of

the non-concave penalized likelihood $Q_n(\beta) = l_n(\beta) - \sum_{j=1}^{p} p_\varpi(|\beta_j|)$ if

$$X_1^\mathsf{T}\mu(\hat{\boldsymbol{\theta}}) - X_1^\mathsf{T}y + np'_\varpi(\hat{\beta}_{01}) = 0 \tag{2.6.1}$$

$$\|X_2^\mathsf{T}(y - \mu(\hat{\boldsymbol{\theta}}))\|_\infty - np'_\varpi(0+) < 0 \tag{2.6.2}$$

$$\lambda_{\min}\left(X_1^\mathsf{T}\Sigma(\hat{\boldsymbol{\theta}})X_1\right) - n\kappa(p_\varpi, \hat{\beta}_{01}) > 0. \tag{2.6.3}$$

The following conditions 4.1-4.4 are for the design matrix X, and they are essentially the same as the corresponding conditions from Fan and Lv [2011]. We first define a few notations used in the following regularity conditions. L_∞ norm of a matrix is the maximum of the L_1 norm of each row. $\lambda_{\max}()/\lambda_{\min}()$ denotes the maximum/minimum eigen-value of a symmetric matrix, respectively. Denote a neighborhood of the non-zero coefficients as $\mathcal{N}_0 = \{\delta \in R^s : \|\delta - \beta_{01}\|_\infty \le d_n\}$.

Condition 4.1. $\|[X_1^\mathsf{T}\Sigma(\boldsymbol{\theta}_0)X_1]^{-1}\|_\infty = O(b_s n^{-1})$, where

$$b_s = O(n^{\gamma_s}) \ll \min(n^{1/2-\gamma_0}, n^{\gamma_0-\nu}(\log n)^{-1/2}) \text{ and } \gamma_s \ge 0.$$

Condition 4.2 $\max_{\delta\in\mathcal{N}_0}\max_{j=1}^{p}\lambda_{\max}[X_1^\mathsf{T}|x_j|\mathrm{diag}\{|\mu''(X_1\delta)|\}X_1] = O(n)$, where the derivative $\mu''(X_1\delta)$ is taken component-wise.

Condition 4.3 $\max_{j=1}^{p}||x_j||_\infty = o(n^{(1-\alpha)/2}(\log n)^{-1/2})$ if the responses are unbounded.

Condition 4.4 $\max_{\delta\in\mathcal{N}_0}\kappa(p_\varpi, \delta) \le \min_{\delta\in\mathcal{N}_0}\lambda_{\min}[n^{-1}X_1^\mathsf{T}\Sigma(X_1\delta)X_1]$, where $\kappa(p_\varpi, \delta)$ is defined as the local concavity of a penalty function at $v = (v_1, ..., v_q)^\mathsf{T}$:

$$\kappa(p_\varpi, v) = \lim_{\epsilon\to 0+}\max_{1\le j\le q}\sup_{t_1<t_2\in(|v_j|-\epsilon,|v_j|+\epsilon)} -\frac{p'_\varpi(t_2) - p'_\varpi(t_1)}{t_2 - t_1}.$$

For the penalties with continuous second derivatives, $\kappa(p_\varpi, v) = \max_{1\le j\le q} -p''_\varpi(v_j)$.

Given conditions 1 to 4, we have the following weak oracle property.

Theorem 2. (Weak oracle property) Given the conditions 1 to 4, with probability at least $P_{\texttt{converage}} = 1 - 2\left[sn^{-1} + (p-s)\exp\left(-n^\alpha \log n\right)\right]$, there exists a penalized likelihood estimator $\hat\beta = (\hat\beta_1^\mathsf{T}, \hat\beta_2^\mathsf{T})^\mathsf{T}$ which satisfies
(a) Sparsity: $P(\hat\beta_2 = \mathbf{0}) \to 1$, (b) L_∞ loss: $\|\hat\beta_1 - \beta_{10}\|_\infty = o(n^{-\gamma_0}\sqrt{\log n})$.

Lemma 1 (for proofs of the propositions 2 and 3)

For condition 3.3, if $s = O(n^\nu) \ll \min(n^{\gamma_0/2}(\log n)^{-1/4}, n^{-\gamma_0+1/2})$, then

$$\max(n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n}) \ll n^{-\gamma_0}\sqrt{\log n} = O(d_n). \tag{2.6.4}$$

Proof:

If $1/3 < \gamma_0 < 1/2$, then $\min(n^{\gamma_0/2}(\log n)^{-1/4}, n^{-\gamma_0+1/2}) = n^{-\gamma_0+1/2}$

$$\begin{aligned}
\max(n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n}) &= \max(s^2 n^{-2\gamma_0}\log n, sn^{-1/2}\sqrt{\log n}) \\
&\ll \max(n^{1-4\gamma_0}\log n, n^{-\gamma_0}\sqrt{\log n}) \\
&= n^{-\gamma_0}\sqrt{\log n} = O(d_n).
\end{aligned}$$

If $0 \leq \gamma_0 \leq 1/3$, then $\min(n^{\gamma_0/2}(\log n)^{-1/4}, n^{-\gamma_0+1/2}) = n^{\gamma_0/2}(\log n)^{-1/4}$,

$$\begin{aligned}\max(n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n}) &= \max(s^2 n^{-2\gamma_0}\log n, sn^{-1/2}\sqrt{\log n}) \\ &\ll \max(n^{-\gamma_0}\sqrt{\log n}, n^{\gamma_0/2-1/2}(\log n)^{1/4}) \\ &= n^{-\gamma_0}\sqrt{\log n} = O(d_n).\end{aligned}$$

Therefore, $\max(n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n}) \ll n^{-\gamma_0}\sqrt{\log n} = O(d_n)$.

Lemma 2 (for proofs of propositions 2 and 3)

For condition 3.3, if $s \ll \min(n^{\gamma_0/2-\gamma_s/2}(\log n)^{-1/4}, n^{-\gamma_0-\gamma_s+1/2})$, then

$$\max(n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n}) \ll n^{-\gamma_0-\gamma_s}\sqrt{\log n} = O(b_s^{-1}d_n). \qquad (2.6.5)$$

Proof:

If $\gamma_0 + \gamma_s/3 > 1/3$, then $\min(n^{\gamma_0/2-\gamma_s/2}(\log n)^{-1/4}, n^{-\gamma_0-\gamma_s+1/2}) = n^{-\gamma_0-\gamma_s+1/2}$, and

$$\begin{aligned}\max(n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n}) &= \max(s^2 n^{-2\gamma_0}\log n, sn^{-1/2}\sqrt{\log n}) \\ &\ll \max(n^{1-4\gamma_0-2\gamma_s}\log n, n^{-\gamma_0-\gamma_s}\sqrt{\log n}) \\ &= n^{-\gamma_0-\gamma_s}\sqrt{\log n} = O(b_s^{-1}d_n).\end{aligned}$$

If $0 \leq \gamma_0+\gamma_s/3 \leq 1/3$, then $\min(n^{\frac{\gamma_0}{2}-\frac{\gamma_s}{2}}(\log n)^{-1/4}, n^{-\gamma_0-\gamma_s+1/2}) = n^{\gamma_0/2-\gamma_s/2}(\log n)^{-1/4}$, and

$$\begin{aligned}\max(n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n}) &= \max(s^2 n^{-2\gamma_0}\log n, sn^{-1/2}\sqrt{\log n}) \\ &\ll \max(n^{-\gamma_0-\gamma_s}\sqrt{\log n}, n^{\gamma_0/2-\gamma_s/2-1/2}(\log n)^{1/4}). \\ &= n^{-\gamma_0-\gamma_s}\sqrt{\log n} = O(b_s^{-1}d_n).\end{aligned}$$

Thus $\max(n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n}) \ll n^{-\gamma_0-\gamma_s}\sqrt{\log n} = O(b_s^{-1}d_n)$

Proof of Proposition 1

For SCAD:

- Given $d_n \gg \eta_p$ and $s \ll \eta_d$, we will show that if $\lambda_n = O(d_n)$ and $d_n \geq a\lambda_n$ (more precisely, $d_n \geq a\lambda_n$ for SCAD_{λ_n} or $d_n \geq a_n\lambda_n$ for $\text{SCAD}_{\lambda_n,a_n}$), conditions 3.1-3.3 and 4.4 are satisfied.

  - Since $d_n \geq a\lambda_n$, $p'_{\texttt{SCAD}_{\lambda_n}}(d_n) = p'_{\texttt{SCAD}_{\lambda_n,a_n}}(d_n) = 0$. Therefore condition 3.1 is satisfied.

  - Because $p'_{\texttt{SCAD}_{\lambda_n}}(0+) = p'_{\texttt{SCAD}_{\lambda_n,a_n}}(0+) = \lambda_n$, condition 3.3 becomes

    $\lambda_n > \dfrac{\sigma^{-1/2}}{(1-K)}n^{-1/2+\alpha/2}\sqrt{\log n}$ and $\lambda_n \gg \max(n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n})$.

    First, $\lambda_n > \frac{\sigma^{-1/2}}{(1-K)}n^{-1/2+\alpha/2}\sqrt{\log n}$ by our choice of $\lambda_n = O(d_n)$, and the assumption

    $d_n \gg n^{-1/2+\alpha/2}\sqrt{\log n}$. Next, $\lambda_n \gg \max(n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n})$ is satisfied by Lemma 1, and the choice of $\lambda_n = O(d_n)$. Therefore condition 3.3 is satisfied.

  - For either $\texttt{SCAD}_{\lambda_n}$ or $\texttt{SCAD}_{\lambda_n,a_n}$, we have $p'_\varpi(0+)/p'_\varpi(d_n) = \infty$ because $p'_\varpi(0+) > 0$ and $p'_\varpi(d_n) = 0$. Therefore condition 3.2 is satisfied.

  - For any $\delta = (\delta_1, ..., \delta_s)^\mathsf{T} \in \mathcal{N}_0 \equiv \{\boldsymbol{\delta} \in \mathbf{R}^s : \|\boldsymbol{\delta} - \beta_{01}\|_\infty \leq d_n\}$, we have $|\delta_j| \geq d_n \geq a\lambda_n$, and thus $\kappa(p_\varpi, \delta_j) = 0$ for either $\texttt{SCAD}_{\lambda_n}$ or $\texttt{SCAD}_{\lambda_n,a_n}$. Therefore condition 4.4 is satisfied.

- Given $d_n \ll O(n^{-1/2+\alpha/2}\sqrt{\log n})$,
  - Condition 3.3 requires $\lambda_n > \frac{\sigma^{-1/2}}{(1-K)} n^{-1/2+\alpha/2}\sqrt{\log n}$. Given $d_n \ll n^{-1/2+\alpha/2}\sqrt{\log n}$, we have $d_n \ll \lambda_n$. Therefore, $p'_{\texttt{SCAD}_{\lambda_n}}(d_n) = p'_{\texttt{SCAD}_{\lambda_n,a_n}}(d_n) = \lambda_n$.
  - Condition 3.1 requires $p'_{\varpi}(d_n) = \lambda_n \ll d_n$ since $b_s^{-1} d_n \ll d_n$.

    Clearly, no such λ_n exists to satisfy $d_n \gg \lambda_n$ and $d_n \ll \lambda_n$ or $d_n < \lambda_n$ simultaneously.

For MCP_{λ_n}:

- Given $d_n \gg \frac{\sigma^{-1/2}}{(1-K)} n^{-1/2+\alpha/2}\sqrt{\log n}$ and $s \ll \min(n^{\gamma_0/2}(\log n)^{-1/4}, n^{-\gamma_0+1/2})$, we will show that if $\lambda_n = O(d_n)$ and $d_n \geq a\lambda_n$, conditions 3.1-3.3 and 4.4 are satisfied.
  - Since $d_n \geq a\lambda_n$, $p'_{\texttt{MCP}_{\lambda_n}}(d_n) = 0$. Therefore condition 3.1 is satisfied.
  - Because $p'_{\texttt{MCP}_{\lambda_n}}(0+) = \lambda_n$, condition 3.3 becomes

    $$\lambda_n > \frac{\sigma^{-1/2}}{(1-K)} n^{-1/2+\alpha/2}\sqrt{\log n} \text{ and } \lambda_n \gg \max(n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n}).$$

    First, $\lambda_n > \frac{\sigma^{-1/2}}{(1-K)} n^{-1/2+\alpha/2}\sqrt{\log n}$ by our choice of λ_n. Next, by Lemma 1, $\lambda_n \gg \max(n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n})$ because $\lambda_n = O(d_n)$. Therefore condition 3.3 is satisfied.
  - For $\texttt{MCP}_{\lambda_n}$, $p'_{\texttt{MCP}}(0+)/p'_{\texttt{MCP}}(d_n) = \infty$ because $p'_{\texttt{MCP}}(0+) > 0$ and $p'_{\texttt{MCP}}(d_n) = 0$. Therefore condition 3.2 is satisfied.
  - Because $d_n \geq a\lambda_n$, $\kappa(p_\varpi, \delta) = 0$ for any $\delta \in \mathcal{N}_0 \equiv \{\boldsymbol{\delta} \in \mathbf{R}^s : \|\boldsymbol{\delta} - \beta_{01}\|_\infty \leq d_n\}$. Thus condition 4.4 is satisfied.

- Given $d_n \ll O(n^{-1/2+\alpha/2}\sqrt{\log n})$,

  - Condition 3.3 requires $\lambda_n > \frac{\sigma^{-1/2}}{(1-K)} n^{-\frac{1}{2}+\frac{\alpha}{2}}\sqrt{\log n}$. Given $d_n \ll n^{-\frac{1}{2}+\frac{\alpha}{2}}\sqrt{\log n}$, it leads $d_n \ll \lambda_n$ or $d_n < \lambda_n$. Therefore, $p'_{\texttt{MCP}_{\lambda_n}}(d_n) = \lambda_n$.
  - Condition 3.1 requires $p'_{\varpi}(d_n) = \lambda_n \ll d_n$ since $b_s^{-1} d_n \ll d_n$.

    Clearly, no such λ_n exists to satisfy both conditions simultaneously.

Proof of Proposition 2

For $\text{MCP}_{\lambda_n, a_n}$, we will show that if (λ_n, a_n) satisfy $\lambda_n > \eta_p$,

$$\lambda_n \gg \max(n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n})$$

, and $a_n\lambda_n < d_n$, conditions 3.1-3.3 and 4.4 are satisfied.

- Since $d_n \geq a_n\lambda_n$, $p'_{\texttt{MCP}_{\lambda_n,a_n}}(d_n) = 0$. Therefore, condition 3.1 is satisfied.
- Because $p'_{\texttt{MCP}_{\lambda_n,a_n}}(0+) = \lambda_n$, condition 3.3 becomes

  $$\lambda_n > \frac{\sigma^{-1/2}}{(1-K)} n^{-1/2+\alpha/2}\sqrt{\log n} \text{ and } \lambda_n \gg \max(n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n}).$$

  By our choice of λ_n, condition 3.3 is satisfied.
- For $\texttt{MCP}_{\lambda_n,a_n}$, $p'_{\texttt{MCP}}(0+)/p'_{\texttt{MCP}}(d_n) = \infty$ because $p'_{\texttt{MCP}}(0+) > 0$ and $p'_{\texttt{MCP}}(d_n) = 0$. Therefore condition 3.2 is satisfied.
- Because $d_n \geq a\lambda_n$, $\kappa(p_\varpi, \delta) = 0$ for any $\delta \in \mathcal{N}_0 \equiv \{\boldsymbol{\delta} \in \mathbf{R}^s : \|\boldsymbol{\delta} - \beta_{01}\|_\infty \leq d_n\}$. Thus condition 4.4 is satisfied.

Proof of Proposition 3

For SICA_{λ_n}:

- $p'_{\texttt{SICA}_{\lambda_n}}(0+) = \lambda_n(1+1/\tau) = O(\lambda_n)$ and $p'_{\texttt{SICA}_{\lambda_n}}(d_n) = \frac{\lambda_n \tau(\tau+1)}{(d_n+\tau)^2} = O(\lambda_n)$. Because $s \ll \min(n^{\gamma_0/2-\gamma_s/2}(\log n)^{-1/4}, n^{-\gamma_0-\gamma_s+1/2})$ and $\alpha < 1-2\gamma_0-2\gamma_s$, we have

$$\max(n^{-1/2+\alpha/2}\sqrt{\log n}, n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n}) \ll n^{-\gamma_s-\gamma_0}\sqrt{\log n}$$

by Lemma 2. In addition, given $\left\|X_2^\mathsf{T}\Sigma(\boldsymbol{\theta}_0)X_1(X_1^\mathsf{T}\Sigma(\boldsymbol{\theta}_0)X_1)^{-1}\right\|_\infty \leq K\,(d_n/\tau+1)^2$, we will show that if

$$\max(n^{-1/2+\alpha/2}\sqrt{\log n}, n^{-2\gamma_0+2\nu}(\log n)^2, n^{\nu-1/2}\sqrt{\log n}) \ll \lambda_n \ll n^{-\gamma_s-\gamma_0}\sqrt{\log n},$$

conditions 3.1-3.3 and 4.4 are satisfied.

  - Since $p'_{\texttt{SICA}_{\lambda_n}}(d_n) = O(\lambda_n) \ll n^{-\gamma_s-\gamma_0}\sqrt{\log n}$, condition 3.1 is satisfied.
  - Because $p'_{\texttt{SICA}_{\lambda_n}}(0+) = O(\lambda_n)$ by the choice of λ_n, condition 3.3 is satisfied by
$$\max(n^{-1/2+\alpha/2}\sqrt{\log n}, n^{-2\gamma_0+2\nu}(\log n)^2, n^{\nu-1/2}\sqrt{\log n}) \ll \lambda_n$$
  - Since $p'_{\texttt{SICA}_{\lambda_n}}(0+)/p'_{\texttt{SICA}_{\lambda_n}}(d_n) = (d_n/\tau+1)^2$, condition 3.2 is satisfied by
$$\left\|X_2^\mathsf{T}\Sigma(\boldsymbol{\theta}_0)X_1(X_1^\mathsf{T}\Sigma(\boldsymbol{\theta}_0)X_1)^{-1}\right\|_\infty \leq K\left(\frac{d_n}{\tau}+1\right)^2.$$
  - Because $p''_{\texttt{SICA}_{\lambda_n}}(d_n) = O(\lambda_n) = o(1)$, condition 4.4 is satisfied.

- Given $d_n \ll O(n^{-1/2+\alpha/2}\sqrt{\log n})$,
  - Condition 3.3 requires $p'_{\texttt{SICA}_{\lambda_n}}(0+) = O(\lambda_n) > \frac{\sigma^{-1/2}}{(1-K)}n^{-\frac{1}{2}+\frac{\alpha}{2}}\sqrt{\log n}$. Given $d_n \ll n^{-\frac{1}{2}+\frac{\alpha}{2}}\sqrt{\log n}$, it leads $d_n \ll \lambda_n$. Therefore, $p'_{\texttt{SICA}_{\lambda_n}}(d_n) = O(\lambda_n)$.

– Condition 3.1 requires $p'_{\varpi}(d_n) = \lambda_n \ll d_n$ since $b_s^{-1} d_n \ll d_n$.

Clearly, no such λ_n exists to satisfy both conditions simultaneously.

For Log_{λ_n}:

- $p'_{\text{Log}_{\lambda_n}}(0+) = \lambda_n/\tau = O(\lambda_n)$ and $p'_{\text{Log}_{\lambda_n}}(d_n) = \lambda_n/(d_n + \tau) = O(\lambda_n)$.

  Given $\alpha < 1 - 2\gamma_0 - 2\gamma_s$ and $s \ll \min(n^{\gamma_0/2 - \gamma_s/2}(\log n)^{-1/4}, n^{-\gamma_0 - \gamma_s + 1/2})$, by Lemma 2, we have

  $$\max(n^{-1/2+\alpha/2}\sqrt{\log n}, n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n}) \ll n^{-\gamma_s-\gamma_0}\sqrt{\log n}.$$

  Given the additional condition $\left\|X_2^{\mathsf{T}}\Sigma(\boldsymbol{\theta}_0)X_1(X_1^{\mathsf{T}}\Sigma(\boldsymbol{\theta}_0)X_1)^{-1}\right\|_\infty \leq K\,(d_n/\tau + 1)$, we will show that if

  $$\max(n^{-1/2+\alpha/2}\sqrt{\log n}, n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n}) \ll \lambda_n \ll n^{-\gamma_s-\gamma_0}\sqrt{\log n},$$

  conditions 3.1-3.3 and 4.4 are satisfied.

  – Since $p'_{\text{Log}_{\lambda_n}}(d_n) = O(\lambda_n) \ll n^{-\gamma_s-\gamma_0}\sqrt{\log n}$ by the choice of λ_n, condition 3.1 is satisfied.

  – Because $p'_{\text{Log}_{\lambda_n}}(0+) = O(\lambda_n)$, by the choice of λ_n, condition 3.3 is satisfied by

  $$\max(n^{-1/2+\alpha/2}\sqrt{\log n}, n^{-2\gamma_0+2\nu}\log n, n^{\nu-1/2}\sqrt{\log n}) \ll \lambda_n.$$

  – Since $p'_{\text{Log}_{\lambda_n}}(0+)/p'_{\text{Log}_{\lambda_n}}(d_n) = d_n/\tau + 1$, condition 3.2 is satisfied given

  $$\left\|X_2^{\mathsf{T}}\Sigma(\boldsymbol{\theta}_0)X_1(X_1^{\mathsf{T}}\Sigma(\boldsymbol{\theta}_0)X_1)^{-1}\right\|_\infty \leq K\left(\frac{d_n}{\tau} + 1\right).$$

  – Because $p''_{\text{Log}_{\lambda_n}}(d_n) = O(\lambda_n) = o(1)$, condition 4.4 is satisfied.

- Given $d_n \ll O(n^{-1/2+\alpha/2}\sqrt{\log n})$,

– Condition 3.3 requires $p'_{\texttt{Log}_{\lambda_n}}(0+) = O(\lambda_n) > \frac{\sigma^{-1/2}}{(1-K)} n^{-\frac{1}{2}+\frac{\alpha}{2}} \sqrt{\log n}$. Given $d_n \ll n^{-\frac{1}{2}+\frac{\alpha}{2}} \sqrt{\log n}$, it leads $d_n \ll \lambda_n$. Therefore, $p'_{\texttt{Log}_{\lambda_n}}(d_n) = O(\lambda_n)$.

– Condition 3.1 requires $p'_{\varpi}(d_n) = \lambda_n \ll d_n$ since $b_s^{-1} d_n \ll d_n$.

Clearly, no such λ_n exists to satisfy both conditions simultaneously.

Proof of Proposition 4

For $\text{SICA}_{\lambda_n, a_n}$:

- Let $\lambda_n = O(n^{\gamma_\lambda})$ and $\tau_n = O(n^{\gamma_\tau})$. Given $0 < \alpha < 1$ and $\nu \leq \gamma_0$, we will show that if $\gamma_\tau < -2\gamma_0 - \gamma_s < \gamma_\lambda < -\gamma_0$, conditions 3.1-3.3 and 4.4 are satisfied.

  – Given $\gamma_\tau < -\gamma_0$, $\exists$ constant C such that $d_n + \tau_n \geq C^{-1} n^{-\gamma_0}$. Therefore

  $$\begin{aligned} p'_{\texttt{SICA}_{\lambda_n, a_n}}(d_n) &= \frac{\lambda_n \tau_n (\tau_n + 1)}{(d_n + \tau_n)^2} \leq C^2 n^{2\gamma_0} \lambda_n \tau_n (\tau_n + 1) = O(n^{2\gamma_0 + \gamma_\lambda + \gamma_\tau}), \\ \kappa(p_\varpi, \delta) &= |p''_{\texttt{SICA}_{\lambda_n, \tau_n}}(d_n)| = \frac{2\lambda_n \tau_n (\tau_n + 1)}{(d_n + \tau_n)^3} \leq 2C^3 n^{3\gamma_0} \lambda_n \tau_n (\tau_n + 1) \\ &= O(n^{3\gamma_0 + \gamma_\lambda + \gamma_\tau}). \end{aligned}$$

  – $p'_{\texttt{SICA}_{\lambda_n, a_n}}(d_n) = O(n^{2\gamma_0 + \gamma_\lambda + \gamma_\tau}) \ll b_s^{-1} d_n = O(n^{-\gamma_s - \gamma_0} \sqrt{\log n})$ by the choice of λ_n, τ_n with $\gamma_\lambda + \gamma_\tau < -3\gamma_0 - \gamma_s$. Therefore, condition 3.1 is satisfied.

  – Since $p'_{\texttt{SICA}_{\lambda_n, \tau_n}}(0+) = O(n^{\gamma_\lambda - \gamma_\tau})$, condition 3.3 becomes

  $$n^{\gamma_\lambda - \gamma_\tau} \gg \max(n^{-2\gamma_0 + 2\nu} \log n, n^{\nu - 1/2} \sqrt{\log n})$$

  and

  $$n^{\gamma_\lambda - \gamma_\tau} > \frac{\sigma^{-1/2}}{(1-K)} n^{-1/2 + \alpha/2} \sqrt{\log n}.$$

  Given $0 \leq \alpha < 1$ and $\nu < \gamma_0 < 1/2$ (conditions 2.1 and 2.2),

  $$\max\left(\frac{\sigma^{-1/2}}{(1-K)} n^{-1/2+\alpha/2} \sqrt{\log n}, n^{-2\gamma_0 + 2\nu} \log n, n^{\nu - 1/2} \sqrt{\log n} \right) \ll \log n$$

  and $n^{\gamma_\lambda - \gamma_\tau} \gg \log n$ by $\gamma_\lambda - \gamma_\tau > 0$. Thus condition 3.3 is satisfied.

– By conditions 2.1 and 2.2, $0 \leq \nu < \gamma_0 < 1/2$, thus $\gamma_\tau < -2\gamma_0 < -\gamma_0 - \nu/2$. Thus $\frac{p'_{\mathtt{SICA}_{\lambda_n,\tau_n}}(0+)}{p'_{\mathtt{SICA}_{\lambda_n,\tau_n}}(d_n)} = (d_n/\tau_n + 1)^2 = O(n^{-2\gamma_0 - 2\gamma_\tau} \log n) \gg O(n^v)$. Therefore condition 3.2 is satisfied

– Condition 4.4 is fulfilled by $\kappa(p_\varpi, \delta) = |p''_{\mathtt{SICA}_{\lambda_n,\tau_n}}(d_n)| = O(n^{3\gamma_0 + \gamma_\lambda + \gamma_\tau}) = o(1)$ because $\gamma_\tau + \gamma_\lambda < -3\gamma_0$.

For $\mathrm{Log}_{\lambda_n,a_n}$:

- Given $0 < \alpha < 1$ and $\nu \leq \gamma_0$, we will show that if $\gamma_\tau < \gamma_\lambda < -2\gamma_0 - \gamma_s$, conditions 3.1-3.3 and 4.4 are satisfied. Given $\gamma_\tau < -\gamma_0$, $\exists$ constant C such that $d_n + \tau_n \geq C^{-1} n^{-\gamma_0}$. Therefore

$$\begin{aligned} p'_{\mathtt{Log}_{\lambda_n,\tau_n}}(d_n) &= \lambda_n/(d_n + \tau_n) \leq C n^{\gamma_0} \lambda_n = O(n^{\gamma_0 + \gamma_\lambda}) \\ \kappa(p_\varpi, \delta) &= |p''_{\mathtt{Log}_{\lambda_n,\tau_n}}(d_n)| = \lambda_n/(d_n + \tau_n)^2 \leq C^2 n^{2\gamma_0} \lambda_n = O(n^{2\gamma_0 + \gamma_\lambda}). \end{aligned}$$

– $p'_{\mathtt{Log}_{\lambda_n,\tau_n}}(d_n) = O(n^{\gamma_0 + \gamma_\lambda}) \ll b_s^{-1} d_n = O(n^{-\gamma_s - \gamma_0} \sqrt{\log n})$ by the choice of λ_n with $\gamma_\lambda < -2\gamma_0 - \gamma_s$. Therefore, condition 3.1 is satisfied.

– Since $p'_{\mathtt{Log}_{\lambda_n}}(0+) = \lambda_n/\tau_n = O(n^{\gamma_\lambda - \gamma_\tau})$. Condition 3.3 becomes

$$n^{\gamma_\lambda - \gamma_\tau} \gg \max(n^{-2\gamma_0 + 2\nu} \log n, n^{\nu - 1/2} \sqrt{\log n})$$

and

$$n^{\gamma_\lambda - \gamma_\tau} > \frac{\sigma^{-1/2}}{(1-K)} n^{-1/2 + \alpha/2} \sqrt{\log n}.$$

Given $0 \leq \alpha < 1$ and $\nu < \gamma_0 < 1/2$ (conditions 2.1 and 2.2),

$$\max\left(\frac{\sigma^{-1/2}}{(1-K)} n^{-1/2 + \alpha/2} \sqrt{\log n}, n^{-2\gamma_0 + 2\nu} \log n, n^{\nu - 1/2} \sqrt{\log n} \right) \ll \log n.$$

Because $\gamma_\lambda - \gamma_\tau > 0$, $n^{\gamma_\lambda - \gamma_\tau} \gg \log n$. Thus condition 3.3 is satisfied.

– By conditions 2.1 and 2.2, $0 \leq \nu < \gamma_0 < 1/2$, thus $\gamma_\tau < -2\gamma_0 < -\gamma_0 - \nu$. Thus $\frac{p'_{\mathtt{Log}_{\lambda_n,\tau_n}}(0+)}{p'_{\mathtt{Log}_{\lambda_n,\tau_n}}(d_n)} = d_n/\tau_n + 1 = O(n^{-\gamma_0-\gamma_\tau}\sqrt{\log n}) \gg O(n^v)$. Therefore condition 3.2 is satisfied

– Condition 4.4 is fulfilled by $\kappa(p_\varpi, \delta) = |p''_{\mathtt{Log}_{\lambda_n,\tau_n}}(d_n)| = O(n^{2\gamma_0+\gamma_\lambda}) = o(1)$ because $\gamma_\lambda < -2\gamma_0$.

Proof of Corollary 1

Given $d_n \ll \eta_p$:

- For $\mathrm{MCP}_{\lambda_n,a_n}$, :
  - Condition 3.3 requires $\lambda_n > \eta_p$.
  - If a_n is tuned such that $d_n < a_n\lambda_n$, then $p'_\varpi(d_n) = \lambda_n + d_n/a_n$. Ccondition 3.3 requires $\lambda_n > \eta_p$ so that condition 3.1 cannot be satisfied due to $p'_\varpi(d_n) > \eta_p \gg d_n$.
  - If a_n is tuned such that $d_n \geq a_n\lambda_n$, then $p'_\varpi(d_n) = 0$. Therefore, condition 3.1 can be satisfied. This restricts the valid range of a_n: $a_n < d_n/\lambda_n < d_n/\eta_p = o(1)$.
- For $\mathrm{SICA}_{\lambda_n,\tau_n}$:
  - Condition 3.3 requires $p'_\varpi(0+) > \eta_p$.
  - Note that $p'_\varpi(d_n)$ of SICA can be expressed as $p'_\varpi(d_n) = p'_\varpi(0+)/(d_n/\tau_n + 1)^2$. Condition 3.1 requires $p'_\varpi(d_n) = p'_\varpi(0+)/(d_n/\tau_n+1)^2 \ll d_n$. Combin-

ing condition with 3.3, it leads that $\eta_p/(d_n/\tau_n+1)^2 \ll p'_\varpi(0+)/(d_n/\tau_n+1)^2 \ll d_n$. Note that condition 3.2 requires $d_n/\tau_n \to \infty$ as shown in proof of proposition 4 so that $O(d_n/\tau_n+1) = O(d_n/\tau_n)$. Therefore, the valid range of τ_n is restricted as $\tau_n < d_n^{3/2}/\eta_p^{1/2} = o(1)$.

- For $\text{Log}_{\lambda_n,\tau_n}$:
  - Condition 3.3 requires $p'_\varpi(0+) > \eta_p$.
  - Note that $p'_\varpi(d_n)$ of Log can be expressed as $p'_\varpi(d_n) = p'_\varpi(0+)/(d_n/\tau_n+1)$. Condition 3.1 requires $p'_\varpi(d_n) = p'_\varpi(0+)/(d_n/\tau_n+1) \ll d_n$. Combined with condition 3.3, it leads that $\eta_p/(d_n/\tau_n+1) \ll p'_\varpi(0+)/(d_n/\tau_n+1) \ll d_n$. Note that condition 3.2 requires $d_n/\tau_n \to \infty$ as shown in proof of proposition 4 so that $O(d_n/\tau_n+1) = O(d_n/\tau_n)$. Therefore, the valid range of τ_n is restricted as $\tau_n < d_n^2/\eta_p = o(1)$.

CHAPTER 3: Prediction of cancer drug sensitivity

3.1 Introduction

Human cancer arises from an accumulation of somatic mutations during the lifetime of a patient. Recent studies have shown that cancer growth is often driven by a few somatic mutations (so-called driver mutations), which may be buried among a large number of "passenger" mutations [Hanahan and Weinberg, 2011]. Interventions targeting these driver mutations or relevant pathways have proved to be effective treatment options. However, not all the patients with the targeted somatic lesions respond to the therapy. Take the targeted breast cancer treatment on the oncogene HER2 as an example. The HER2 gene encodes a protein product that promotes the growth of cancer cells. The amplification of the HER2 gene in breast cancer increases the aggressiveness of the tumor. A drug, Trastuzumab, has been developed to target HER2 amplification. However, among those breast cancer patients with HER2 over-expression, only 30% respond to Trastuzumab therapy [De Palma and Hanahan, 2012]. It is believed that genome-wide genetic heterogeneity among cancer patients is one of the main reasons for the diverse treatment responses. In other words, patients with HER2 amplification may have very different genomic features (e.g., DNA alterations, gene expression, and epigenetic marks) in their genomes, and these differences may lead to diverse treatment responses.

Preclinical model systems such as cancer cell lines that reflect the genomic diversity of human cancers can be used to identify predictive genomic features/biomarkers for drug sensitivity [Caponigro and Sellers, 2011]. Recently, two groups (Garnett et al.

[2012] and Barretina et al. [2012]) have studied drug sensitivity in a large number of cancer cell lines and measured several types of genomic features including mutations of cancer genes, genome-wide copy number alterations, and gene expression. The sample size ranges from 200 to 500 per drug, while the number of genomic features is greater than 10,000. The authors conducted drug-by-drug analysis to identify associated genomic features, and they demonstrated that these cell line systems can capture expected molecular targets of cancer drugs and provide novel findings on the genomic basis of drug sensitivity.

It is expected that drugs with the same targets may have some common genomic features in addition to their individual features. The results of the aforementioned studies [Garnett et al., 2012; Barretina et al., 2012] support this speculation. Therefore, a joint analysis of the drugs sharing a target may improve the sensitivity and specificity with which we can identify their shared genomic features. To this end, we consider the feature selection method for multivariate responses to identify predictive genomic features of drugs with the same target.

Two types of methods have been developed for feature selection for multivariate responses: group-wise selection and bi-level selection. Group-wise variable selection methods, such as group Lasso [Yuan and Lin, 2006] or group adaptive Lasso [Wang and Leng, 2008], assume that all the response variables within a group are associated with the same set of covariates [Huang et al., 2012]. This assumption is not reasonable for cancer drug-sensitivity studies. For example, Garnett et al. [2012] and Barretina et al. [2012] have shown that drugs with the same target may have some shared genomic features, but they also have individual features. In contrast, bi-level selection methods encourage the selection of covariates associated with all

the response variables, but they also allow some covariates to be associated with one or a few response variables [Breheny and Huang, 2009]. These flexibilities in bi-level selection methods are desirable for cancer drug-sensitivity applications. A few methods have been developed for bi-level selection, such as group bridge [Huang et al., 2009] and composite MCP [Breheny and Huang, 2009]. Although these methods work satisfactorily in many real-data analyses, we find that their performance is limited in some genomic applications where the genomic features have strong correlations. These issues motivate us to develop a new method to construct predictive models of cancer drug sensitivity using genomic features.

In this paper, we propose a new bi-level selection method called BipLog. Simulation studies show that it has substantially higher sensitivity and specificity than existing methods. We apply BipLog to identify the genomic features associated with drug sensitivity for two sets of real data [Garnett et al., 2012; Barretina et al., 2012]. We seek to answer a few important questions in our data analysis. First, by splitting the data from Garnett et al. [2012] into training and testing sets, we assess the variation in the drug sensitivity that can be explained by our predictive model. Second, we use all the data from Garnett et al. [2012] to select genomic features associated with each drug target, and we evaluate their prediction performance using independent data from Barretina et al. [2012]. There are substantial differences in these two studies in terms of the drugs studied and the method to estimate the drug sensitivity. Therefore, this between-study comparison helps to evaluate the robustness and generality of our method. Third, we use this between-study comparison to compare the results of BipLog with those of the "drug-by-drug" analysis using the elastic net [Zou and Hastie, 2005].

The remainder of this paper is organized as follows. We introduce BipLog and its implementation in Section 2. We present the simulation studies and real-data analyses in Sections 3 and 4. Section 5 provides concluding remarks.

3.2 Method

3.2.1 Objective function

Suppose in a group of n samples, we observe q response variables, denoted $y_k = (y_{1k}, ..., y_{nk})^T$ $(1 \leq k \leq q)$, and p covariates, denoted $\mathbf{x}_j = (x_{1j}, ..., x_{nj})^T$ $(1 \leq j \leq p)$. We assume that q is much smaller than the sample size n, but p is often larger or much larger than n. After standardizing y_k and $\mathbf{x}_j$ to have mean 0 and $\|y_k\|_2^2 = \|\mathbf{x}_j\|_2^2 = 1$, we assume a linear system: $E(y_k) = X\beta_k = \sum_{j=1}^{p} \mathbf{x}_j \beta_{jk}$, where $X_{n\times p} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ and $\beta_k = (\beta_{1k}, ..., \beta_{pk})^T$. Let $\beta = (\beta_1, ..., \beta_q)$, and denote each row of β by $\mathbf{b}_j = (\beta_{j1}, ..., \beta_{jq})$. Let $|\mathbf{b}_j| = \sum_{k=1}^{q} |\beta_{jk}|$. The objective function that we aim to minimize is a penalized least squares:

$$Q(\beta) \;=\; \frac{1}{2n} \sum_{k=1}^{q} \|y_k - X\beta_k\|_2^2 + \sum_{j=1}^{p} \sum_{k=1}^{q} p_{\boldsymbol{\theta}_1}(|\beta_{jk}|) + \sum_{j=1}^{p} p_{\boldsymbol{\theta}_2}(|\mathbf{b}_j|), \quad (3.2.1)$$

where $p_{\boldsymbol{\theta}_1}(|\beta_{jk}|) = \lambda_1 \log(|\beta_{jk}| + \tau_1)$, $p_{\boldsymbol{\theta}_2}(|\mathbf{b}_j|) = \lambda_2 \log(|\mathbf{b}_j| + \tau_2)$, $\boldsymbol{\theta}_1 = (\lambda_1, \tau_1)$, and $\boldsymbol{\theta}_2 = (\lambda_2, \tau_2)$.

In its general form, $p_{\varpi}(\beta) = \lambda \log(|\beta| + \tau)$ is the Log penalty for a parameter β with tuning parameters $\varpi = (\lambda, \tau)$. The Log penalty is a nonconvex penalty, or more precisely a folded concave penalty [Fan and Lv, 2010] in the sense that it is concave for $\beta \in [0, \infty)$, with continuous derivative $p'_{\varpi}(\beta) \geq 0$, and $p'_{\varpi}(0+) > 0$. Friedman [2008] originally proposed the Log penalty in an alternative form: $\lambda \log[(1-r)|\beta| + r]$, with $0 < r < 1$. Friedman [2008] observed that the Log penalty bridges the L_1 penalty

(Lasso) and the L_0 penalty (all-subset selection) as r changes from 1 to 0. To illustrate the characteristics of the Log penalty, we compare it with one of the most commonly used penalties, the Lasso penalty (Figure 1). Lasso gives biased penalized estimation since the penalty increases linearly as the regression coefficients increase. The Log penalty mitigates this issue by reducing the penalty increase rate for larger regression coefficients. In a previous study of penalized estimation with a univariate response variable, we have shown that the Log penalty has better performance than Lasso in some genomic applications [Sun et al., 2010].

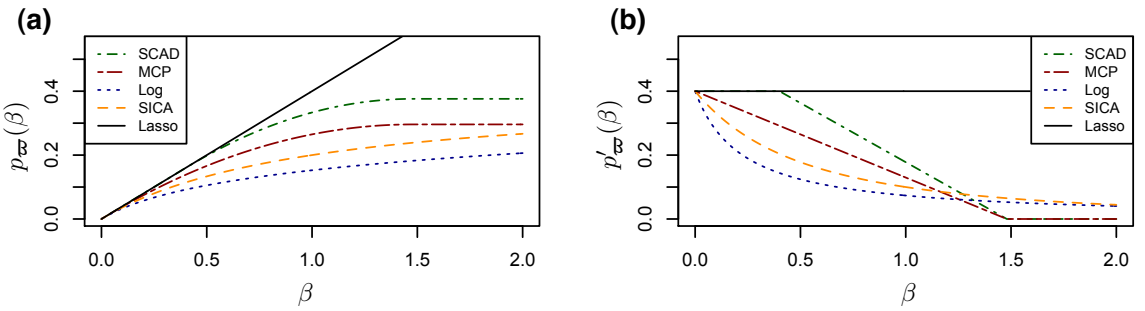

Figure 3.1: (a) The Log and Lasso penalty functions. For the Log penalty, $\lambda = 0.09, \tau = 0.225$ and for Lasso, $\lambda = 0.4$. These two sets of tuning parameters are comparable in the sense that they provide the same penalty derivative at 0+. (b) The derivatives of these two penalty functions for the tuning parameters of Figure 1(a).

We achieved bi-level selection by applying Log penalties to each coefficient and each group of coefficients (i.e., the coefficients of the same covariate across all responses) through $\sum_{j=1}^{p}\sum_{k=1}^{q} p_{\boldsymbol{\theta}_1}(|\beta_{jk}|)$ and $\sum_{j=1}^{p} p_{\boldsymbol{\theta}_2}(|\mathbf{b}_j|)$, respectively. Given that the observations of the responses and covariates have been standardized, the magnitudes of the regression coefficients are comparable across different responses and covariates. Here we choose to use a group penalty of the form $p_{\boldsymbol{\theta}_2}(|\mathbf{b}_j|) = \lambda_2 \log(|\mathbf{b}_j| + \tau_2) = \lambda_2 \log(\sum_{k=1}^{q} |\beta_{jk}| + \tau_2)$. In the following section, we will explain why we use this group penalty and why an alternative form of the L_2 penalty (i.e., $\lambda_2 \log(\|\mathbf{b}_j\|_2 + \tau_2)$

where $\|\mathbf{b}_j\|_2 \equiv (\sum_{k=1}^{q} \beta_{jk}^2)^{1/2})$ does not lead to desirable penalization.

3.2.2 Computation

We estimate the β that minimizes $Q(\beta)$ in Equation (3.2.1) using a combination of local linear approximation (LLA) [Zou and Li, 2008] and a coordinate descent algorithm. Specifically, given initial values of β, or the estimates from the tth iteration, denoted $\{\hat{\beta}_j^{(k)}\}$, we apply LLA to the Log penalty functions $p_{\boldsymbol{\theta}_1}(|\beta_{jk}|)$ and $p_{\boldsymbol{\theta}_2}(|\mathbf{b}_j|)$ to update them at the $(t+1)$th iteration:

$$\begin{aligned}
p_{\boldsymbol{\theta}_1}(|\beta_{jk}|) &\approx p_{\boldsymbol{\theta}_1}\left(|\hat{\beta}_j^{(k)}|\right) + \left.\frac{\partial p_{\boldsymbol{\theta}_1}(|\beta_{jk}|)}{\partial|\beta_{jk}|}\right|_{|\beta_{jk}|=|\hat{\beta}_j^{(k)}|}\left(|\beta_{jk}| - |\hat{\beta}_j^{(k)}|\right) = \frac{\lambda_1|\beta_{jk}|}{|\hat{\beta}_j^{(k)}| + \tau_1} + C_1, \\
p_{\boldsymbol{\theta}_2}(|\mathbf{b}_j|) &\approx p_{\boldsymbol{\theta}_2}\left(|\hat{\mathbf{b}}_j^{(t)}|\right) + \sum_{k=1}^{q} \left.\frac{\partial p_{\boldsymbol{\theta}_2}(|\mathbf{b}_j|)}{\partial|\beta_{jk}|}\right|_{|\beta_{jk}|=|\hat{\beta}_j^{(k)}|}\left(|\beta_{jk}| - |\hat{\beta}_j^{(k)}|\right) \\
&= \sum_{k=1}^{q} \frac{\lambda_2|\beta_{jk}|}{|\hat{\mathbf{b}}_j^{(t)}| + \tau_2} + C_2,
\end{aligned}$$

where C_1 and C_2 are constants with respect to β_{jk}. Then the objective function at the $(t+1)$th iteration, denoted $\tilde{Q}^{(t+1)}(\beta)$, can be written

$$\tilde{Q}^{(t+1)}(\beta) = \frac{1}{2n}\sum_{k=1}^{q}\|y_k - X\beta_k\|_2^2 + \sum_{j=1}^{p}\sum_{k=1}^{q}\frac{\lambda_1|\beta_{jk}|}{|\hat{\beta}_j^{(k)}| + \tau_1} + \sum_{j=1}^{p}\sum_{k=1}^{q}\frac{\lambda_2|\beta_{jk}|}{|\hat{\mathbf{b}}_j^{(t)}| + \tau_2}. \quad (3.2.2)$$

$\tilde{Q}^{(t+1)}(\beta)$ should be understood as a working objective function, which is a function of the regression coefficients of interest together with estimates of these coefficients at previous iteration. Therefore, it is different from the objective function $Q(\beta)$ specified in Equation (3.2.1). We use a coordinate descent approach to find each regression coefficient β_{jk} sequentially. To solve for β_{jk}, we minimize the following objective function

$$\tilde{Q}(\beta_{jk}) = \frac{1}{2}\left(\beta_{jk} - \bar{\beta}_{jk}^{(t)}\right)^2 + \left\{\frac{\lambda_1}{|\hat{\beta}_j^{(k)}| + \tau_1} + \frac{\lambda_2}{\sum_{k=1}^{q}|\hat{\beta}_j^{(k)}| + \tau_2}\right\}|\beta_{jk}|, \quad (3.2.3)$$

where $\bar{\beta}_j^{(k)} = (1/n)\sum_{i=1}^{n} x_{ij}\left(y_{ik} - \sum_{l\neq j} x_{il}\hat{\beta}_{lk}^{(t)}\right)$. In summary, this "LLA + coordinate descent" algorithm alternates through different iterations indexed by t, and within each iteration, it estimates all the regression coefficients sequentially. Finally, this algorithm is considered to have converged if the maximum difference in the coefficient estimates between consecutive iterations is less than a predefined threshold, say 10^{-4}.

The penalty term for each step of the coordinate descent algorithm (Equation (3.2.3)) can be written as an adaptive Lasso form of $\lambda_1 \hat{w}_{jk}|\beta_{jk}|$ where the weight is

$$\hat{w}_{jk} = \left[\left(|\hat{\beta}_j^{(k)}| + \tau_1\right)^{-1} + (\lambda_2/\lambda_1)\left(\sum_{k=1}^{q} |\hat{\beta}_j^{(k)}| + \tau_2\right)^{-1}\right]. \tag{3.2.4}$$

Therefore, the above computational algorithm is reminiscent of adaptive Lasso [Zou, 2006] rather than adaptive group Lasso [Wang and Leng, 2008], which uses the L_2 norm of β as a group-level penalty. In contrast to the adaptive Lasso, which adapts a weight function $1/|\hat{\beta}_j^{(k)}|$, our weight function in Equation (3.2.4) is a weighted sum of the contributions of the individual coefficient estimates $(|\hat{\beta}_j^{(k)}| + \tau_1)^{-1}$ and the group-level estimates $(\sum_{k=1}^{q} |\hat{\beta}_j^{(k)}| + \tau_2)^{-1}$, with weights 1 and λ_2/λ_1, respectively. Another important difference between our penalty and the adaptive Lasso penalty is the inclusion of the tuning parameters τ_1 and τ_2, which prevent an infinite penalty for any regression coefficient with a previous estimate of 0. This is necessary for the iterative estimation procedure to proceed with one or more regression coefficients penalized to 0. The sizes of τ_1 and τ_2 can be adjusted to apply penalties of the appropriate strength.

In summary, the penalty terms used in the intermediate steps of our algorithm (namely in Equations (3.2.2) and (3.2.3)) are different from those of adaptive Lasso

and adaptive group Lasso. More importantly, they are part of the intermediate objective function that is updated at each iteration. The ultimate objective function is that in Equation (3.2.1), with bi-level Log penalties.

3.2.3 A Bayesian interpretation of BipLog

The following Bayesian interpretation provides additional insight into our method and the role of the tuning parameters. Recall that $\mathbf{b}_j = (\beta_{j1}, ..., \beta_{jq})^T$ are the regression coefficients for the jth covariate across the q response variables. Our BipLog penalty can be derived from a Bayesian setup using the following priors:

$$\begin{aligned} p(\mathbf{b}_j|\omega_{j1}, ..., \omega_{jq}, \omega_j) &= \left\{\prod_{k=1}^{q} \frac{1}{2}(\omega_{jk}^{-1} + \omega_j^{-1}) \exp\left(-\frac{|\beta_{jk}|}{\omega_{jk}}\right)\right\} \exp\left(-\frac{\sum_{k=1}^{q}|\beta_{jk}|}{\omega_j}\right), \\ p(\omega_{jk}|\delta_1, \tau_1) &= \text{inv-Gamma}(\omega_{jk}; \delta_1, \tau_1) = \frac{\tau_1^{\delta_1}}{\Gamma(\delta_1)}\omega_{jk}^{-1-\delta_1} \exp\left(-\frac{\tau_1}{\omega_{jk}}\right), \\ p(\omega_j|\delta_2, \tau_2) &= \text{inv-Gamma}(\omega_j; \delta_2, \tau_2) = \frac{\tau_2^{\delta_2}}{\Gamma(\delta_2)}\omega_j^{-1-\delta_2} \exp\left(-\frac{\tau_2}{\omega_j}\right), \end{aligned}$$

where $\delta_1 > 0$, $\delta_2 > 0$, $\tau_1 > 0$, and $\tau_2 > 0$ are four hyperparameters. Given the above specification, after integrating out ω_{jk} and ω_j, we obtain the density of $\mathbf{b}_j$:

$$f(\mathbf{b}_j|\delta_1, \tau_1, \delta_2, \tau_2) \propto \frac{\tau_2^{\delta_2}\delta_2}{2(\sum_{i=1}^{q}|\beta_{jk}| + \tau_2)^{1+\delta_2}} \prod_{i=1}^{q} \frac{\tau_1^{\delta_1}\delta_1}{2(|\beta_{jk}| + \tau_1)^{1+\delta_1}}. \tag{3.2.5}$$

This Bayesian interpretation illustrates the similarities and differences of our method and adaptive Lasso. The priors for $\mathbf{b}_j$ include the Laplace prior for each regression coefficient β_{jk} and the Laplace prior for the L_1 norm of $\mathbf{b}_j$. The Laplace prior corresponds to the Lasso penalty [Tibshirani, 1996; Park and Casella, 2008]. The fact that we assign different parameters ω_{jk} and ω_j for the Laplace priors of the regression coefficients β_{jk} and $|\mathbf{b}_j|$ implies that these priors correspond to adaptive Lasso penalties. In the high-dimensional setting where the number of covariates p

is much larger than n, one cannot obtain a good initial estimate of each regression coefficient to decide the prior distribution of β_{jk}. Therefore, we further assign prior distributions for ω_{jk} and ω_j, and after integrating out ω_{jk} and ω_j, we obtain the density of $\mathbf{b}_j$ (Equation (3.2.5)) in terms of the hyperparameters $\delta_1 > 0$, $\delta_2 > 0$, $\tau_1 > 0$, and $\tau_2 > 0$. This density of $\mathbf{b}_j$ is connected to the Log penalty. In fact, $-\log\{f(\mathbf{b}_j|\delta_1, \tau_1, \delta_2, \tau_2)\}$ gives exactly the same form of the BipLog penalty as in Equation (3.2.1) if we set $n\lambda_1 = 1 + \delta_1$ and $n\lambda_2 = 1 + \delta_2$. This also gives more insight into the scale of the tuning parameters of λ_1 and λ_2. Empirically, the grids of possible values of λ_1 and λ_2 could be set at the scale of n^{-1} since both δ_1 and δ_2 are constant $O(1)$.

3.2.4 A penalized maximum likelihood estimation perspective

A generalized form of PMLE is:

$$n^{-1}l_n(\beta) - \rho(\beta),$$

where $l_n(\beta)$ is the log-likelihood function and $\rho(.)$ is a general form of the penalty function. The goal of PMLE is to select the important variables for which the penalized coefficient estimates are nonzero [Fan and Lv, 2010].

If we assume that the q response variables follow a multivariate Gaussian distribution, the penalized least squares estimation problem addressed in this paper is closely related to the PMLE. The difference is that in the PMLE estimation, we also need to estimate the inverse covariance matrix of the q response variables given the covariates, and the strength of the penalization is related to not only the tuning parameters but also the residual variance. A smaller residual variance leads to a smaller penalization.

This is reasonable since a smaller residual variance means a better model fitting, and thus it deserves less penalization. However, in a high-dimensional setting, where p is larger or much larger than n, the relationship between the residual variance and the penalization strength may introduce instability into the iterative updating algorithm. During the iterations, if several covariates are mistakenly included in the model because of their spurious correlations with the residual errors, the residual variance will become smaller, which further attenuates the penalty strength. This then encourages the selection of more covariates and may lead to overfitting problems. We have observed this type of overfitting for PMLE in our preliminary work. Therefore, although PMLE is expected to be more efficient than penalized least squares estimation, we choose to use the latter since it is more robust.

3.2.5 Tuning parameter selection

We choose the best set of tuning parameters by a grid search over an initial pool of parameters. Based on the Bayesian interpretation of BipLog, we set $\lambda_1 = (1+\delta_1)/n$ and $\lambda_2 = (1+\delta_2)/n$. The initial values for δ_1 and δ_2 range from 0 to 5.0 with a 0.5 increment. Let $\hat{\beta}_{ji}^{mls}$ denote the estimates of the marginal regression coefficients. The initial values for τ_1 and τ_2 are from 1^{-3} to $\max_{j,k}\{|\hat{\beta}_{jk}^{mls}|\}$ and from 1^{-3} to $\max_j\{\sum_{k=1}^{q}|\hat{\beta}_{jk}^{mls}|\}$, respectively, in predefined numbers. We select a combination of tuning parameters using the extended BIC [Chen and Chen, 2008]. The extended BIC for a model m is:

$$\texttt{BIC}_{\varrho}(m) = -2\log l_n\{\hat{\boldsymbol{\theta}}(m)\} + \kappa_m \log n + 2\varrho \log \varsigma(S_{\kappa_m}),$$

where $l_n\{\hat{\boldsymbol{\theta}}(m)\}$ is the log likelihood, $\hat{\boldsymbol{\theta}}(m)$ are estimates of all the parameters, κ_m is the degree of freedom for model m, and $\varsigma(S_{\kappa_m})$ is the number of models with degree of freedom equal to κ_m. Specifically, $l_n\{\hat{\boldsymbol{\theta}}(m)\}$ is calculated using the penalized coefficient estimates assuming a multivariate Gaussian distribution. Given

the coefficient estimates, we can calculate the determinant of the residual covariance matrix, denoted $|\hat{\Sigma}|$, and then the log likelihood is simply $-(1/2)\log|\hat{\Sigma}|$. Note that the calculation of $|\hat{\Sigma}|$ is straightforward because of our assumption that the number of response variables is much smaller than the sample size n. We set the number of nonzero coefficient estimates to κ_m and $\varsigma(S_{\kappa_m}) = \binom{pq}{\kappa_m}$, i.e., the number of choices of κ_m coefficients from a total of pq regression coefficients. In addition, following Chen and Chen [2008], we set $\varrho \approx 1 - 1/[2\log(pq)/\log n]$.

3.3 Simulation Studies

3.3.1 Simulation setup

We used simulated data to evaluate our method and two existing methods for bi-level variable selection. The major challenges of feature selection using genomic data are the high dimensionality and the correlations among the genomic features. It is difficult to simulate high-dimensional genomic data with a realistic correlation structure except in a few special cases. One such case is the correlation structure among SNPs (single nucleotide polymorphisms) in an F2 cross. The R package QTL provides a set of utility functions for such simulations [Broman et al., 2003]. Using the function `sim.map` in R/QTL, we first simulated a genetic marker map of 2,000 SNPs from 20 chromosomes of length 90 cM, with 100 SNPs per chromosome. Then we used the function `sim.cross` in R/QTL to simulate the genotype data of an F2 cross with sample size $n = 200$ based on the simulated marker map. As expected, the simulated genotypes show strong correlations for nearby SNPs (average R^2 is 0.96 for SNPs within 1 cM) and no correlation for SNPs from different chromosomes. We randomly selected $p = 600$ SNPs from the 2,000 SNPs for the following simulation of quantitative traits.

We simulated a total of $q = 30$ quantitative traits from the multivariate linear model

$$\mathbf{Y}_{n\times q} = X_{n\times p}\beta_{p\times q} + \mathcal{E}_{n\times q}, \tag{3.3.1}$$

where $\mathbf{Y} = (y_1, ..., y_q)$. The residuals $\mathcal{E}$ were simulated from a multivariate Gaussian distribution with mean 0 and compound symmetry covariance structure with diagonal variance $(0.25+0.5)$ and off-diagonal covariance 0.5. Traits 1 to 10 share a pair of causal SNPs, and each has its own causal SNP. Traits 11 to 30 do not have individual causal SNPs. Traits 11 to 20 share two pairs of causal SNPs, and traits 21 to 30 share one pair of causal SNPs. The pairs of causal SNPs shared across traits may be located in different chromosomes (unlinked) or at the same chromosome with the effect sizes being (η, η) (SNPs linked in coupling) or (η, -η) (SNPs linked in repulsion). We set the genotype effect size $\eta = 0.3$ or 0.6. Given the three relationships between the causal SNP pairs and the two effect sizes, there are six simulation scenarios in total.

We compared BipLog with group bridge and composite MCP [Huang et al., 2012]. We used the implementation of group bridge and composite MCP in R/GRPREG, with the default choice of 100 possible values of the tuning parameter λ, which were uniformly distributed on a log scale. We used an oracle criterion to select the tuning parameters to avoid confounding the feature-selection performance with the criterion for the tuning-parameter selection. Specifically, let s be the number of causal SNPs, and let `D` be the number of discoveries, i.e., the number of nonzero regression coefficient estimates. `D` = `TD` + `FD`, where `TD` and `FD` are the number of true and false discoveries. Under nonnull simulations, the oracle criterion evaluates a model based on three measures: the false discovery rate `FD/D`, the power `TD`/s, and the sum of the squared errors of the regression coefficient estimates $\sum_{j=1}^{p}\sum_{k=1}^{q}|\hat{\beta}_{jk} - \tilde{\beta}_{jk}|^2$,

where $\tilde{\beta}_{jk}$ is the true value of β_{jk}. We select the model with the smallest value of $\texttt{wt}(\texttt{FD/D}-\texttt{TD}/s)+\sum_{j=1}^{p}\sum_{k=1}^{q}|\hat{\beta}_{jk}-\tilde{\beta}_{jk}|^2$, where $\texttt{wt}$ is a weight to balance the number of true/false discoveries and the bias. We set $\texttt{wt}$ to 10 so that we select the models mainly based on $(\texttt{FD/D}-\texttt{TD}/s)$, and the sum of squared errors is a secondary criterion. We also set $\texttt{wt} = 1$ or 0.1 for comparison purposes, and the conclusions are consistent with the results for $\texttt{wt} = 10$ (results not shown).

Table 3.1 summarizes the empirical performance of the three bi-level selection methods: BipLog, group bridge (gBridge), and composite MCP (cMCP). When the shared SNPs are unlinked or linked in coupling, the three methods have a comparable number of true discoveries while BipLog has many fewer false discoveries. When the shared SNPs are linked in repulsion and the effect size is relatively small ($\eta = 0.3$), gBridge and cMCP fail to identify most of the true discoveries while BipLog finds almost 50% of them. In general, BipLog can identify true signals with a smaller bias in the coefficient estimates.

3.4 Genomic signatures of cancer drug sensitivity

For a panel of 639 human cancer cell lines, Garnett et al. [2012] measured the mutation statuses of 64 commonly mutated cancer genes (exon-sequencing), the genome-wide copy number alterations (Affymetrix SNP array 6.0), and the gene expression (Affymetrix HT-U133A microarray). A total of 130 drugs, including those for targeted cancer therapy or broad-spectrum chemotherapy, were selected for the analysis. Each drug was studied in a range of 275 to 505 cell lines. For a panel of 947 human cell lines, Barretina et al. [2012] measured the mutation statuses of 1600 genes (targeted-sequencing), the genome-wide copy number alterations (Affymetrix SNP array 6.0), and the gene expression (Affymetrix U133 plus 2.0 array). Twenty-four

Table 3.1: Comparisons of three bi-level selection methods (group bridge (gBridge), composite MCP (cMCP), and BipLog) via simulation studies. For each of the 6 simulation scenarios, 30 traits are considered. The total number of true trait-SNP associations is 90, which includes 10 associations due to SNPs affecting only one trait (**individual SNPs**) and 80 associations due to SNP pairs shared across traits (**shared SNPs**). The tuning parameters are selected to minimize $10(\texttt{FD/D} - \texttt{TD}/s) + \sum_{j=1}^{p}\sum_{k=1}^{q}|\hat{\beta}_{jk} - \beta_{jk}|^2$. We present the median number of true discoveries (**TD**) and false discoveries (**FD**) and the average bias of the regression coefficient estimates (in brackets []) for the true signals over 100 simulations.

		gBridge	cMCP	BipLog
Shared SNPs are unlinked				
$\eta = 0.3$	individual-SNPs:TD [bias]	9 [0.14]	9 [0.20]	7 [0.10]
	shared-SNPs:TD [bias]	72 [0.16]	68 [0.20]	80 [0.076]
	total FD	172	39	3
$\eta = 0.6$	individual-SNPs:TD [bias]	10 [0.26]	10 [0.29]	10 [0.071]
	shared-SNPs:TD [bias]	78 [0.29]	80 [0.30]	80 [0.059]
	total FD	117	11	0
Shared SNPs are linked in coupling				
$\eta = 0.3$	individual-SNPs:TD [bias]	1 [0.15]	8 [0.21]	7 [0.10]
	shared-SNPs:TD [bias]	72 [0.15]	69 [0.14]	77 [0.082]
	total FD	60	17	4
$\eta = 0.6$	individual-SNPs:TD [bias]	1 [0.30]	10 [0.31]	10 [0.086]
	shared-SNPs:TD [bias]	77 [0.27]	80 [0.18]	80 [0.075]
	total FD	44	3	1
Shared SNPs are linked in repulsion				
$\eta = 0.3$	individual-SNPs:TD [bias]	2 [0.14]	6 [0.24]	7 [0.061]
	shared-SNPs:TD [bias]	0 [0.13]	0 [0.28]	36 [0.12]
	total FD	1	1	7
$\eta = 0.6$	individual-SNPs:TD [bias]	10 [0.21]	10 [0.19]	10 [0.077]
	shared-SNPs:TD [bias]	72 [0.28]	69 [0.46]	80 [0.14]
	total FD	127	87	4

anticancer drugs were screened for 500 cell lines on average. In both studies, the drug sensitivity was assessed by IC_{50}, which is half-maximal inhibitory concentration.

3.4.1 Evaluation of prediction model using training/testing data

Of the 130 drugs analyzed by Garnett et al. [2012], 41 have non-missing IC_{50} values in fewer than 331 cell lines, while the other 89 drugs have non-missing IC_{50} values in more than 561 cell lines. These 89 drugs were grouped by their targets, and two drugs were excluded from our analysis because they do not group with any other drugs. We will first study these 87 drugs since a larger sample size is necessary for the following studies using training/testing sets.

Of the 87 drugs, 57, 69, and 56 are grouped by targeted family, targeted process, and targeted molecule, respectively. One drug is often grouped in multiple ways. There are four targeted families: chemotherapy, CTK (cytoplasmic/non-receptor tyrosine kinase), RTK (receptor tyrosine kinase), and S/T Kinase (serine/threonine protein kinase), which include 12, 7, 10, and 30 drugs respectively. There are 18 targeted processes and 24 targeted molecules groups. Most groups based on the targeted processes have fewer than 10 drugs, and the two largest groups have 17 and 20 drugs, respectively. For the targeted molecules, most of the groups have fewer than 5 drugs, and the largest group has 7 drugs. The three grouping strategies have a semi-hierarchical order: targeted family > targeted process > targeted molecule. For example, the group for the RTK targeted family includes the groups for targeted processes such as ERK Signaling and PI3K/MTOR. Furthermore, the ERK signaling targeted process includes targeted molecules such as EGFR and MET.

To evaluate the validity of the selected features, we split the cell lines into training and testing sets. For the testing set, we randomly selected 65 cell lines from those with non-missing IC_{50} values for all 87 drugs. We used the remaining cell lines as the

training set. Both the training and testing data were standardized to have a mean of 0 and a standard deviation of 1 for the response and covariates. The training data were then used for feature selection. If a drug belonged to more than one group, we took the union of the genomic features associated with that drug across the groups. Given the genomic features selected for each drug, we re-estimated the regression coefficients using the training data (denoted $\hat{\beta}_{train}$) and thus obtained a predictive model for each drug. Next, we used the testing data to estimate the percentage of the variance explained by the predictive model. Let SS_z be the sum of squares of z, and let y_{test} and X_{test} be the standardized $\log(\text{IC}_{50})$ and genomic features in the testing set. Then

$$\text{Prediction R-square} \equiv 1 - \frac{SS\boldsymbol{\epsilon}_{test}}{SSy_{test}}, \text{ where } \boldsymbol{\epsilon}_{test} = y_{test} - X_{test}\hat{\beta}_{train}.$$

The possible range for prediction R-square is $(-\infty, 1]$. A negative value clearly indicates a bad prediction, and the significance of a positive value was evaluated by the following approach. Given a drug with k associated features, we randomly chose k features from the candidate 13,847 features including 84 binary variables of cancer gene mutation statuses, 426 copy number alterations, 13,321 gene expressions, and 16 binary variables for cancer types, under the null scenario where these k features are irrelevant to the drug sensitivity. We estimated their regression coefficients using the training data, and then evaluated the prediction R-square using the testing data. We constructed the null distribution of the prediction R-square by repeating the above procedure 1,000 times, and then we calculated the p-value as the percentage of the null simulations where the prediction R-squares were greater than or equal to the observed prediction R-square. This approach is computationally efficient: it took 8 seconds on average to generate the p-value for each drug.

BipLog identified that 70 of the 87 drugs were associated with at least one ge-

nomic feature. Figures 3.2a and 3.2b show the distributions of the number of selected features and the prediction R-squares across these 70 drugs. Forty-nine (70%) of the 70 drugs had prediction R-squares greater than 0, and 17 (24%) had prediction R-squares greater than 20%. Forty-one of the drugs had significant prediction R-squares at the 0.05 significance level (Figure 3.2c), which corresponds to an estimate of FDR $= 87 \times 0.05/41 \approx 0.1$. As expected, there is a strong correlation between the prediction R-squares and their p-values, although the relationship is not monotonic (Figure 3.2d). Therefore, in practice it is helpful to consider both the size of the prediction R-square and its p-values. Overall, these results suggest that the identified genomic features could provide useful predictions of drug sensitivity. We will give specific examples in the next subsection.

3.4.2 Construction of prediction model

Next, we combined the training and testing sets and selected the genomic features using all the available data for the 87 drugs. A feature was selected by a group if it had a nonzero coefficient for at least one drug in the group. For the drugs grouped by target family, we selected 10, 4, 6, and 0 features for the groups Chemotherapy, RTK, CTK, and S/T Kinase, respectively. Figure 4.2 shows the distribution of the number of features selected per group for the targeted processes and targeted molecules.

Next we discuss a few examples, shown in Figure 3.4; the complete results can be found in the supplementary materials. Somatic mutations may lead to the fusion of two genes. The abnormal gene BCR-ABL is formed by the fusion of genes BCR and ABL; this is often observed in chronic myeloid leukemia (CML). BCR-ABL encodes a tyrosine kinase that is not regulated by cellular signals and thus causes unregulated

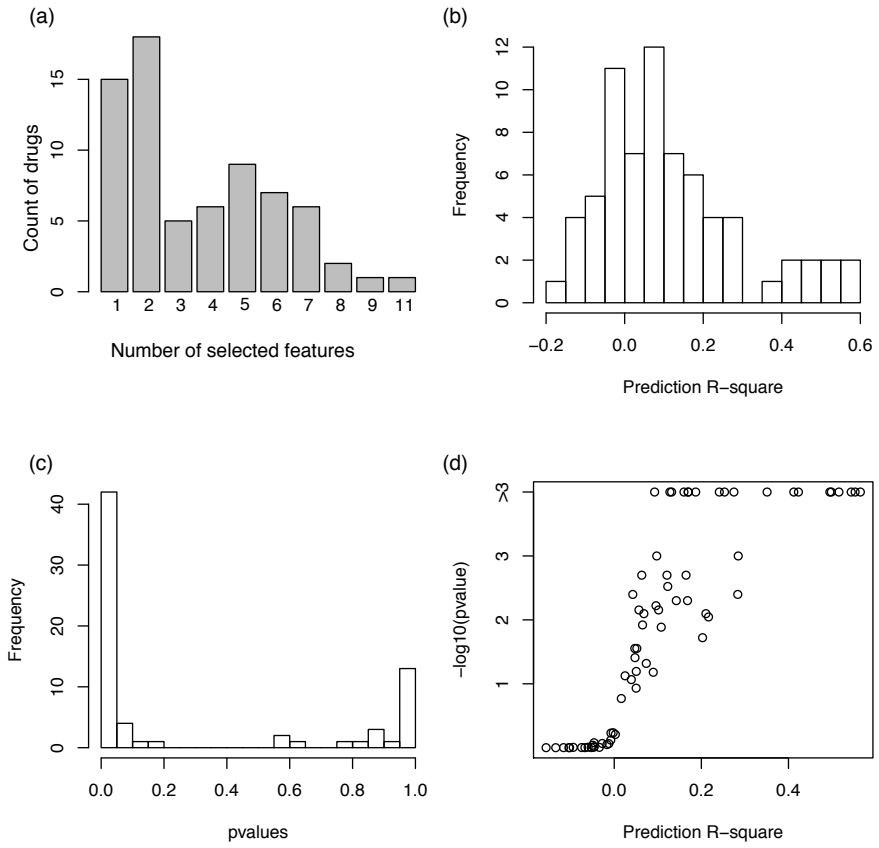

Figure 3.2: Summary of the genomic feature selection results of the within-study analysis. (a) Distribution of the number of genomic features selected per drug. (b) Distribution of the prediction R-squares for each drug. (c) Distribution of the p-values of the prediction R-squares. (d) Scatter plot of the prediction R-squares and their corresponding p-values. Since the null distribution was simulated as 1000 null samples, the smallest p-value is 0.001.

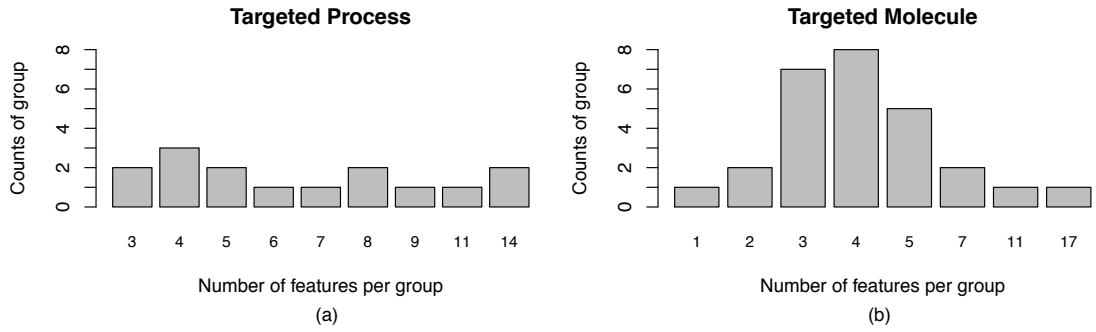

Figure 3.3: Distribution of the number of selected features by the two grouping strategies.

A. BCR_ABL

	AP-24534	Nilotinib	Bosutinib
EGFR	0.00	0.00	-0.19
AZU1	-0.20	0.00	0.00
CAV2	0.00	0.00	-0.19
BCR_ABL_MUT	-0.39	-0.67	0.00

B. MEK1/MEK2

	RDEA119	CI-1040	PD-0325901	AZD6244
PHLDA1	-0.49	-0.39	-0.43	-0.36

C. ERBB2

	Lapatinib	BIBW2992
C1ORF116	-0.31	-0.36
CYR61	-0.35	0.00
ERBB2_CN	0.00	-0.26
ERBB2_MUT	-0.31	0.00
STAM2	0.00	-0.16

D. Mitosis

	Vinorelbine	EpothiloneB	Vinblastine	Docetaxel	BX-795	SL0101-1	BI-D1870	ZM-447439	RO-3306
ABCB1	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
YAP1	0.00	0.00	-0.25	-0.24	0.00	0.00	-0.10	0.00	0.00
AXL	0.00	-0.13	0.00	-0.17	-0.36	0.00	0.00	-0.19	-0.15
blood	0.00	0.00	0.00	0.28	0.18	0.00	0.33	0.21	0.15

Figure 3.4: Genomic features associated with four groups of drugs that share the molecular targets BCR_ABL (A), MEK1/MEK2 (B), ERBB2 (C) or the process target Mitosis (D). For each group, the regression coefficient matrix is shown for those genomic features with at least one nonzero coefficient, where a row corresponds to a genomic feature and a column corresponds to a drug. The feature X_MUT is a binary indicator showing whether or not gene X has mutation; ERBB2_CN is the copy number of the gene ERBB2; blood is a binary indicator showing whether or not the cell line is derived from a blood tumor. The remaining features are gene expressions.

cell proliferation, which may lead to cancer. Three drugs that target BCR-ABL protein products are included in this study (Figure 3.4A). The sensitivity of two of these drugs is negatively correlated with the occurrence of the BCR-ABL mutation, which is expected. The negative correlation indicates that the presence of the BCR-ABL mutation is related to a reduction in $\log(\mathrm{IC}_{50})$, hence an increase in the drug sensitivity. There are two interesting new findings in this example. (1) The sensitivity of AP-24534 also increases as the expression of AZU1 increases, which is consistent with the tumor suppressor role of AZU1 [Chen et al., 2000]. (2) The sensitivity of Bosutinib is associated with the expression of two cancer-related genes, EGFR and CAV2, instead of the BCR-ABL mutation. EGFR is a signaling protein that plays an important role in many types of cancer, and CAV2 is potentially a tumor suppressor [Lee et al., 2011].

Figure 3.4B shows that when the gene encoding PHLDA1 has a higher expression, all four drugs that target MEK1/MEK2 (mitogen-activated protein (MAP) kinase) have higher sensitivity. Several previous studies have suggested that PHLDA1 may be functionally important in cancer, and some studies have shown that it functions in the MEK1/MEK2 pathway [Oberst et al., 2008]. This finding suggests that the expression of PHLDA1 could be an informative biomarker for the efficacy of cancer drugs targeting MEK1/MEK2.

The ERBB2 (also known as the HER2) gene encodes a protein product that promotes the growth of cancer cells. The amplification of the ERBB2 gene in breast cancer increases the aggressiveness of the tumor. Our analysis identifies several genes related to the two drugs targeting ERBB2 (Figure 3.4C): BIBW2992 and Lapatinib. BIBW2992 has been approved by the U.S. Food and Drug Administration for use against non-small cell lung carcinoma (NSCLC), and its efficacy for breast cancer treatment is being evaluated. Lapatinib has been approved for treatment in advanced HER2-receptor-positive breast cancer patients. As expected, we identified the ERBB2 mutation or the ERBB2 copy number variation as genomic features associated with these drugs. The novel findings are the association with the gene expressions of C1ORF116, CYR61, and STAM2. C1ORF116 interacts with SMD2 and SMD3, which are both closely related to growth-factor signaling and tumorigenesis [Tian et al., 2003]. Several studies have shown that CYR61 is involved with breast cancer tumorigenesis and progression [Tsai et al., 2002; Planque and Perbal, 2003]. Furthermore, the gene expression CYR61 has been found to be associated with stage, tumor size, and estrogen receptor expression in breast cancer patients [Xie et al., 2001]. In addition, STAM2 may be involved in "signaling by EGFR in cancer" [Croft et al., 2011]. Therefore, the combined information from the ERBB2 mutation (or copy number alterations) and gene expression of C1ORF116, CYR61,

and STAM2 may provide a more accurate prediction of drug efficacy than the ERBB2 mutation/copy number alteration alone.

Figure 3.4D presents the estimated coefficient matrix for nine drugs that target the Mitosis process. The features shared by several drugs include the expression of genes YAP1 and AXL and the blood-tissue indicator. YAP1 encodes "YES-associated protein 1," which has been shown to be related to different types of cancer [Wang et al., 2012; Rosenbluh et al., 2012]. AXL encodes a receptor tyrosine kinase, which is also involved with tumorigenesis [Hong et al., 2013]. Previous studies have shown that the protein products of YAP1 and AXL may function together [Cui et al., 2011].

3.4.3 Validation of the prediction model

We treated the data of Garnett et al. [2012] and Barretina et al. [2012] as the training and testing study data, respectively, constructed the prediction models from the training data, and then evaluated them using the testing data. Of the 24 drugs analyzed by Barretina et al. [2012], 12 were analyzed in the training study. Five of the 12 drugs had missing values in more than 325 cell lines in the training study, so they were not included in the 87 drugs in the above analysis. To address this issue, we conducted another group-wise analysis using the training data for groups involving any of these 12 drugs. Then we chose the features associated with each drug as the union of the features selected in this new analysis and those from the above analysis, whenever possible. For the 12 drugs that were analyzed using the testing data but not the training data, we fitted the prediction models using the features selected for their drug targets. For example, for the drug Topotecan, which targets the molecule TOP1, we used the features from the training study associated with the drug group that targeted TOP1 as the features associated with Topotecan.

To determine whether at least one of the selected features is associated with drug sensitivity in the testing data, we used an F-test to compare the intercept-only model and the model with all the identified genomic features. The analysis results are presented in Figure 3.5. The F-test p-values are smaller than 0.05 in most cases (note that only drugs with at least one identified genomic feature are included in the figure). The drugs PLX4720 and Lapatinib are particularly significant, with p-values of 10^{-39} and 10^{-17} respectively.

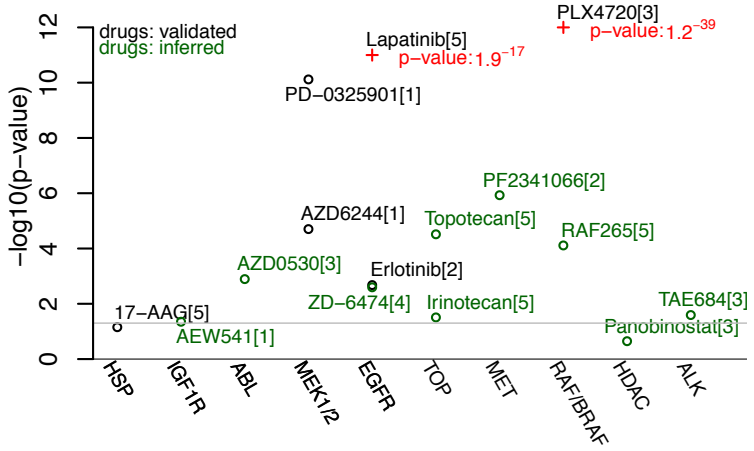

Figure 3.5: Evaluation of the predictive model in the study of Barretina et al. [2012]; the models themselves were constructed using the data of Garnett et al. [2012]. The "validated drugs" are the drugs that were analyzed in both studies. The inferred drugs are the drugs that were analyzed only in the study of Barretina et al. [2012]. The x-axis shows the drug targets, and the y-axis shows the $-\log 10$(p-values) from the F-test that compares the model with all the identified genomic features to the intercept-only model using the data of Barretina et al. [2012]. The numbers in brackets are the number of features in the corresponding prediction model.

To compare the genomic features identified by our method and the elastic net analysis in the study of Garnett et al. [2012], we calculated the prediction R-square of $\log(\text{IC}_{50})$ in the testing study for the 12 drugs that were analyzed in both the training and testing studies. Because of the disparate ranges and scales of $\log(\text{IC}_{50})$ in the two studies, as shown in Figure 3.6, we could not directly use the regression coefficients estimated from the data of Garnett et al. [2012]. Instead, we used the

following procedure to estimate the prediction R-squares. First, for each drug, we

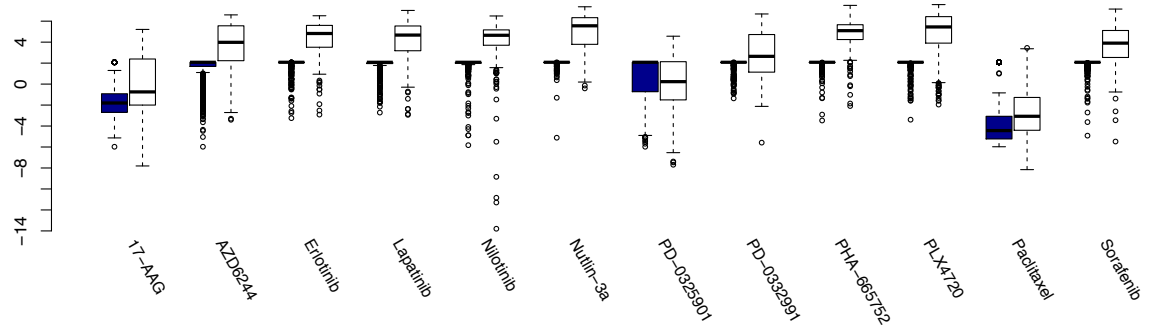

Figure 3.6: Pairwise box-plots of $\log IC_{50}$ for the 12 drugs that were analyzed in both Barretina et al. [2012] and Garnett et al. [2012]. For each drug, the blue box-plot corresponds to the study of Barretina et al. [2012], and the transparent box-plot corresponds to the study of Garnett et al. [2012].

randomly split the cell lines in the testing data into two groups of equal size, and denoted them set1 and set2. We used the cell lines of set1 to estimate the linear regression coefficients of the features identified by the training data. Then we used set2 to estimate the prediction R-square. We repeated this procedure 100 times to obtain median prediction R-squares as our final estimates. Then we applied Monte Carlo simulations to evaluate the significance of the prediction R-squares, similarly to our approach in Section 4.1.

Of the 12 drugs, 6 had a prediction R-square greater than 0 using a set of features selected by our method or the elastic net analysis in the training study. The results for these 6 drugs are presented in Table 3.2. In general, BipLog tended to choose more parsimonious models than those chosen by the elastic net, and the estimates of the prediction R-square were statistically significant in all but one case. Some of the models selected by the elastic net had greater prediction R-squares than those from BipLog, such as AZD6244 and PLX4720. However, because more variables were included in the model, they were not significantly larger than what was expected from

the null distributions.

Table 3.2: Prediction R-squares in the study of Barretina et al. [2012].

Groupwise analysis by BipLog						
Drug	17-AAG	AZD6244	PD-0325901	PLX4720	Erlotinib	Lapatinib
Prediction R^2 [Num of X]	15% [5]	2.5% [1]	7.3% [1]	27% [3]	$< 0.0\%$ [2]	21% [5]
p-value	< 0.001	< 0.001	< 0.001	< 0.001	1.00	< 0.001
Drug-by-drug analysis by Elastic Net						
Drug	17-AAG	AZD6244	PD-0325901	PLX4720	Erlotinib	Lapatinib
Prediction R^2 [Num of X]	15% [16]	10% [7]	29% [17]	29% [5]	1.5% [7]	14% [16]
p-value	< 0.001	0.275	< 0.001	0.525	0.949	0.430

CHAPTER 4: Models that are subject to unidentifiable parameters.

4.1 Introduction.

In this paper, we consider an estimation problem where a certain parameter values such as $\beta = 0$ will cause an identifiability issue. Similar problem has been addressed in the hypothesis testing framework, where a nuisance parameter ζ is present only under the alternative hypothesis ($\beta \neq 0$). Therefore, the nuisance parameter is not identifiable under the null hypothesis ($\beta = 0$). It is a non-regular testing framework since the nuisance parameter only present under the alternative hypothesis. Therefore, the standard large sample asymptotic theory cannot be directly applicable (Davtes [1977], Davies [1987]).

For the estimation problems of change points, there are extensive literatures. For instance, Bai [1997] establishes the convergence rate and asymptotic distribution for the least square estimation of a change point in multiple regression. Muggeo [2003] considers the regression models with one or more break-points parameters and utilizes a linearization technique for fitting piecewise terms in the models. He and Severini [2010] studies the theoretical properties of maximum likelihood estimators of the parameters of a multiple change-point model.

In the maximum likelihood estimation framework, due to the unidentifiable parameter (ζ) issue under null hypothesis ($\beta = 0$), the maximum likelihood estimator (MLE) have regular properties only if the likelihood function is specified correctly

with respect to the parameter value of β. Specifically, when $\beta = 0$, the parameters ζ and β should be both absent from the likelihood function; then the MLE for the rest parameters are regular. On the contrary, when $\beta \neq 0$, the parameters ζ and β are both present in the likelihood function; the MLE for all parameters are regular as well. Unlike the methods for estimation of the change points, we are interested in designing an estimation procedure which can automatically take care of the specification of correct likelihood function with respect to $\beta = 0$ or $\beta \neq 0$.

Since whether β equals to 0 plays a key role in determining the form of likelihood function, we utilize the idea of penalization estimation procedure and apply adaptive Lasso penalty to β. The adaptive Lasso penalty incorporated to β has the form: $\lambda|\beta|w$, where λ is a tuning parameter, and w stands for the adaptive weight associated to β. As shown in [Zou, 2006], given a proper chosen w, adaptive lasso performs as well as if the true underlying likelihood were given in advance.

To choose a proper weight for β, we apply the idea of constructing a test statistics in (Davtes [1977], Davies [1987]), where the author considers a test statistic, which is a function of the nuisance parameter ζ. If ζ is unknown, the test statistics takes supremum over a range of possible values of ζ. For large values of test statistic, the null hypothesis ($\beta = 0$) will be rejected. Similarly, we take the supremum of profile likelihood estimates of $\hat{\beta}(\zeta)$ over a range of possible values of ζ to be the weight for β. The weight shares similar properties of the test statistic. For large values of the weight, the true value of β is more likely to be non-zero.

The paper is organized as following. In the section 2, we present the asymptotic results for our penalized estimation procedure. Section 3 shows the simulation study

and section 4 presents a real data analysis.

4.2 Asymptotic results.

4.2.1 Notations

Let Y represent a random sample $(y_1, ...y_n)$ of observations and each $y_i, i = 1, ..., n$ is independently and identically distributed with density $\{f(y_i; \theta, \zeta) : \theta \in \Theta, \zeta \in \Xi\}$ with respect to some σ-finite measure μ. The parameter spaces Θ and Ξ are assumed to be a compact subset of metric spaces $\mathcal{R}^s$ and $\mathcal{R}^1$ respectively. The parameter θ takes the form $\theta = (\beta', \gamma')'$, and $\beta \in \mathcal{B}, \gamma \in \Gamma, \Theta = \mathcal{B} \times \Gamma$, where $\mathcal{B} \in \mathcal{R}^1$, and $\Gamma \in \mathcal{R}^{s-1}$. The likelihood function is $\mathcal{L}^{(n)}(\theta, \zeta) = \Pi_{i=1}^n f(y_i; \theta, \zeta)$ and the log-likelihood function is $l^{(n)}(\theta, \zeta) = \sum_{i=1}^n \log f(y_i; \theta, \zeta)$. Note that given that $\beta \in \mathcal{B}_0 = 0$, the parameter ζ is absent from the likelihood function so that it renders ζ to be unidentifiable, i.e. the densities are equivalent to all values of ζ at fixed values γ and $\beta = 0$. Let $\Theta_0 = \mathcal{B}_0 \times \Gamma$, and let θ_0 denote the $\theta \in \Theta_0$. In this case, since β is realized at 0, the density $f(y_i; \theta_0, \zeta)$ is independent of ζ and let $f_0(y_i; \gamma)$ denote this special class density. The corresponding likelihood function is given by $\mathcal{L}_0^{(n)}(\gamma) = \Pi_{i=1}^n f_0(y_i; \gamma)$ and $l_0^{(n)}(\gamma) = \sum_{i=1}^n \log f(y_i; \gamma)$. On the other hand, given that $\beta \in \mathcal{B}_0^c$, $\theta \in \mathcal{B}_0^c \times \Gamma = \Theta_1$, and $\zeta \in \Xi$, all parameters (θ, ζ) are identifiable.

Additionally, let $\dot{l}^{(n)}(\theta; \zeta)$ be the s-vector of partial derivatives of $l^{(n)}(\theta, \zeta)$ with respect to θ, and $\ddot{l}^{(n)}(\theta; \zeta)$ be the $s \times s$ matrix of second partial derivatives of $l^{(n)}(\theta, \zeta)$ with respect to θ. Note that $\dot{l}^{(n)}(\theta; \zeta)|_{\theta=\theta_0}$ and $\ddot{l}^{(n)}(\theta; \zeta)|_{\theta=\theta_0}$ depend on ζ in general even though $l^{(n)}(\theta, \zeta)|_{\theta=\theta_0}$ and $\mathcal{L}^{(n)}(\theta, \zeta)|_{\theta=\theta_0}$ do not [Andrews and Ploberger, 1995].

4.2.2 The estimation procedure

To address the non-identifiability issues for $\theta \in \Theta_0$ in parameter estimation, we apply the idea of penalized estimation with adaptive Lasso penalty [Zou, 2006]. Specifically, we consider the penalized log-likelihood function

$$Q^{(n)}(\theta;\zeta) = l^{(n)}(\theta;\zeta) - \lambda_n \hat{w}_n |\beta| - \lambda_n n^{-1/2} I(\beta = 0)|\zeta|, \tag{4.2.1}$$

where $\lambda_n > 0$ is a tuning parameter, and the weight $\hat{w}_n = |\hat{\beta}^*|^{-\tau}$ for some $\tau > 0$, $\hat{\beta}^*$ is $\sup_{\zeta\in\Xi} \hat{\beta}(\zeta)$, the supremum over the profiled maximum likelihood estimator of $\mathcal{L}^{(n)}(\theta;\zeta)$ at any $\zeta \in \Xi$. $\hat{\beta}(\zeta)$ is the element in $\hat{\theta}(\zeta) = (\hat{\beta}(\zeta), \hat{\gamma}(\zeta))$, which satisfies $l^{(n)}(\hat{\theta}(\zeta);\zeta) = \sup_{\theta\in\Theta} l^{(n)}(\theta;\zeta) \quad \forall \zeta \in \Xi$ with probability to 1 for $\theta \in \Theta$.

Next, we study the theoretical properties of the penalized estimator of (4.2.1) when the underlying true model is either ($\beta = 0$; unidentifiable ζ) or ($\beta \neq 0$; identifiable (θ, ζ)). The weight $\hat{w}_n$ is the key in dealing with the non-identifiability issue for $\theta \in \Theta_0$. First, we give the assumption 1 for the weight of β to ensure the desirable properties of the penalized estimator under the potential identifiability issue.

Assumption 1.

1.1 Given the likelihood function with $\theta \in \Theta_0$, where $\beta = 0$, $\sup_{\zeta\in\Xi} |\hat{\beta}(\zeta) - 0| \to_p 0$ as n goes to ∞ with rate n^α, where $\alpha > 0$.

1.2 Given the true distribution with parameter $\theta \in \Theta_1$, where $\beta \neq 0$, $\sup_{\zeta\in\Xi} |\hat{\beta}(\zeta)| \to_p c_\beta$ as n goes to ∞, where c_β is some nonzero constant.

Similar to [Andrews and Ploberger, 1995], we assume the parametric model is sufficiently regular such that the MLE $\hat{\theta}(\zeta)$ is consistent for $\theta \in \Theta_0$ uniformly over $\zeta \in \Xi$

for the above Assumption 1.1. Assumption 1.2 holds for any regular likelihood function since $\sup_{\zeta\in\Xi}|\hat{\beta}(\zeta)|$ covers the maximum likelihood estimator at the true value of ζ, which is consistent for the true parameter non-zero β. Under the assumption 1, when $\beta = 0$, the weight inflates to infinity as the sample size grows, this induces the estimator of β to be penalized to zero. Once the estimator of β is penalized to be 0, the realized equation (4.2.1) is equivalent to $l_0^{(n)}(\gamma)$, and the penalized estimator of γ is equivalent to the maximum likelihood estimator of $l_0^{(n)}(\gamma)$.

For the following theoretical discussion, we first consider the case where $(\beta = 0; \text{unidentifiable } \zeta)$, if Assumption 1 and Assumption 2 (presented in the Appendix) are hold, we will show that the penalized estimator is equivalent to the MLE of the likelihood function $\mathcal{L}_0^{(n)}(\gamma) = \Pi_{i=1}^n f_0(y_i;\gamma)$. Next, for the case where (identifiable (θ,ζ)), if Assumptions 1 and 3 (presented in the Appendix) are hold and without assuming the differentiability of likelihood function with respect to ζ, we will show that the penalized estimator is consistent. In addition, if the likelihood function is differentiable with respect to ζ, we provide a standard argument for the convergence rate and asymptotic normality of the penalized estimator. For the case where the likelihood function it not differentiable with respect to ζ, we consider the change-point model as an example, and provide specialized arguments for it.

4.2.3 Model of $(\beta = 0; \textbf{unidentifiable } \zeta)$

We establish that if $\beta = 0$, then the penalized estimation procedure is equivalent to the maximum likelihood estimation for the true likelihood $\mathcal{L}_0^{(n)}(\theta_0)$ with probability going to one asymptotically in the following Theorem 1. In other words, the estimator of the non-zero parameters γ is asymptotically equivalent to the maximum likelihood estimator for $\mathcal{L}_0^{(n)}(\theta_0)$. Its proof can be found in the Appendix.

Theorem 1. Given $\theta \in \Theta_0$, and under the assumptions 1 and 2 (presented in the Appendix) for the likelihood function, if $\lambda_n n^{-1/2} \to 0$ and $\lambda_n n^{(2\alpha\tau-1)/2} \to \infty$, then with probability to 1, there exists a maximizer $(\tilde{\theta}^{(n)}, \tilde{\zeta}^{(n)})$ of $Q^{(n)}(\theta, \zeta)$ such that $P(\tilde{\beta}^{(n)} = 0) \to 1$ as $n \to \infty$. In addition, $\|\tilde{\gamma} - \gamma_0\| = O_p(n^{-1/2})$ with $\sqrt{n}(\tilde{\gamma}^{(n)} - \gamma_0) \to_d N(0, \mathrm{I}(\gamma_0)^{-1})$, where $\mathrm{I}(\gamma_0)$ is the Fisher information matrix corresponding to $\mathcal{L}_0^{(n)}(\theta_0)$.

4.2.4 Model of ($\beta \neq 0$; identifiable (θ, ζ))

Next, we study the properties of the penalized estimation procedure given that the true β is not 0. First, we show that the penalized estimator is consistent if assumptions 1 and 3 (presented in the Appendix) are hold. Note that assumption 3 for the likelihood function does not put constraints on the differentiability of the likelihood function with respect to the parameter ζ. It is because the differentiability with respect to ζ might not be hold in general. For examples, if we consider the parametric model of the example 3 in [Davtes, 1977], let $Y = (y_1, ..., y_n)$ represent a random n independently and identically distributed sample from the density $f(y; \beta, \zeta) = (1 - \beta)e^{-y} + \beta\zeta e^{-\zeta y}$, where $1 < \zeta < \infty$. The likelihood function is differentiable with respect to parameter ζ. When $\beta = 0$, it renders ζ to be unidentifiable. However, for the linear model with change point in [Bacon and Watts, 1971],

$$Y = (\gamma_0 + \gamma_1 X) + \beta(X - \zeta)I(X > \zeta) + \epsilon, \tag{4.2.2}$$

where X is covariate, ϵ is random error following standard normal distribution, and $I(X > \zeta)$ is an indicator variable: $I(X > \zeta) = 1$ if $X > \zeta$. The likelihood function for this model is not differentiable with respect to ζ. Similarly, if $\beta = 0$, it renders ζ to be unidentifiable.

We first establish consistency property of the penalized estimator $(\tilde{\theta}^{(n)}, \tilde{\zeta}^{(n)})$ without assuming the differentiability of likelihood function with respect to ζ in the Theorem 2.

Theorem 2. Given $\beta \neq 0$, and under the assumptions 1 and 3 (the assumption 3 is from Wald [1949] presented in the Appendix), then $P(\lim_{n\to\infty}(\tilde{\theta}^{(n)}, \tilde{\zeta}^{(n)}) = (\theta, \zeta)) = 1$, where $(\theta, \zeta) \in \Theta_1 \times \Xi_1$. (The proof of Theorem 2 can be found in the Appendix.)

The argument to establish the property of the asymptotic normality of the penalized estimator depends on the differentiability of the likelihood function with respect to ζ. If Assumptions 1 3, and 4 (presented in the Appendix): the likelihood function is second order differentiable with respect to ζ, holds, Theorem 3 gives that the asymptotic normality of the penalized estimator $(\tilde{\theta}^{(n)}, \tilde{\zeta}^{(n)})$.

Theorem 3. Under the assumptions 1, 2, and 4 for likelihood functions and assume $(\theta, \zeta) \in \Theta_1 \times \Xi_1$, if $\lambda_n n^{-1/2} \to 0$, then $\sqrt{n}((\tilde{\theta}^{(n)}, \tilde{\zeta}^{(n)}) - (\theta, \zeta)) \to_d N(0, \mathrm{I}(\theta, \zeta)^{-1})$, where $\mathrm{I}(\theta, \zeta)$ is the fisher information matrix. The proof of Theorem 3 is a direct adaption from Fan and Li [2001]; therefore, it is omitted.

However, if the Assumption 4 does not hold, it is required to address this issue specifically to prove the asymptotic normality of the penalized estimator. In this paper, we take the one change point model as an example, and establish the asymptotic normality of the penalized estimator for this particular model, where the likelihood function is not differentiable with respect to ζ.

Asymptotic normality of the penalized estimator for the one change point model

Consider $(y_1, ..., y_n)$ to represent a independent n samples, where $y_i = (z_i, w_i)$, z_i is a random vectors, and w_i is a 1-dimension random variable associated to the change point ζ. Specifically, the sample points y_i are assumed to be independently drawn from the parametric model:

$$f_1(y_i;\gamma) \quad \text{for } w_i \leq \zeta; \quad f_2(y_i;\gamma,\beta) \quad \text{for } w_i > \zeta, \tag{4.2.3}$$

This model describes that the density function f of each sample i depends on the value of w_i with respect to the thresholding parameter ζ. Furthermore, to adapt a similar set-up from [He and Severini, 2010], the n sample can be rearranged to an order as drawn from the one change point model by $0 < n_1 < n$ based on w_i and ζ,

$$f_1(y_i;\gamma) \quad \text{for } 1 \leq i \leq n_1, \text{ where } w_i \leq \zeta \tag{4.2.4}$$

$$f_2(y_i;\gamma,\beta) \quad \text{for } n_1 + 1 \leq i \leq n, \text{ where } w_i > \zeta. \tag{4.2.5}$$

Clearly, both n_1 and ζ are change point parameters. An known parameter ζ and the observed values of w_i for each sample can transform to an known parameter n_1. The parameter γ, a vector with $s-1$ elements is unknown common parameter for f_1 and f_2, and β is an one-dimensional unknown within-segment parameter for f_2. Note that if $\beta = 0$, $f_2(y_i;\gamma,\beta=0)$ is equivalent to $f_1(y_i;\gamma)$, which renders ζ to be unidentifiable. The corresponding log likelihood function is

$$\acute{l} \equiv l(\theta,\zeta) = \sum_{i=1}^{n} \log\{I(w_i \leq \zeta) f_1(y_i;\gamma) + I(w_i > \zeta) f_2(y_i;\gamma,\beta)\} \tag{4.2.6}$$

. (4.2.7)

Theorem 4. Given $\beta \neq 0$ and under the Assumptions 5 from [He and Severini, 2010], which are presented in the Appendix for likelihood function, if $\lambda_n n^{-1/2} \to 0$,

then $\sqrt{n}(\tilde{\theta}^{(n)} - \theta) \to_d N(0, \mathrm{I}(\theta)^{-1})$, where $\theta \in \Theta_1$.

Theorem 4 establishes the asymptotic normality property of the penalized estimator for one change point model. The corresponding proof can be found in the Appendix.

The following simulation section is based on the one change point model to evaluate the empirical performance of the penalized estimator.

4.3 Simulation studies

To evaluate the performance of the penalization estimation approach, we compare our method to the regular maximum likelihood estimation procedure by a set of simulated data under the non-identifiable or identifiable model scenarios. Specifically, we consider $y_i = (z_i, w_i)$ following the one change point model:

$$z_i = \delta_0 + w_i\delta_1 + w_i\beta I(w_i > \zeta) + \epsilon_i. \tag{4.3.1}$$

Covariates $X = (x_1, ...x_n)$ and $W = (w_1, ...w_n)$ are both simulated as n i.i.d standard normal random variables respectively. The change point ζ is set as 0, the intercept δ_0 is set as 0.5, and the slope δ_1 is set as 0.25. Additionally, each residual ϵ_i is independently and identically distributed random variable following normal distribution with mean 0 and variance 0.5. We consider 16 situations involving different combinations of sample size n and effect sizes of β: $n = 50$, 200, 1000, or 3000, and $\beta = 0$, 0.5, 1.0, or 2.0. The response variables $z = (z_1, ..., z_n)$ are simulated accordingly based on (4.3.1).

The likelihood incorporated into the penalized estimation procedure is from one

change point model for all 16 simulation scenario. When $\beta = 0$, which corresponds to no change point model, the maximum likelihood estimation is based on the likelihood from the model without change point or the model with one change point with unknown change point position. On the other hand, when $\beta \neq 0$, the maximum likelihood estimation is based on the likelihood from one change point model with known or unknown change point position.

To estimate the position of the unknown change point, an arbitrary interval $(-2, 2)$ by 0.01 increment is provided for both penalized and regular maximum likelihood estimation procedures. In addition, an initial set of 25 tuning parameter λ is provided for penalized estimation. Bayesian Information Criterion (BIC) is used for the selection of tuning parameter.

Each simulation scenario consists of 1000 replications. We calculate the mean, median, mean and median of the model based variance, empirical variance and coverage probability of the true values of δ_1 and β. Let PMLE denote our penalized estimation procedure. Additionally, let MLE^n, MLE^c and MLE^{uc} denote regular maximization likelihood estimation approach for model without change point, one change point model with known position or unknown position respectively.

The results shown in Table 1 are for the not identifiable model where $\beta = 0$. They suggest that as sample size becomes larger, the performance of the penalized estimator becomes similar to that of regular maximum likelihood estimation based on model without change point.

The results shown in Tables 2 to 4 are for the identifiable model where $\beta \neq 0$. They suggest that as sample size and the effect size of β become larger, the performance of the penalized estimator becomes similar to that of regular maximum likelihood estimation based on one change point model with unknown change point position.

Table 4.1: Empirical studies of the penalized estimation procedure for model without change point, with sample size n = (50, 200, 1000, or 3000), $\delta_1 = 0.25$ and $\beta = 0$. For penalization estimation approach, the tuning parameter is selected to minimize BIC. We compare the penalization estimation approach to regular maximum likelihood function estimation for one change point model with known or unknown change point. We present the mean of the estimates of δ_1, β and ζ across 1000 simulations. In addition, we present the mean of model-based variance estimator, and empirical variance estimator for δ_1 and β. Moreover, the coverage probabilities for δ_1 and β are presented.

$\delta_1 = 0.25$	$\tilde{\delta}_1$	$\text{var}^m(\tilde{\delta}_1)$	$\text{var}^e(\tilde{\delta}_1)$	cover	# sim	$\tilde{\beta}$	$\text{var}^m(\tilde{\beta})$	$\text{var}^e(\tilde{\beta})$	$\tilde{\zeta}$
$\beta = 0$	mean	mean		P	$\tilde{\beta} = 0$	mean	mean		mean
$n = 50$									
MLE^n	0.245	0.01019	0.0117	0.925	–	–	–	–	–
MLE^{uc}	0.246	0.02074	0.04109	0.83	–	-0.005	0.07312	0.21068	-0.304
PMLE^{bic}	0.247	0.01047	0.01483	0.909	966	-0.002	0.05699	0.02047	-0.018
$n = 200$									
MLE^n	0.252	0.00274	0.00263	0.955	–	–	–	–	–
MLE^{uc}	0.257	0.00779	0.02145	0.755	–	-0.008	0.01716	0.06346	-0.031
PMLE^{bic}	0.251	0.00281	0.00363	0.944	987	0	0.01708	0.00304	-0.001
$n = 1000$									
MLE^n	0.252	0.00052	0.00053	0.946	–	–	–	–	–
MLE^{uc}	0.254	0.00163	0.00505	0.752	–	-0.002	0.00352	0.01487	-0.094
PMLE^{bic}	0.252	0.00053	0.00062	0.942	995	-0.001	0.00329	0.00027	-0.002
$n = 3000$									
MLE^n	0.25	0.00017	0.00018	0.936	–	–	–	–	–
MLE^{uc}	0.251	0.00051	0.00147	0.769	–	-0.003	0.00112	0.00444	0.042
PMLE^{bic}	0.25	0.00017	0.00019	0.934	998	0	0.00097	3e-05	0

–: Not applicable.
MLE^n: MLE with likelihood for no change point model.
MLE^{uc}: MLE with likelihood for unknown change point position.
PMLE^{bic}: PMLE using BIC to select tuning parameter.
var^m: Model based variance estimator.
var^e: Empirical variance estimator.

4.4 Real data analysis

To evaluate the performance of the penalized estimation procedure, we analyze three data sets.

Table 4.2: Empirical studies of the penalized estimation procedure for one change point model, with sample size $n = (50, 200, 1000, \text{or } 3000)$, $\delta_1 = 0.25$ and $\beta = 0.5$. For penalization estimation approach, the tuning parameter is selected to minimize BIC. We compare the penalization estimation approach to regular maximum likelihood function estimation for one change point model with known or unknown change point. We present the mean of the estimates of δ_1, β and ζ across 1000 simulations. In addition, we present the mean of model-based variance estimator, and empirical variance estimator for δ_1 and β. Moreover, the coverage probabilities for δ_1 and β are presented.

$\delta_1 = 0.25$	$\tilde{\delta}_1$	$\text{var}^m(\tilde{\delta}_1)$	$\text{var}^e(\tilde{\delta}_1)$	cover	# sim	$\tilde{\beta}$	$\text{var}^m(\tilde{\beta})$	$\text{var}^e(\tilde{\beta})$	cover	$\tilde{\zeta}$
$\beta = 0.5$	mean	mean		P	$\tilde{\beta} = 0$	mean	mean		P	mean
$n = 50$										
MLE^c	0.249	0.03271	0.03533	0.933	–	–	–	–	–	–
MLE^{uc}	0.244	0.02224	0.03151	0.892	–	0.505	0.0565	0.08905	0.889	-0.144
PMLE^{bic}	0.396	0.01177	0.04144	0.396	759	0.199	0.05326	0.13256	0.198	-0.029
$n = 200$										
MLE^c	0.258	0.01001	0.01028	0.955	–	–	–	–	–	–
MLE^{uc}	0.274	0.00709	0.00956	0.868	–	0.451	0.01825	0.02231	0.912	-0.023
PMLE^{bic}	0.353	0.00495	0.02312	0.492	467	0.294	0.01868	0.08198	0.518	0.014
$n = 1000$										
MLE^c	0.253	0.00185	0.00191	0.94	–	0.496	0.00528	0.00551	0.948	–
MLE^{uc}	0.272	0.00158	0.00223	0.859	–	0.457	0.00419	0.00642	0.812	0.007
PMLE^{bic}	0.272	0.00158	0.00223	0.859	0	0.457	0.00419	0.00642	0.812	0.007
$n = 3000$										
MLE^c	0.249	0.00061	0.00063	0.942	–	0.502	0.00185	0.00185	0.954	–
MLE^{uc}	0.26	0.00057	0.00081	0.876	–	0.48	0.00165	0.00249	0.858	-0.002
PMLE^{bic}	0.26	0.00057	0.00081	0.876	0	0.48	0.00165	0.00249	0.858	-0.002

–: Not applicable.
MLE^c: MLE with likelihood for known change point position.
MLE^{uc}: MLE with likelihood for unknown change point position.
PMLE^{bic}: PMLE using BIC to select tuning parameter.
var^m: Model based variance estimator.
var^e: Empirical variance estimator.

4.4.1 Stagnant band height data example

The first dataset we consider is from [Bacon and Watts, 1971], which were originally obtained from the Ph.D. thesis of R. A. Cook. The dataset is from the experiments to study the relationship between the stagnant-band-height and the controlled various flow rate of water down an inclined channel using different surfactants. The response variable is the logarithm of stagnant surface layer height in centimeters and the predictor variable is the logarithm of the water flow rate in grams per centimeter

Table 4.3: Empirical studies of the penalized estimation procedure for one change point model, with sample size $n = (50, 200, 1000, \text{or } 3000)$, $\delta_1 = 0.25$ and $\beta = 1.0$. For penalization estimation approach, the tuning parameter is selected to minimize BIC. We compare the penalization estimation approach to regular maximum likelihood function estimation for one change point model with known or unknown change point. We present the mean of the estimates of δ_1, β and ζ across 1000 simulations. In addition, we present the mean of model-based variance estimator, and empirical variance estimator for δ_1 and β. Moreover, the coverage probabilities for δ_1 and β are presented.

$\delta_1 = 0.25$	$\tilde{\delta}_1$	$\text{var}^m(\tilde{\delta}_1)$	$\text{var}^e(\tilde{\delta}_1)$	cover	# sim	$\tilde{\beta}$	$\text{var}^m(\tilde{\beta})$	$\text{var}^e(\tilde{\beta})$	cover	$\tilde{\zeta}$
$\beta = 1.0$	mean	mean		P	$\tilde{\beta} = 0$	mean	mean		P	mean
$n = 50$										
MLE^c	0.246	0.03756	0.03958	0.93	–	–	–	–	–	–
MLE^{uc}	0.28	0.02729	0.03544	0.9	–	0.939	0.07679	0.09672	0.905	0.009
PMLE^{bic}	0.374	0.02195	0.07847	0.63	320	0.738	0.07501	0.29628	0.651	0.016
$n = 200$										
MLE^c	0.254	0.00916	0.00863	0.955	–	–	–	–	–	–
MLE^{uc}	0.301	0.00737	0.01068	0.843	–	0.903	0.02148	0.03362	0.802	-0.026
PMLE^{bic}	0.301	0.00736	0.01111	0.843	3	0.902	0.0215	0.03521	0.802	-0.023
$n = 1000$										
MLE^c	0.25	0.00178	0.00172	0.952	–	–	–	–	–	–
MLE^{uc}	0.267	0.00167	0.00203	0.894	–	0.965	0.00498	0.00631	0.866	-0.007
PMLE^{bic}	0.267	0.00167	0.00203	0.894	0	0.965	0.00498	0.00631	0.866	-0.007
$n = 3000$										
MLE^c	0.249	0.00065	0.00064	0.955	–	1.002	0.00183	0.00172	0.959	–
MLE^{uc}	0.257	0.00063	0.00072	0.922	–	0.985	0.00176	0.00208	0.903	0
PMLE^{bic}	0.257	0.00063	0.00072	0.922	0	0.985	0.00176	0.00208	0.903	0

–: Not applicable.
MLE^c: MLE with likelihood for known change point position.
MLE^{uc}: MLE with likelihood for unknown change point position.
PMLE^{bic}: PMLE using BIC to select tuning parameter.
var^m: Model based variance estimator.
var^e: Empirical variance estimator.

per second.

We fit a linear regression model to this data: $z_i = \delta_0 + w_i'\delta_1 + w_i\beta I(w_i > \zeta) + \epsilon_i$, where z_i and w_i stand for the response and predictor variables respectively. The obtained model is $E(z_i) = 0.46 - 0.46w_i - 0.53w_i I(w_i > 0.29)$; the left plot in figure 1 shows the sample points with the fitted line.

Table 4.4: Empirical studies of the penalized estimation procedure for one change point model, with sample size $n = (50, 200, 1000, \text{or } 3000)$, $\delta_1 = 0.25$ and $\beta = 2.0$. For penalization estimation approach, the tuning parameter is selected to minimize BIC. We compare the penalization estimation approach to regular maximum likelihood function estimation for one change point model with known or unknown change point. We present the mean of the estimates of δ_1, β and ζ across 1000 simulations. In addition, we present the mean of model-based variance estimator, and empirical variance estimator for δ_1 and β. Moreover, the coverage probabilities for δ_1 and β are presented.

$\delta_1 = 0.25$	$\tilde{\delta}_1$	$\text{var}^m(\tilde{\delta}_1)$	$\text{var}^e(\tilde{\delta}_1)$	cover	# sim	$\tilde{\beta}$	$\text{var}^m(\tilde{\beta})$	$\text{var}^e(\tilde{\beta})$	cover	$\tilde{\zeta}$
$\beta = 2.0$	mean	mean		P	$\tilde{\beta} = 0$	mean	mean		P	mean
$n = 50$										
MLE^c	0.257	0.03159	0.03223	0.939	–	–	–	–	–	–
MLE^{uc}	0.323	0.02612	0.03595	0.865	–	1.839	0.07382	0.11555	0.795	-0.058
PMLE^{bic}	0.323	0.02612	0.03595	0.865	0	1.839	0.07382	0.11555	0.795	-0.058
$n = 200$										
MLE^c	0.249	0.01041	0.01049	0.95	–	2.005	0.03322	0.03494	0.94	–
MLE^{uc}	0.292	0.00947	0.0126	0.889	–	1.912	0.02919	0.04387	0.845	0.003
PMLE^{bic}	0.292	0.00947	0.0126	0.889	0	1.912	0.02919	0.04387	0.845	0.003
$n = 1000$										
MLE^c	0.25	0.00185	0.00184	0.956	–	2.002	0.00552	0.00544	0.952	–
MLE^{uc}	0.263	0.0018	0.0021	0.922	–	1.975	0.00534	0.0064	0.911	0.002
PMLE^{bic}	0.263	0.0018	0.0021	0.922	0	1.975	0.00534	0.0064	0.911	0.002
$n = 3000$										
MLE^c	0.251	0.00065	0.00064	0.95	–	1.998	0.00182	0.0018	0.953	–
MLE^{uc}	0.258	0.00064	0.00066	0.936	–	1.985	0.0018	0.00194	0.921	-0.003
PMLE^{bic}	0.258	0.00064	0.00066	0.936	0	1.985	0.0018	0.00194	0.921	-0.003

–: Not applicable.
MLE^c: MLE with likelihood for known change point position.
MLE^{uc}: MLE with likelihood for unknown change point position.
PMLE^{bic}: PMLE using BIC to select tuning parameter.
var^m: Model based variance estimator.
var^e: Empirical variance estimator.

4.4.2 Metabolic pathways data example

The second data is from [Julious, 2001], which aims to study the switch of metabolic pathways in persons during physical exercise. To examine whether the metabolic pathways change from aerobic to anaerobic when people produce energy during exercise, volume of carbon dioxide exhaled and volume of oxygen inhaled are measured for outcome and predictor variables. The measurements are taken on a single person per every 30 seconds up to a maximum of 17.5 minutes. A linear regression

model with a change point between two segmented lines was fitted to this data, where the change point, if exists, represents the switch between metabolic pathways. As shown in [Julious, 2001], the best fitting two segmented lines model is

$$z_i = 0.076 + 0.042 w_i \ \ (w_i \leq 39.46); \quad z_i = -1.659 + 0.086 w_i \ \ (39.46 < w_i).$$

We also fit a linear regression model with two segmented lines to this data, and the obtained model is the same as above. The right plot in figure 1 shows the sample points with the fitted line.

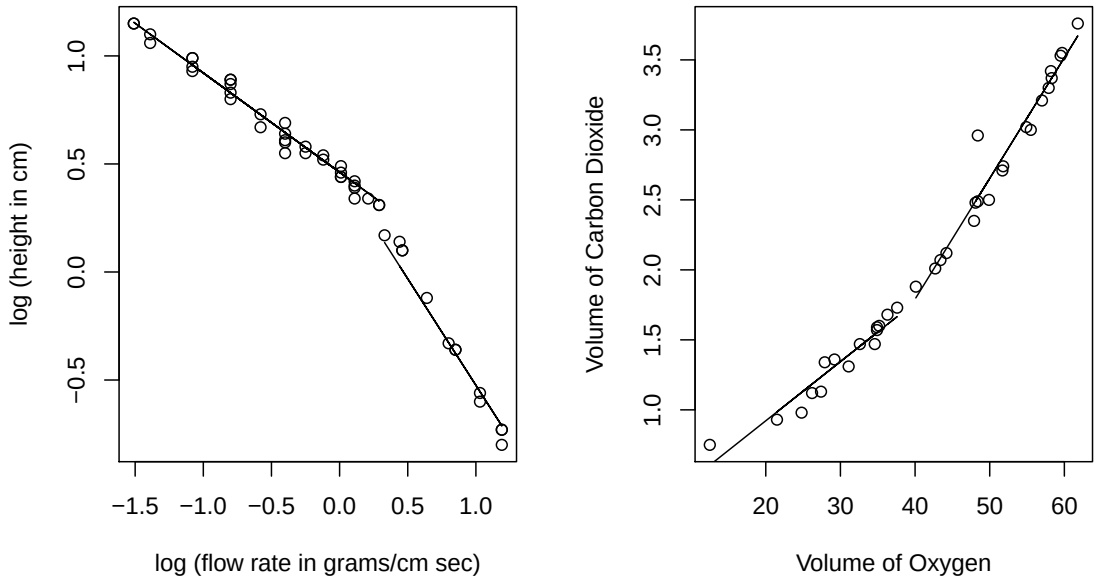

Figure 4.1: The left plot is for the stagnant band height data from [Bacon and Watts, 1971] with the fitted line by the penalized estimation procedure. The right plot is for the metabolic pathways data from [Bacon and Watts, 1971] with the fitted line by the penalized estimation procedure.

4.4.3 Drug sensitivity data example

In a panel of several hundred human cancer cell lines, [Garnett et al., 2012] measured the sensitivities of drugs and genomic factors. We have applied our method to

the data of drug 17-AAG to identify the change point linear relationship between the response variable, the half-maximal inhibitory concentration (IC_{50}), i.e., the amount of drugs to kill 50% of the cancer cells on the log scale, and the logarithm of gene expression. Drug 17-AAG is an HSP90 inhibitor, and has shown its significant anti-tumor activity for various types of cancer in clinical studies Usmani et al. [2009]. Figure 2 presents the relationship between the $\log(IC_{50})$ of 17-AAG and two gene expression variables NQO1 and ZNF273 on the log scale. The gene expression of NQO1 is upregulated in livers of hepatocarcinoma patients, and has been studied its role in the cancer development Joseph et al. [1994]. Moreover, it has been suggested a promising therapeutic target for pancreatic cancer Ough et al. [2005]. ZNF273 is a member of the zinc-finger protein family and involved in transcriptional regulation (NCBI Gene ID: 10793). The analysis results by our method suggest that the relationship between $\log(IC_{50})$ of 17-AAG and NQO1 gene expression on log scale is simple linear trend while a change-point linear model fits better for the relationship between $\log(IC_{50})$ of 17-AAG and ZNF273 gene expression on log scale.

4.5 Additional conditions and asymptotic results

Following [Andrews and Ploberger, 1995],the likelihood function $\mathcal{L}^{(n)}(\theta, \zeta)$ is assumed to satisfy the following assumption 2. The Assumption 2.3 is the simplified version assuming for nontrending data.

Assumption 2.

2.1 $\mathcal{L}^{(n)}(\theta, \zeta)$ does not depend on ζ for all $\theta \in \Theta_0$.

2.2 $l^{(n)}(\theta; \zeta)$ is twice differentiable with respect to θ for all $(\theta, \zeta) \in \Theta_0 \times \Xi$.

2.3 $\sup_{\zeta \in \Xi,\ \theta \in \Theta_0} | - n^{-1} \ddot{l}^{(n)}(\theta; \zeta) - \mathbf{I}(\theta; \zeta)| \to_p 0$, where $\mathbf{I}(\theta; \zeta)$ is the asymptotic information matrix for $\theta \in \Theta_0$ at a given $\zeta \in \Xi$, which depends on both θ and

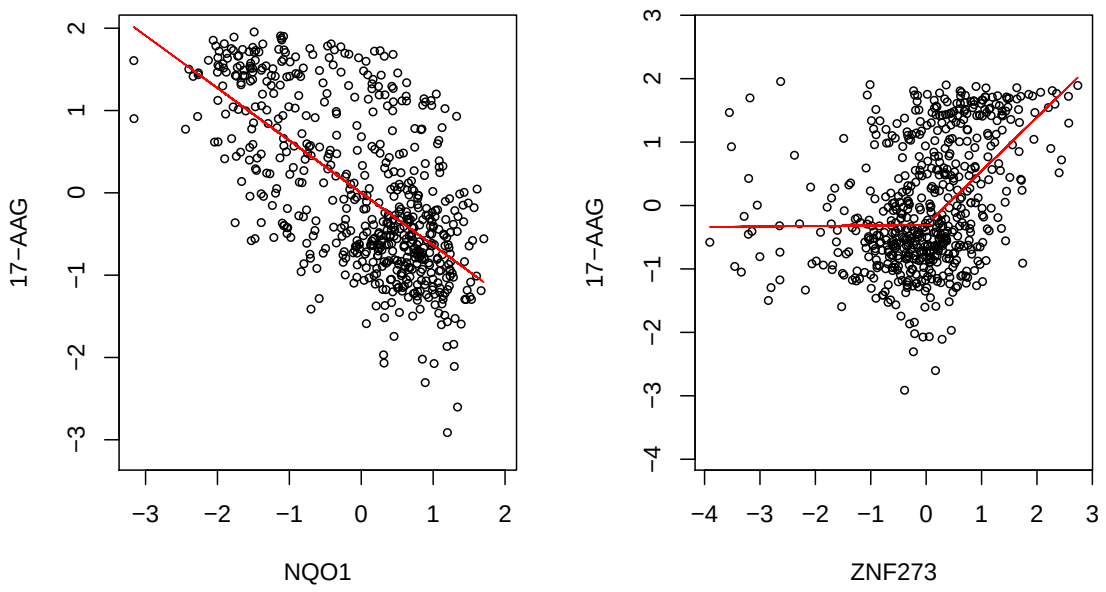

Figure 4.2: The left plot shows the fitted line for $\log(\text{IC}_{50})$ of 17-AAG and the gene expression of NQO1 on log scale by the penalized estimation procedure. The right plot is for $\log(\text{IC}_{50})$ of 17-AAG and the gene expression of ZNF273 on log scale.

ζ, and is assumed to be positive definite.

2.4 For each $\theta \in \Theta_0$, $n^{-1/2}\dot{l}^{(n)}(\theta;\cdot) \to G(\theta,\cdot)$, as processes indexed by $\zeta \in \Xi$ for some mean zero $\mathcal{R}^s$-valued Gaussian stochastic process $\{G(\theta,\zeta) : \zeta \in \Xi\}$ that has $E[G(\theta,\zeta)G(\theta,\zeta)^{\mathsf{T}}] = \mathbf{I}(\theta;\zeta) \quad \forall \zeta \in \Xi$ and has continuous sample path as functions of ζ for a fixed θ with probability 1. Moreover, we assumed that $\sup_{\zeta\in\Xi} n^{-1/2}\dot{l}^{(n)}(\theta;\zeta) = O_p(1)$

2.5 For all $j,k,l = 1,...s$, and all θ in some neighborhood of $\theta \in \Theta_0$,

$$\sup_{\zeta\in\Xi} \left|\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial\theta_l} \log f(y_i;\theta,\zeta)\right|$$

are dominated by integrable functions.

Assumption 3. Following [Wald, 1949], let $F(y;(\theta,\zeta))$ denote the cumulative distribution function of y_i; i.e., $F(y;(\theta,\zeta)) = P(y_i < y)$. Additionally, let $f(y;(\theta,\zeta),\rho)$ be the supremum of $f(y;(\theta',\zeta'))$ with respect to (θ',ζ') when $\|(\theta,\zeta) - (\theta',\zeta')\| \leq \rho$. For any positive r, let $\pi(y,r)$ be the supremum of $f(y;(\theta',\zeta'))$ with respect to (θ',ζ') when $\|(\theta,\zeta)\| \geq r$. In addition, let $f^*(y;(\theta,\zeta),\rho) = f(y;(\theta,\zeta),\rho)$ when $f(y;(\theta,\zeta),\rho) > 1$, and $= 1$ otherwise. Also, let $\pi^*(y,r) = \pi(y,r)$ when $\pi(y,r) > 1$, and $= 1$ otherwise. (The following assumptions 3.1 to 3.4 are from [Wald, 1949].)

3.1 $F(y;(\theta,\zeta))$ is either discrete or absolutely continuous for all $(\theta,\zeta) \in \Theta_1 \times \Xi_1$.

3.2 For sufficiently small ρ and large r, the expected values $\int_{-\infty}^{\infty} \log f^*(y;(\theta,\zeta),\rho) dF(y;(\theta_1,\zeta_1))$ and $\int_{-\infty}^{\infty} \log \pi^*(y,r) dF(y;(\theta_1,\zeta_1))$ are finite, where $(\theta_1,\zeta_1) \in \Theta_1 \times \Xi_1$ denote the true parameter.

3.3 If (θ_j,ζ_j) is a parameter point different from the true parameter $(\theta,\zeta) \in \Theta_1 \times \Xi_1$, then $F(y;(\theta_j,\zeta_j)) \neq F(y;(\theta,\zeta))$ for at least one value of y.

3.4 For $(\theta,\zeta) \in \Theta_1 \times \Xi_1$, $\int_{-\infty}^{\infty} |\log f(y;(\theta,\zeta))| dF(y;(\theta,\zeta)) < \infty$.

3.5 $\lambda_n \mathcal{L}^{(n)}(\theta_1;\zeta_1)^{-1} \to 0$.

Assumption 4

$l^{(n)}(\theta,\zeta)$ is second order differentiable with respect to ζ.

Proof of Theorem 1

The proof consists of three parts. The first part is to show that there exists a local maximizer $(\tilde{\theta}^{(n)},\tilde{\zeta}^{(n)})$ of $Q^{(n)}(\theta,\zeta)$ satisfying $\|\tilde{\theta}^{(n)} - \theta_0\| = O_p(n^{-1/2})$, and the second part is to prove $P(\tilde{\beta}^{(n)} = 0) \to 1$ as $n \to \infty$. The final part is to show the asymptotic

normality of $\tilde{\gamma}^{(n)}$.

First, we show that there exists a local maximizer $(\tilde{\theta}^{(n)}, \tilde{\zeta}^{(n)})$ of $Q^{(n)}(\theta, \zeta)$ satisfying $\|\tilde{\theta}^{(n)} - \theta_0\| = O_p(n^{-1/2})$. Letting $\tilde{\theta}^{(n)} = \theta_0 + n^{-1/2}\mathbf{u}$ for a fixed $\mathbf{u} = (u_1, u_2)^\mathsf{T}$, where u_1 corresponds to β, and u_2 corresponds to γ. In particular, for any $\epsilon > 0$, there exists a large enough constant C such that

$$P\left\{\sup_{\zeta\in\Xi}\left[\sup_{\|\mathbf{u}\|=C} Q^{(n)}(\theta_0 + n^{-1/2}\mathbf{u};\zeta) - Q^{(n)}(\theta_0;\zeta)\right] < 0\right\} \geq 1-\epsilon, \tag{4.5.1}$$

where $Q^{(n)}(\theta_0;\zeta) = \mathcal{L}^{(n)}(\theta_0) - \frac{\lambda_n}{\sqrt{n}}|\zeta|$.

The first step is to consider at a fixed ζ, we want to show that there exists a large enough constant C_ζ for any $\epsilon > 0$ such that

$$P\left\{\sup_{\|\mathbf{u}\|=C_\zeta}\left[Q^{(n)}(\theta_0 + n^{-1/2}\mathbf{u};\zeta) - Q^{(n)}(\theta_0;\zeta)\right] < 0\right\} \geq 1-\epsilon. \tag{4.5.2}$$

Define

$$D^{(n)}(\mathbf{u};\zeta) = Q^{(n)}(\theta_0 + n^{-1/2}\mathbf{u};\zeta) - Q^{(n)}(\theta_0;\zeta).$$

By Taylor expansion of $\mathcal{L}^{(n)}(\theta_0 + n^{-1/2}\mathbf{u};\zeta)$ at θ_0, we can obtain

$$\begin{aligned}
D^{(n)}(\mathbf{u};\zeta) &= n^{-1/2}\dot{l}_n(\theta_0;\zeta)^{\mathsf{T}}\mathbf{u} \\
&\quad - - \frac{1}{2n}\mathbf{u}^{\mathsf{T}}\left\{\ddot{l}_n(\theta_0;\zeta)\right\}\mathbf{u} \\
&\quad + (n^{-3/2}/6)\sum_{j,k,l=1}^{s}\dddot{l}_n(\check{\theta};\zeta)\mathbf{u}_j\mathbf{u}_k\mathbf{u}_l \\
&\quad + \lambda_n n^{-1/2}\hat{w}_n n^{1/2}\left\{|\beta_0| - |\beta_0 + n^{-1/2}u_1|\right\} \\
&\quad + \frac{\lambda_n}{\sqrt{n}}\left\{I(\beta_0 = 0) - I(\beta_0 + n^{-1/2}u_1 = 0)\right\}|\zeta| \\
&= n^{-1/2}\dot{l}_n(\theta_0;\zeta)^{\mathsf{T}}\mathbf{u} \\
&\quad - - \frac{1}{2n}\mathbf{u}^{\mathsf{T}}\left\{\ddot{l}_n(\theta_0;\zeta)\right\}\mathbf{u} \\
&\quad + (n^{-3/2}/6)\sum_{j,k,l=1}^{s}\dddot{l}_n(\check{\theta};\zeta)\mathbf{u}_j\mathbf{u}_k\mathbf{u}_l \\
&\quad + \lambda_n n^{-1/2}\hat{w}_n\left\{-|u_1|\right\} \\
&\quad + \frac{\lambda_n}{\sqrt{n}}\left\{1 - I(\beta_0 + n^{-1/2}u_1 = 0)\right\}|\zeta| \\
&\equiv \Gamma_1^{(n)}(\zeta) - \Gamma_2^{(n)}(\zeta) + \Gamma_3^{(n)}(\zeta) + \Gamma_4^{(n)} + \Gamma_5^{(n)}(\zeta),
\end{aligned}$$

for some $\check{\theta}$ between θ_0 and $\theta_0 + n^{-1/2}\mathbf{u}$.

By the assumptions 2.3 and 2.4, $\Gamma_1^{(n)} = O_p(1)$, $\Gamma_2^{(n)} \to \mathrm{I}(\theta_0;\zeta)$, the asymptotic information matrix. Both $\Gamma_3^{(n)}$ and $\Gamma_5^{(n)}$ converge to zero by assumption. If $u_1 \neq 0$, then $\Gamma_4^{(n)} = \lambda_n n^{(2\alpha\tau-1)/2}|n^{\alpha}\hat{\beta}^*|^{-\tau}|u_1| \to -\infty$ by $n^{\alpha}\hat{\beta}^* = O_p(1)$ and $\lambda_n n^{(2\alpha\tau-1)/2} \to \infty$. Since the rest terms are all finite, $D^{(n)}(\mathbf{u};\zeta) \to -\infty$ and (4.5.2) holds for any ζ. On the other hand, if $u_1 = 0$, then $\Gamma_4^{(n)} = 0$. Since $\Gamma_{22}^{(n)}$ is positive, by choosing a sufficiently large C_ζ to have $\Gamma_1^{(n)}$ dominated by the rest terms, (4.5.2) holds.

To establish equation (4.5.1), note that if $u_1 \neq 0$, since $\Gamma_4^{(n)} \to -\infty$ regardless of the value of ζ, equation (4.5.1) is satisfied automatically. Therefore, we only need consider the situation where $u_1 = 0$ so that $\Gamma_4^{(n)} = 0$, and the event we are interested

in is

$$\begin{aligned} & \sup_{\zeta\in\Xi}\sup_{\|\mathbf{u}\|=C_\zeta}\left[Q^{(n)}(\theta_0+\mathbf{u};\zeta)-Q^{(n)}(\theta_0;\zeta)\right] \\ = & \sup_{\|\mathbf{u}\|=C}\sup_{\zeta\in\Xi}\left[D^{(n)}(\mathbf{u};\zeta)\right]. \end{aligned}$$

Note that

$$\begin{aligned} \sup_{\zeta\in\Xi} D^{(n)}(\mathbf{u};\zeta) &= \sup_{\zeta\in\Xi}\left\{\Gamma_1^{(n)}(\zeta)-\Gamma_2^{(n)}(\zeta)+\Gamma_3^{(n)}(\zeta)+\Gamma_5^{(n)}(\zeta)\right\} \\ &\leq \sup_{\zeta\in\Xi}\Gamma_1^{(n)}(\zeta)+\sup_{\zeta\in\Xi}[-\Gamma_2^{(n)}(\zeta)]+\sup_{\zeta\in\Xi}\Gamma_3^{(n)}(\zeta)+\sup_{\zeta\in\Xi}\Gamma_5^{(n)}(\zeta) \end{aligned}$$

Similar to the above argument, both $\sup_{\zeta\in\Xi}\Gamma_3^{(n)}(\zeta)$ and $\sup_{\zeta\in\Xi}\Gamma_5^{(n)}(\zeta)$ converges to 0. By assumptions 2.3 and 2.4, $\sup_{\zeta\in\Xi}\Gamma_1^{(n)}(\zeta)$ is $O_p(1)$, and $\sup_{\zeta\in\Xi}[-\Gamma_{22}^{(n)}(\zeta)]$ is negative. Taking a large enough C so that $\sup_{\zeta\in\Xi}[-\Gamma_{22}^{(n)}(\zeta)]$ dominates the rest of terms; then

$$\sup_{\|\mathbf{u}\|=C}\sup_{\zeta\in\Xi} D^{(n)}(\mathbf{u};\zeta)<0.$$

Therefore, for any $\epsilon>0$, there exists some constant C such that equation 4.5.1 holds.

For the second part, consider the event $\left\{\tilde{\beta}^{(n)}\neq 0\right\}$. By the Karush-Kuhn-Tunker optimality conditions, we have

$$\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\log f(y_i;X,\tilde{\theta}^{(n)},\tilde{\zeta}^{(n)})=\lambda_n\hat{w}_n\mathrm{sgn}(\tilde{\beta}^{(n)}).$$

Therefore, $P(\tilde{\beta}^{(n)}\neq 0)\leq P\left\{\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\log f(y_i;X,\tilde{\theta}^{(n)},\tilde{\zeta}^{(n)})=\lambda_n\hat{w}_n\mathrm{sgn}(\tilde{\beta}^{(n)})\right\}$. To study the event $\left\{n^{-1/2}\frac{\partial}{\partial\beta}\log f(y_i;X,\tilde{\theta}^{(n)},\tilde{\zeta}^{(n)})=n^{-1/2}\lambda_n\hat{w}_n\mathrm{sgn}(\tilde{\beta}^{(n)})\right\}$, by Taylor ex-

pansion on the left-had side,

$$\begin{aligned} & n^{-1/2}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\log f(y_i;X,\tilde{\theta}^{(n)},\tilde{\zeta}^{(n)}) \\ = \; & n^{-1/2}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\log f(y_i;\theta_0,\tilde{\zeta}^{(n)}) \\ + \; & n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial\beta^2}\log f(y_i;X,\theta_0,\tilde{\zeta}^{(n)})n^{1/2}(\tilde{\theta}^{(n)}-\theta_0)(1+o_p(1)) \\ \equiv \; & U_1^{(n)}+U_2^{(n)}. \end{aligned}$$

Both $U_1^{(n)}$ and $U_2^{(n)}$ are $O_p(1)$ by assumptions 2.3 and 2.4, and $(\tilde{\theta}^{(n)}-\theta_0)=O_p(n^{-1/2})$ as shown in the first part of the proof. On the right-hand side, $n^{-1/2}\lambda_n\hat{w}_n\text{sgn}(\tilde{\beta}^{(n)}))=\lambda_n n^{(2\alpha\tau-1)/2}|n^{\alpha}\hat{\beta}^{*}|^{-\tau}\to\infty$; as a consequence, $P(\tilde{\beta}^{(n)}\neq 0)\to 0$.

Finally, to show the normality of $\tilde{\gamma}^{(n)}$, since the likelihood $\mathcal{L}^{(n)}(\theta,\zeta)$ is unidentifiable when $\theta=\theta_0$. We cannot use the same approach to establish the normality of the nonzero estimator $\tilde{\gamma}^{(n)}$ as [Fan and Li, 2001]. They conduct the partial Taylor expansion with respect to γ of $\frac{\partial Q^{(n)}(\theta,\zeta)}{\partial\gamma}|_{(\theta,\zeta)=((0,\tilde{\gamma}),\tilde{\zeta})}=0$ as follows.

$$\begin{aligned} 0 \; = \; & \frac{1}{\sqrt{n}}\frac{\partial\mathcal{L}^{(n)}(\theta,\zeta)}{\partial\gamma}|_{\theta=\theta_0,\zeta=\zeta_0} \\ & +\frac{1}{n}\left[\sum_{i=1}^{n}\frac{\partial^2}{\partial\gamma^2}\log f(y_i;\theta,\zeta)|_{\theta=\theta_0,\zeta=\zeta_0}+o_p(n)\right]\sqrt{n}(\tilde{\gamma}^{(n)}-\gamma_0). \end{aligned}$$

Clearly, due to the unidentifiability issue, when $\theta=\theta_0$, there is no existence of ζ_0. Therefore, the limits of $\frac{1}{\sqrt{n}}\frac{\partial\mathcal{L}^{(n)}(\theta,\zeta)}{\partial\gamma}|_{\theta=\theta_0,\zeta=\zeta_0}$ and $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\gamma^2}\log f(y_i;\theta,\zeta)|_{\theta=\theta_0,\zeta=\zeta_0}$ do not exist. Only the point-wise limits of $\frac{1}{\sqrt{n}}\frac{\partial\mathcal{L}^{(n)}(\theta,\zeta)}{\partial\gamma}|_{\theta=\theta_0}$ and $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\gamma^2}\log f(y_i;\theta,\zeta)|_{\theta=\theta_0}$ at any fixed point of $\zeta\in\Xi$ exist.

Since we already show that there exists a large enough N, for any $n\geq N$, $\tilde{\beta}^{(n)}=0$, it can be observed that $Q^{(n)}(\tilde{\beta}^{(n)}=0,\tilde{\gamma}^{(n)};\tilde{\zeta}^{(n)})=l^{(n)}(\tilde{\gamma}^{(n)})$. Therefore, $l^{(n)}(\tilde{\gamma}^{(n)})$ is

no longer involving with β and ζ, and equivalent to $l_0^{(n)}(\tilde{\gamma}^{(n)})$. At this point, $\tilde{\gamma}^{(n)}$ essentially maximizes $l_0^{(n)}(\gamma)$.

By Taylor expansion of $\sum_{i=1}^n \frac{\partial}{\partial\gamma}\log f_0(y_i;\tilde{\gamma}^{(n)})$ around γ_0,

$$0 \;=\; \sum_{i=1}^n \frac{\partial}{\partial\gamma}\log f_0(y_i;\gamma_0) + \sum_{i=1}^n \frac{\partial^2}{\partial\gamma^2}\log f_0(y_i;\gamma_0)(\tilde{\gamma}^{(n)}-\gamma_0)(1+o_p(n)),$$

After rearrangement

$$n^{-1/2}\sum_{i=1}^n \frac{\partial}{\partial\gamma}\log f_0(y_i;\gamma_0) = -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\gamma^2}\log f_0(y_i;\gamma_0)\sqrt{n}(\tilde{\gamma}^{(n)}-\gamma_0) + o_p(1).$$

Therefore, $\sqrt{n}(\tilde{\gamma}^{(n)}-\gamma_0) \to_d N(0, \mathrm{I}(\gamma_0)^{-1})$, where $\mathrm{I}(\gamma_0)$ is the Fisher information matrix corresponding to $\mathcal{L}_0^{(n)}(\gamma_0)$

Proof of Theorem 2

To show $P(\lim_{n\to\infty}(\tilde{\theta}^{(n)},\tilde{\zeta}^{(n)}) = (\theta_1,\zeta_1)) = 1$, we follow the strategy of proof for the Theorem 2 in [Wald, 1949].

Since $(\tilde{\theta},\tilde{\zeta})$ satisfies

$$\frac{Q^{(n)}(\tilde{\theta},\tilde{\zeta})}{Q^{(n)}(\theta_1,\zeta_1)} \geq 1 > 0, \text{ for all } n \text{ and for all } y_1,...,y_n.$$

It is sufficient to show that for any $\epsilon > 0$ the probability is one that all limit points $(\tilde{\theta},\tilde{\zeta})$ of $(\tilde{\theta}^{(n)},\tilde{\zeta}^{(n)})$ satisfying $\|(\tilde{\theta},\tilde{\zeta})-(\theta_1,\zeta_1)\| \leq \epsilon$.

Consider the event that there exists a limit point $(\breve{\theta},\breve{\zeta})$ such that $\|(\breve{\theta},\breve{\zeta})-(\theta_1,\zeta_1)\| > \epsilon$. This implies that $\sup_{\|(\theta,\zeta)-(\theta_1,\zeta_1)\|>\epsilon} Q^{(n)}(\theta,\zeta) \geq Q^{(n)}(\breve{\theta},\breve{\zeta})$ so that

$$\frac{\sup_{\|(\theta,\zeta)-(\theta_1,\zeta_1)\|>\epsilon} Q^{(n)}(\theta,\zeta)}{Q^{(n)}(\theta_1;\zeta_1)} \geq c > 0$$

for infinitely many n. Therefore, it is sufficient to show that this is an event with probability 0.

To establish this, we will show that

$$P\left\{\lim_{n\to\infty}\frac{\sup_{(\theta,\zeta)\in\omega}Q^{(n)}(\theta;\zeta)}{Q^{(n)}(\theta_1;\zeta_1)}=0\right\}=1, \tag{4.5.3}$$

where ω be any closed subset of the parameter space $\Theta\times\Xi$ which does not contain the true parameter point (θ_1,ζ_1).

Clearly, $\sup_{(\theta,\zeta)\in\omega}Q^{(n)}(\theta;\zeta)\leq\sup_{(\theta,\zeta)\in\omega}\mathcal{L}^{(n)}(\theta;\zeta)$. Therefore, equation (4.5.3) is proved if we can show that

$$P\left\{\lim_{n\to\infty}\frac{\sup_{(\theta,\zeta)\in\omega}\mathcal{L}^{(n)}(\theta;\zeta)}{Q^{(n)}(\theta_1;\zeta_1)}=0\right\}=1, \tag{4.5.4}$$

Based on the assumption 3 and the Theorem 1 in [Wald, 1949], the likelihood part of $Q^{(n)}(\theta;\zeta)$ satisfies

$$\begin{aligned}
& P\left\{\lim_{n\to\infty}\frac{\sup_{(\theta,\zeta)\in\omega}\mathcal{L}^{(n)}(\theta;\zeta)}{\mathcal{L}^{(n)}(\theta_1;\zeta_1)}=0\right\}\\
= \;& P\left\{\lim_{n\to\infty}\log\left\{\frac{\sup_{(\theta,\zeta)\in\omega}\mathcal{L}^{(n)}(\theta;\zeta)}{\mathcal{L}^{(n)}(\theta_1;\zeta_1)}\right\}=-\infty\right\}\\
= \;& 1.
\end{aligned}$$

Therefore,

$$
\begin{aligned}
& P\left\{\frac{\sup_{(\theta,\zeta)\in\omega}\mathcal{L}^{(n)}(\theta;\zeta)}{Q^{(n)}(\theta_1;\zeta_1)}=0\right\} \\
= \; & P\left\{\frac{\sup_{(\theta,\zeta)\in\omega}\mathcal{L}^{(n)}(\theta;\zeta)}{\mathcal{L}^{(n)}(\theta_1;\zeta_1)[1-(\lambda_n\hat{w}_n|\beta_1|+\frac{\lambda_n}{\sqrt{n}}I(\beta_1=0)|\zeta_1|)/\mathcal{L}^{(n)}(\theta_1;\zeta_1)]}=0\right\} \\
= \; & P\left\{\lim_{n\to\infty}\left\{\log\left[\frac{\sup_{(\theta,\zeta)\in\omega}\mathcal{L}^{(n)}(\theta;\zeta)}{\mathcal{L}^{(n)}(\theta_1;\zeta_1)}\right]+\log\left[\frac{1}{(1-\delta_n(\theta_1,\zeta_1))}\right]\right\}=-\infty\right\} \\
\approx \; & P\left\{\lim_{n\to\infty}\log\left\{\frac{\sup_{(\theta,\zeta)\in\omega}\mathcal{L}^{(n)}(\theta;\zeta)}{\mathcal{L}^{(n)}(\theta_1;\zeta_1)}\right\}=-\infty\right\}=1 \text{ since}
\end{aligned}
$$

$$
\begin{aligned}
& 1-\delta_n(\theta_1,\zeta_1) \\
= \; & 1-\left\{\left[\lambda_n\hat{w}_n|\beta_1|+\frac{\lambda_n}{\sqrt{n}}I(\beta_1=0)|\zeta_1|\right]/\mathcal{L}^{(n)}(\theta_1;\zeta_1)\right\} \\
= \; & 1-\left\{\lambda_n\hat{w}_n|\beta_1|/\mathcal{L}^{(n)}(\theta_1;\zeta_1)\right\}\to 1
\end{aligned}
$$

by the assumptions 1.2 and 3.5, $\hat{w}_n \to |c_\beta|^{-\tau}$ and $\lambda_n/\mathcal{L}^{(n)}(\theta_1;\zeta_1)\to 0$.

Therefore, with probability 1, $\lim_{n\to\infty}(\tilde{\theta}^{(n)},\tilde{\zeta}^{(n)})=(\theta_1,\zeta_1)$.

Notations

Following [He and Severini, 2010], let $\Lambda_1 = n_1/n$ be the true percentage of sample with the observed w_i less than or equal to ζ for a sample with size n, and Λ_1^0 be the constant fraction value as $n\to\infty$. Additionally, let $\tilde{n}_1^{(n)}$ denote the estimate of the percentage of sample with the observed w_i less than or equal to ζ, $\tilde{n}_1^{(n)}=\sum_{i=1}^n I(w_i\le\tilde{\zeta}^{(n)})$. The expected information matrix is given by $I(\theta)=E[-\frac{\partial^2}{\partial\theta^2}\acute{l}]$, where $\theta\in\Theta_1$.

Assumption 5

5.1 $f_1(y_i;\gamma) \neq f_2(y_i;\gamma,\beta)$ on a set of non-zero measure given $\beta \neq 0$.

5.2 $\acute{l}$ is third-order continuously differentiable with respect to θ.

5.3 The expectations of the first and second order derivatives of $\acute{l}$ with respect to θ exist for θ in its parameter space.

Proof of Theorem 4

The penalized likelihood function is differentiable at $\beta \neq 0$,

$$\frac{\partial}{\partial\theta}Q^{(n)}(\theta,\zeta) = \frac{\partial}{\partial\theta}l_n(\theta,\zeta) - \lambda_n\hat{w}_n\text{sgn}(\beta).$$

Since $(\tilde{\theta}^{(n)},\tilde{\zeta}^{(n)})$ maximizes $Q^{(n)}(\theta;\zeta)$,

$$\begin{aligned} 0 = \frac{\partial}{\partial\theta}Q^{(n)}(\tilde{\theta}^{(n)},\tilde{\zeta}^{(n)}) &= \frac{\partial}{\partial\theta}l_n(\tilde{\theta}^{(n)},\tilde{\zeta}^{(n)}) - \lambda_n\hat{w}_n\text{sgn}(\tilde{\beta}^{(n)}) \\ &= \frac{\partial}{\partial\theta}l_n(\tilde{\theta}^{(n)},\tilde{n}_1^{(n)}) - \lambda_n\hat{w}_n\text{sgn}(\tilde{\beta}^{(n)}). \end{aligned}$$

Then expand $\frac{\partial}{\partial\theta}l_n(\tilde{\theta}^{(n)},\tilde{n}_1^{(n)})$ around $\frac{\partial}{\partial\theta}l_n(\theta,\tilde{n}_1^{(n)})$,

$$\sqrt{n}(\tilde{\theta}^{(n)}-\theta) = \left[-\frac{1}{n}\frac{\partial^2}{\partial\theta^2}l_n(\theta,\tilde{n}_1^{(n)}) + o_p(1)\right]^{-1}\frac{1}{\sqrt{n}}\left[\frac{\partial}{\partial\theta}l_n(\theta,\tilde{n}_1^{(n)}) - \lambda_n\hat{w}_n\text{sgn}(\tilde{\beta}^{(n)})\right].$$

Since $n^{-1/2}\lambda_n\hat{w}_n\text{sgn}(\tilde{\beta}^{(n)}) \to 0$, only $\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}l_n(\theta,\tilde{n}_1^{(n)})$ plays a critical role in determining the limiting distribution of $\sqrt{n}(\tilde{\theta}^{(n)}-\theta)$. This makes the proof for $\sqrt{n}(\tilde{\theta}^{(n)}-\theta) \to_d N(0,\text{I}(\theta)^{-1})$ be a special case of Theorem 2.2 and 2.3 in [He and Severini, 2010].

First, we study the behavior of $\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}l_n(\theta,\tilde{n}_1^{(n)})$. Consider

$$\frac{1}{\sqrt{n}}\left[\frac{\partial}{\partial\theta}l_n(\theta,\tilde{n}_1^{(n)}) - \frac{\partial}{\partial\theta}l_n(\theta,n_1)\right] + \frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}l_n(\theta,n_1),$$

and study the limiting behavior of $\frac{1}{\sqrt{n}}\left[\frac{\partial}{\partial\theta}l_n(\theta,\tilde{n}_1^{(n)})-\frac{\partial}{\partial\theta}l_n(\theta,n_1)\right]$:

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\left[\frac{\partial}{\partial\theta}l_n(\theta,\tilde{n}_1^{(n)})-\frac{\partial}{\partial\theta}l_n(\theta,n_1)\right]\\
=\;&\frac{1}{\sqrt{n}}\left\{I(\tilde{n}_1^{(n)}\geq n_1)\left[\sum_{i=1}^{\tilde{n}_1^{(n)}}\frac{\partial}{\partial\theta}\log f_1(y_i;\gamma)+\sum_{i=\tilde{n}_1^{(n)}+1}^{n}\frac{\partial}{\partial\theta}\log f_2(y_i;\gamma,\beta)\right]\right\}\\
=\;&\frac{1}{\sqrt{n}}\left\{I(\tilde{n}_1^{(n)}\geq n_1)\sum_{i=n_1+1}^{\tilde{n}_1^{(n)}}\left[\frac{\partial}{\partial\theta}\log f_1(y_i;\gamma)-\frac{\partial}{\partial\theta}\log f_2(y_i;\gamma,\beta)\right]+\right.\\
&\left.I(\tilde{n}_1^{(n)}<n_1)\sum_{i=\tilde{n}_1^{(n)}+1}^{n_1-1}\left[-\frac{\partial}{\partial\theta}\log f_1(y_i;\gamma)+\frac{\partial}{\partial\theta}\log f_2(y_i;\gamma,\beta)\right]\right\}
\end{aligned}$$

By the consistency of $\tilde{\zeta}^{(n)}$ in the Theorem 2 and the Theorem 2.2 in [He and Severini, 2010], $\frac{1}{\sqrt{n}}\left[\frac{\partial}{\partial\theta}l_n(\theta,\tilde{n}_1^{(n)})-\frac{\partial}{\partial\theta}l_n(\theta,n_1)\right]\to 0$. Using similar argument, it can be shown that $-\frac{1}{n}\frac{\partial^2}{\partial\theta^2}l_n(\theta,\tilde{n}_1^{(n)})\to\bar{I}(\theta)$ as $n\to\infty$. Furthermore, since $\frac{\partial}{\partial\theta}l^0(\theta)\to_d N(0,\bar{I}(\theta))$, $\sqrt{n}(\tilde{\theta}^{(n)}-\theta)\to_d N(0,\mathrm{I}(\theta)^{-1})$, where $\theta\in\Theta_1$.

CHAPTER 5: Conclusion

To summarize, the first paper investigates the applicability of the penalty functions in challenging high dimensional settings such as genomic studies. We conducted a theoretical analysis on the roles of tuning parameters with respect to the dimension of the problem and minimum effect size. The results suggest that the derivatives of the penalty function around 0 and the minimum effect size are two important quantities to be considered. A good performance of the penalized estimation requires that these two quantities be asymptotically different. Among the four penalties discussed in this paper, tuning one regularization parameter is sufficient to exploit the advantages of SCAD. In contrast, MCP, SICA and Log's performances can be significantly improved if two instead of one (λ) regularization parameter is tuned. These theoretical conclusions are well supported in our empirical studies. In our simulations, we also observe that a penalized estimation using SICA or Log appears to be computationally more efficient than using MCP. The good performance of tuning two regularization parameters comes with the cost of added computational time. In real data analysis, one needs to judge the difficulty of the penalization problem in terms of effect size and dimensionality in order to choose whether one or two regularization parameters are needed, and the theoretical results of this paper can guide such choices. These theoretical results are based on the sufficient conditions of the weak oracle property, and thus they could be refined if the sufficient and necessary conditions of the weak oracle property are available, though deriving such conditions itself is a very challenging task.

Based on the results in the first paper, we designed a new method, BipLog, for the bi-level selection of genomic features related to cancer drug sensitivity. BipLog can select the covariates shared by a group of response variables as well as the covariates that are associated with one or a few of the response variables. The application of BipLog to real-data analysis reveals many interesting results. This is partly due to the strong effect size in the data. In contrast to genome-wide association studies where a genetic variant may explain only a few percentage of the variation in the trait of interest, the genomic features measured in tumor tissues have a strong influence on the cancer progression and its response to drug treatment. This makes cancer genomic studies one of a few areas where statistical methods can make a major contribution in the near future to disease prevention and treatment.

For the third paper, we constructed an estimation procedure for the models where a certain parameter values such as $\beta = 0$ will cause an identifiability issue. We utilize the idea of penalization estimation procedure and apply adaptive Lasso penalty. In addition, we established the asymptotic results for our penalized estimation procedure, and evaluated its performances in the simulation study and real data analysis.

BIBLIOGRAPHY

Andrews, D. and Ploberger, W. (1995). Admissibility of the likelihood ratio test when a nuisance parameter is present only under the alternative. *The Annals of Statistics*, pages 1609–1629.

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, pages 821–856.

Bacon, D. and Watts, D. (1971). Estimating the transition between two intersecting straight lines. *Biometrika*, 58(3):525–534.

Bai, J. (1997). Estimation of a change point in multiple regression models. *Review of Economics and Statistics*, 79(4):551–563.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607.

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and its Interface*, 2(3):369–380.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232–253.

Broman, K. W., Wu, H., Sen, Ś., and Churchill, G. A. (2003). R/qtl: Qtl mapping in experimental crosses. *Bioinformatics*, 19(7):889–890.

Browning, S. and Browning, B. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097.

Bühlmann, P. and Mandozzi, J. (2012). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Computational Statistics*, pages 1–24.

Caponigro, G. and Sellers, W. R. (2011). Advances in the preclinical testing of cancer therapeutic hypotheses. *Nature Reviews Drug Discovery*, 10(3):179–187.

Chen, H.-M., Schmeichel, K. L., Mian, I. S., Lelievre, S., Petersen, O. W., and Bissell, M. J. (2000). Azu-1: A candidate breast tumor suppressor and biomarker for tumor progression. *Molecular Biology of the Cell*, 11(4):1357–1367.

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

Chen, J. and Chen, Z. (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, 22(2):555.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., et al. (2011). Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(suppl 1):D691–D697.

Cui, Z., Han, F., Peng, X., Chen, X., Luan, C., Han, R., Xu, W., Guo, X., et al. (2011). YES-associated protein 1 promotes adenocarcinoma growth and metastasis through activation of the receptor tyrosine kinase axl. *International Journal of Immunopathology and Pharmacology*, 25(4):989–1001.

Davies, R. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 74(1):33–43.

Davtes, R. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 64(2):247–254.

De Palma, M. and Hanahan, D. (2012). The biology of personalized cancer medicine: Facing individual complexities underlying hallmark capabilities. *Molecular Oncology*, 6(2):111–127.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55(4):997–1004.

Fan, J. (1997). Comments on 'wavelets in statistics: A review' by a. antoniadis. *Statistical Methods & Applications*, 6(2):131–138.

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

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions on*, 57(8):5467–5484.

Friedman, J. (2008). Fast sparse regression and classification. Technical report, Stanford University.

Garnett, M., Edelman, E., Heidorn, S., Greenman, C., Dastur, A., Lau, K., Greninger, P., Thompson, I., Luo, X., Soares, J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674.

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica: Journal of the econometric society*, pages 413–430.

He, H. and Severini, T. (2010). Asymptotic properties of maximum likelihood estimators in models with multiple change points. *Bernoulli*, 16(3):759–779.

Hoelder, S., Clarke, P. A., and Workman, P. (2012). Discovery of small molecule cancer drugs: successes, challenges and opportunities. *Molecular oncology*, 6(2):155–176.

Hong, J., Peng, D., Chen, Z., Sehdev, V., and Belkhiri, A. (2013). Abl regulation by axl promotes cisplatin resistance in esophageal cancer. *Cancer Research*, 73(1):331–340.

Huang, D., Sherman, B., and Lempicki, R. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57.

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high dimensional models. *arXiv preprint arXiv:1204.6491*.

Huang, J., Ma, S., Xie, H., and Zhang, C. (2009). A group bridge approach for variable selection. *Biometrika*, 96(2):339–355.

Joseph, P., Xie, T., Xu, Y., Jaiswal, A. K., et al. (1994). Nad (p) h: quinone oxidoreductase1 (dt-diaphorase): expression, regulation, and role in cancer. *Oncology research*, 6(10-11):525.

Julious, S. (2001). Inference and estimation in a changepoint regression problem. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 50(1):51–61.

Lee, S., Kwon, H., Jeong, K., Pak, Y., et al. (2011). Regulation of cancer cell proliferation by caveolin-2 down-regulation and re-expression. *International Journal of Oncology*, 38(5):1395.

Liu, Y. and Wu, Y. (2007). Variable selection via a combination of the l 0 and l 1 penalties. *Journal of Computational and Graphical Statistics*, 16(4):782–798.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, 37(6A):3498–3528.

Mazumder, R., Friedman, J. H., and Hastie, T. (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495).

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman & Hall/CRC.

Mei, Y., Wang, Z., Zhang, L., Zhang, Y., Li, X., Liu, H., Ye, J., and You, H. (2012). Regulation of neuroblastoma differentiation by forkhead transcription factors foxo1/3/4 through the receptor tyrosine kinase pdgfra. *Proceedings of the National Academy of Sciences*, 109(13):4898–4903.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Muggeo, V. M. (2003). Estimating regression models with unknown break-points. *Statistics in medicine*, 22(19):3055–3071.

Oberst, M. D., Beberman, S. J., Zhao, L., Yin, J. J., Ward, Y., and Kelly, K. (2008). Tdag51 is an erk signaling target that opposes erk-mediated hme16c mammary epithelial cell transformation. *BMC Cancer*, 8(1):189.

Ough, M., Lewis, A., Bey, E. A., Gao, J., Ritchie, J. M., Bornmann, W., Boothman, D. A., Oberley, L. W., and Cullen, J. J. (2005). Efficacy of beta-lapachone in pancreatic cancer treatment: exploiting the novel, therapeutic target nqo1. *Cancer biology & therapy*, 4(1):95–102.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Planque, N. and Perbal, B. (2003). A structural approach to the role of ccn (cyr61/ctgf/nov) proteins in tumourigenesis. *Cancer Cell International*, 3(1):15.

Rosenbluh, J., Nijhawan, D., Cox, A. G., Li, X., Neal, J. T., Schafer, E. J., Zack, T. I., Wang, X., Tsherniak, A., Schinzel, A. C., et al. (2012). β-catenin-driven cancers require a yap1 transcriptional complex for survival and tumorigenesis. *Cell*, 151:1457–1473.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shi, J., Levinson, D., Duan, J., Sanders, A., Zheng, Y., Peâ, I., et al. (2009). Common variants on chromosome 6p22. 1 are associated with schizophrenia. *Nature*, 460(7256):753–757.

Sun, W., Ibrahim, J., and Zou, F. (2010). Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. *Genetics*, 185(1):349–359.

Tian, F., Byfield, S. D., Parks, W. T., Yoo, S., Felici, A., Tang, B., Piek, E., Wakefield, L. M., and Roberts, A. B. (2003). Reduction in smad2/3 signaling enhances tumorigenesis but suppresses metastasis of breast cancer cell lines. *Cancer Research*, 63(23):8284–8292.

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

Tsai, M.-S., Bogart, D. F., Castaneda, J. M., Li, P., and Lupu, R. (2002). Cyr61 promotes breast tumorigenesis and cancer progression. *Oncogene*, 21:8178–8185.

Usmani, S. Z., Bona, R., and Li, Z. (2009). 17 aag for hsp90 inhibition in cancer-from bench to bedside. *Current molecular medicine*, 9(5):654–664.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601.

Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis*, 52(12):5277–5286.

Wang, X., Su, L., and Ou, Q. (2012). YES-associated protein promotes tumour development in luminal epithelial derived breast cancer. *European Journal of Cancer*, 48(8):1227–1234.

Wright, F., Sullivan, P., Brooks, A., Zou, F., Sun, W., Xia, K., Madar, V., Abdellaoui, A., Batista, S., Butler, C., Chen, G., Chen, T., W., C., et al. (2014). Heritability and Genomics of Gene Expression In Peripheral Blood. *Nature Genetics*, in press.

Xie, D., Nakachi, K., Wang, H., Elashoff, R., and Koeffler, H. P. (2001). Elevated levels of connective tissue growth factor, wisp-1, and cyr61 in primary breast cancers associated with more advanced features. *Cancer Research*, 61(24):8917–8923.

Yap, T. A. and Workman, P. (2012). Exploiting the cancer genome: strategies for the discovery and clinical development of targeted molecular therapeutics. *Annual review of pharmacology and toxicology*, 52:549–573.

Yi, N. and Xu, S. (2008). Bayesian LASSO for Quantitative Trait Loci Mapping. *Genetics*, 179:1045–1055.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

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

Zhang, C. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

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

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533.