Robust variable selection with stability selection

by

Gongshun Yang

M.S., Kansas State University, 2020

A Report

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Approved by:

Major Professor
Cen Wu

Copyright

Abstract

Heterogeneity is the hallmark of cancer. In the presence of high-dimensional cancer genomics features, robust variable selection is critical to accommodate the heavy-tailed distributions and outliers rising due to cancer heterogeneity when selecting important omics features associated with the disease phenotype. However, it is challenging to appropriately choose the tuning parameters that can maximize the performance of the regularization approaches given the data heterogeneity. In published studies, assisted tuning has been proposed as an effective strategy to conduct regularized variable selection by overcoming the obstacle of selecting the optimal tuning parameters for non-robust variable selection methods such as LASSO. Nevertheless, we have shown in this study that the heavy-duty nature of this strategy makes it infeasible for omics feature selection when the data heterogeneity exists. Such a limitation has motivated us to develop the robust variable selection with stability selection that significantly outperforms a myriad of variable selection methods. In extensive simulation studies, we have demonstrated the advantage of the proposed method over alternative methods in terms of multiple criteria that are widely used in literature to determine the performance of machine learning methods. Furthermore, all the methods under comparison have been applied to the breast cancer and skin cancer data from TCGA. The proposed method has also shown superior performances in the case studies.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest appreciation to my major advisor, Dr. Cen Wu, for his invaluable guidance, unwavering support, and insightful feedback throughout the course of this journey. Dr. Wu's expertise in the field and commitment to mentorship have been pivotal to my personal growth and the success of my research project. His encouragement and critical input have greatly shaped my academic and professional journey.

I am also immensely grateful to my committee members, Dr. Perla Reyes and Dr. Weixin Song, for their constructive critiques, enlightening discussions, and the significant time they dedicated to reviewing my work. The diverse perspectives and expert feedback provided by Dr. Reyes and Dr. Song have significantly enriched the depth and quality of this research. Their guidance has been instrumental in refining my ideas and enhancing the overall outcome of my study.

Furthermore, I extend my heartfelt gratitude to my family for their endless love, encouragement, and sacrifices. My family's unwavering belief in my abilities and constant support at every step of this academic endeavor have been the bedrock of my resilience and determination. It is their faith in me that has been my greatest source of strength and motivation, enabling me to pursue my academic goals with vigor and passion.

In summary, the support and guidance I have received from my advisors, committee members, and family have been invaluable. Their collective contributions have not only facilitated the successful completion of my project but have also significantly contributed to my personal and academic development

Chapter 1

Introduction

Linear regression is a powerful statistical tool in many fields for investigating the associations between the response variable and a set of predictors through the following linear model: $Y = X\beta + \epsilon$, where Y is the response variable, X is the matrix of predictor variables (usually including the intercept), β is the vector of regression coefficients, and ϵ is the vector of model errors. Fitting a least square regression model seeks to minimize the sum of squared errors: $\min_\beta ||Y - X\beta||_2^2$. However, such a model fitting method is sensitive to the presence of outliers and influential observations that will lead to biased and unreliable estimates of regression parameters. Furthermore, the independent and identical normal distributional assumption on model errors no longer holds even under a single outlying observation in the data. In practice, the irregular data patterns have been widely observed in bioinformatics studies due to the heterogeneity of cancer and other complex diseases[1;2].

Addressing the non-robustness of the classical linear regression analysis has motivated the development of robust statistical methods[3], such as the Least Absolute Deviation (LAD) regression[4]. The LAD regression, in contrast to least square regression, minimizes the sum of absolute errors: $\min_\beta ||Y - X\beta||_1$. The robustness of the LAD loss lies in the L1 form of the objective function, which downweighs the residuals corresponding to extreme observations. The lease square loss, on the other hand, results in the amplification of the residuals with respect to the outliers because of the squared form. Furthermore, while the heteroscedas-

ticity in the data violates the constant variance assumption in least square regression and jeopardizes the validity of inferential procedures, the robust LAD regression is less sensitive to heteroscedasticity and shows advantage not only in estimation but also in statistical inference[4].

In large scale bioinformatics studies, regression analysis has been constantly challenged by the high-dimensionality of data given the "large dimensionality, small sample size" nature of the multi-omics data[1;2]. When dealing with high-dimensional data, traditional regression methods suffer from the curse of dimensionality and are not capable of yielding satisfactory estimation and uncertainty quantification results[5;6]. To combat these challenges, various regularized variable selection techniques have been developed. For example, the LASSO (Least Absolute Shrinkage and Selection Operator) is a regularized least square model which is based on the sum of least square loss and a L1 penalty function[7]. The penalty term in the regularized loss function facilitates the shrinkage of regression coefficients with respect to the irrelevant predictors towards zero, which can reduce overfitting of the models under high-dimensional omics data and potentially enhance the predictive performance. The family of LASSO methods have later been extended to the Bayesian framework, including the Bayesian LASSO[8;9]. A prominent characteristic of the fully Bayesian methods is that the exact statistical inference can be enabled even with a small sample size. Therefore, quantifying uncertainty of findings are feasible for highly structured omics features[10;11].

To simultaneously handle the high-dimensionality and heterogeneity in the omics data, robust variable selection methods have been extensively developed[1;12]. The defining feature of robust regularization methods lies in the objective functions in the form of "unregularized robust loss + penalty function". For example, the LAD–LASSO integrates the L1 penalty with the LAD loss, which is therefore more robust than LASSO[13]. Although frequentist robust variable selection methods have already been widely examined in published literature over the past two decades, their counterpart in the Bayesian framework, especially those under structured sparsity such as in the gene-environment interaction studies, have just been explored in recent studies[14–16].

The validity of variable selection procedures hinges on uncertainty quantification pro-

cedures. Without the inferential guarantee, regularization approaches are likely to yield false findings. A well known example is the graphical LASSO[17;18]. A trustworthy analysis using graphical LASSO on genoimics data demands the presentation of multiple graphical structures along the solution path, as suggested by Friedman et al.[17]. Choosing a single graphical structure as the important finding for downstream analysis may result in misleading conclusions. As shown by Meinshausen and Bühlmann[19], sparse graphical structures obtained by a direct application of graphical LASSO are generally not reproducible. On the other hand, the stability selection-assisted regularization not only stabilizes the shrinkage estimates using the graphical structure but also provides control on error rates of falsely discovered signals. Empirically, this approach subsamples the data multiple times and selects variables that consistently have nonzero estimates across subsamples beyond certain cutoff[19;20]. In addition, knockoff filters are also among the variable selection methods that control the False Discovery Rate (FDR)[21;22]. Next, we will briefly review published studies on variable selection in the succeeding sections.

1.1 Regularized variable selection

In high-throughput cancer studies, it is common to encounter datasets with a large scale of omics features, many of which are irrelevant to the phenotypic trait of interest. Regularized variable selection is critical to address the scientific question on how to identify a small subset of features that is potentially associated with the disease phenotype of the study[1]. The principle behind penalized variable selection is the incorporation of a penalty term to the loss function for the model fitting, which imposes a constraint on the coefficients during model fitting. It effectively reduces model complexity and the risk of overfitting. The general form of the penalized objective function is defined as follows:

$$\min_{\beta} \left\{ L(\beta; Y, X) + \lambda \mathrm{Pen}(\beta) \right\}, \tag{1.1}$$

where $L(\beta; Y, X)$ represents the loss function, such as the least square loss or quantile check loss[1]. Pen(β) denotes the penalty term, and λ is the tuning parameter that controls the bias-variance trade-off[5].

With the formulation specified in (1.1), we can consider the penalization as an optimization problem with a regularized loss function. Two widely applied penalization methods are LASSO (Least Absolute Shrinkage and Selection Operator)[7] and Ridge regression[23;24]. Both methods incorporate a penalty term in the loss function, but in different ways. LASSO employs an L1 penalty which can shrink all regression coefficients towards zero. If the tuning parameter is chosen properly, a subset of predictors will be excluded from the model. Ridge regression, on the other hand, adopts an squared L2 penalty for regularized estimation, resulting in a model with small, but non-zero coefficients. This property implies that while ridge regression effectively shrinks the magnitude of coefficients, it does not perform variable selection, since it cannot shrink coefficients to zero. Besides, elastic net is a method that combines the L1 penalty and Ridge penalties, which accommodate the selection of features with strong correlations[25]. In general, the formulation in (1.1) leads to non-robust variable selection methods if unpenalized loss $L(\cdot)$ is a least square function. In Table 1.1, we provide a partial list of non-robust penalization methods from published studies. Among the tabulated methods, adaptive LASSO improves over LASSO by assigning individual weights to each coefficient in the L1 penalty term. This weighted L1 penalty promotes adaptive sparsity, resulting in a sparse model with better identification performance over LASSO[26].

In addition to the aforementioned LASSO family of methods with convex penalty functions, Fan and Li (2001)[27] have developed the non-convex Smoothly Clipped Absolute Deviation (SCAD) penalty. SCAD differs from LASSO by using a non-convex penalty function that ensures coefficients with large magnitude are less penalized, while those with small absolute values are reduced to zero, which avoids unnecessary penalization on regression coefficients corresponding to more important predictors and enjoys the unbiasedness property. On the other hand, the Minimax Concave Penalty (MCP) has been proposed as a penalty function with control on the amount of concavity, which also enjoys the advantage of SCAD[28].

Table 1.1: *Partial list of non-robust regularized variable selection methods*

Method	Loss Function	Penalty Function
LASSO[7]	Least Square Loss	$\text{Pen}(\beta) = \lambda\|\|\beta\|\|_1$
Ridge[23]	Least Square Loss	$\text{Pen}(\beta) = \lambda^2\|\|\beta\|\|_2^2$
Elastic Net[25]	Least Square Loss	$\text{Pen}(\beta) = \lambda((1-a)/2\|\|\beta\|\|_2^2 + a\|\|\beta\|\|_1)$, for $0 \le a \le 1$
SCAD[27]	Least Square Loss	$\text{Pen}(\beta_j; a) = \begin{cases} \lambda\|\beta_j\| & \text{if } \|\beta_j\| \le \lambda, \\ -\frac{\beta_j^2 - 2a\lambda\|\beta_j\| + \lambda^2}{2(a-1)} & \text{if } \lambda < \|\beta_j\| \le a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } \|\beta_j\| > a\lambda \end{cases}$
MCP[28]	Least Square Loss	$\text{Pen}(\beta_j; a) = \begin{cases} \lambda(\|\beta_j\|) - \frac{\beta_j^2}{2a} & \text{if } \|\beta_j\| \le a\lambda, \\ \frac{a\lambda^2}{2} & \text{if } \|\beta_j\| > a\lambda \end{cases}$
Adaptive LASSO[26]	Least Square Loss	$\text{Pen}(\beta_j) = \lambda w_j\|\beta_j\|$, for $w_j > 0$

Meanwhile, robust variable selection methods have been extensively developed for high-dimensional heterogeneous data, which requires a robust unregularized loss function $L(\beta; Y, X)$ under the specification in (1.1). Table 1.2 shows an incomplete list of published robust regularization methods. Among them, the LAD LASSO incorporates the robustness of LAD regression with the variable selection induced by the L1 norm on regression parameters. Fitting LAD-LASSO model minimizes the sum of absolute residuals and shrinks all coefficients toward zero to achieve variable selection. Therefore, LAD LASSO is particularly useful when dealing with outliers or heavy-tailed error distributions in the high-dimensionality setting[13]. Besides, the Huber Elastic-net adopts the Huber loss to account for robustness, and conducts regularization through the elastic net penalty[29]. The Huber loss is defined as follows for the ith observation:

$$L_\delta(\beta; Y_i, X_i) = \begin{cases} \frac{1}{2}(Y_i - X_i\beta)^2, & \text{if } |Y_i - X_i\beta| \le \delta, \\ \delta\left(|Y_i - X_i\beta| - \frac{1}{2}\delta\right) & \text{otherwise,} \end{cases} . \tag{1.2}$$

which is a quadratic loss when residuals are close to zero, and becomes a least absolute loss on residuals corresponding to outliers. The positive constant δ is a tuning parameter to be determined in a data-driven manner. The structure of Huber loss ensures the combination of analytically tractable square loss under least squares and the robustness of least absolute loss against outlying observations. To achieve sparse identification, the elastic net penalty has been adopted to select features with strong correlations. When Huber loss integrates with the elastic net penalty, it forms a powerful approach for regression in high-dimensional data, balancing robustness against outliers with efficient coefficient estimation and variable selection. In addition, Lambert-Lacroix et al.[30] have developed Huber's loss with the adaptive LASSO resistant to heavy-tailed errors and outliers while performing variable selection. More recently, Kepplinger[31] has established the oracle property of adaptive PENSE under the robust S-loss and adaptive elastic net.

In high-dimensional bioinformatics studies, quantile regression based variable selection methods have been broadly explored[1]. For the ith observation, the quantile loss is defined as:

$$L_\tau(\beta; Y_i, X_i) = \begin{cases} \tau(Y_i - X_i\beta), & \text{if } Y_i - X_i\beta > 0, \\ (1-\tau)(X_i\beta - Y_i), & \text{otherwise,} \end{cases} \tag{1.3}$$

which assigns a weight of τ to positive residuals and $1-\tau$ to negative residuals, targeting the τ-th quantile of the response variable. Alongside this, the L1-norm regularization reduces some coefficients to zero, enhancing model sparsity. This method is especially effective in supervised data settings with heterogeneous response variables, Y, providing robust variable selection and sparse coefficient estimation in diverse conditions. In literature, Li and Zhu[32] have developed an efficient solution-path algorithm for quantile regression with L1-norm regularization, and shown that the entire solution path is piecewise linear. Furthermore, Peng and Wang[33] have developed the coordinate descent algorithm for penalized quantile regression with nonconvex penalty functions such as MCP and SCAD, where each coordinate-wise update in the coordinate descent framework proceeds based on the weighted median regression. Such a strategy has been further extended to robust network based regularization

and variable selection to handle more complicated sparse network omics data[34;35].

A common limitation in robust variable selection methods is the lack of statistical inferential procedures to quantify uncertainty. While the oracle properties of sparse estimator, including the consistency in parameter estimation and asymptotic distribution of the regularized estimator, have been extensively established in non-robust penalization methods[26;27], they are not available in many robust regularization methods, partially due to the challenging in deriving the theoretical results under the non-smooth robust loss functions. Such a gap has further driven the development of robust Bayesian variable selection methods to guarantee the validity of exact statistical inference[14–16].Beyond the methods under continuous response variable, or disease phenotype, discussed here, robust variable selection methods have also been proposed under other types of responses such as the multivariate responses[36], survival outcomes[37;38] and longitudinal traits[39–42].

In general, penalized variable selection methods have been widely adopted. They enhance the prediction accuracy by reducing overfitting, simplify models by removing irrelevant variables, and can handle multicollinearity issues and high-dimensional data where the number of predictors is larger than the number of observations. Despite their strengths, penalized variable selection methods are facing many limitations. The choice of the penalty term, controlled by a tuning parameter λ, is important in determining the degree of regularization applied to the model's coefficients. Selecting an inappropriate λ value may result in overfitting or underfitting, thus affecting the performance of the model. Additionally, these methods may struggle to reliably identify and select the correct variables that truly impact the response variable, especially in the presence of correlated predictors.

Table 1.2: *Partial list of robust regularized variable selection methods*

Method	Loss Function	Penalty Function
Quantile LASSO[43]	Check Loss	$\text{Pen}(\beta) = \lambda\|\beta\|_1$
LAD–LASSO[13]	LAD Loss	$\text{Pen}(\beta) = \lambda\|\beta\|_1$
Huber Elastic-Net[30]	Huber Loss	$\text{Pen}(\beta, a) = \lambda((1-a)/2\|\|\beta\|\|_2^2 + a\|\|\beta\|\|_1)$ for $0 \leq a \leq 1$
QICD-SCAD[33]	Check Loss	$\text{Pen}(\beta_j; a) = \begin{cases} \lambda\|\beta_j\| & \text{if } \|\beta_j\| \leq \lambda, \\ -\frac{\beta_j^2 - 2a\lambda\|\beta_j\| + \lambda^2}{2(a-1)} & \text{if } \lambda < \|\beta_j\| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } \|\beta_j\| > a\lambda \end{cases}$
QICD-MCP[33]	Check Loss	$\text{Pen}(\beta_j; a) = \begin{cases} \lambda(\|\beta_j\|) - \frac{\beta_j^2}{2a} & \text{if } \|\beta_j\| \leq a\lambda, \\ \frac{a\lambda^2}{2} & \text{if } \|\beta_j\| > a\lambda \end{cases}$
Huber[29] Adaptive-LASSO	Huber Loss	$\text{Pen}(\beta_j) = \lambda w_j \|\beta_j\|$, where $w_j > 0$
LAD[34] - Network	LAD Loss	$\text{Pen}(\beta_j) = \sum_{m=1}^{p} \rho_{\lambda_1,\gamma}(\|\beta_m\|) +$ $\lambda_2 \sum_{1 \leq m < k \leq p} \|a_{mk}\| \|\beta_m - \text{sgn}(a_{mk})\beta_k\|$, where $\rho_{\lambda_1,\gamma}(\|\beta_m\|)$ is the MCP penalty.

1.2 Controlling selection of false variables

Within the frequentist framework, quantifying uncertainty in high-dimensional settings is generally more challenging than shrinkage estimation itself, and thus much less developed. Nevertheless, multiple methods of conducting high-dimensional inference by providing confidence intervals (regions) and p-values have been developed[44]. Representative examples include the post-selection inference[45;46], the de-biased LASSO[47], and residual-type bootstrap methods[48;49].

Furthermore, to provide control on the expected number of false positive variables, Meinshausen and Bühlmann have developed the stability selection procedure[19], which has been partially motivated by the irreproducibility issues in high-dimensional regression problems regarding complicated graphical structures. Meinshausen and Bühlmann have shown that when applying graphical LASSO for variable selection, even a small perturbation in the magnitude of tuning parameters can dramatically change the selected graphical structure. To tackle this issue, the stability selection procedure has been developed to control the expected number of false selections while choosing the regularization parameter, leading to significantly improved stability on the selected model.

Specifically, the procedure operates by repeatedly sub-sampling the original data, applying a regularization method such as LASSO or graphical LASSO to each subsample, and then summarizing the selection frequency of each predictor across all the subsamples. Those predictors that are selected with high frequency across many subsamples are deemed truly important, while those selected less frequently are likely the result of noise or chance correlations in the data. The main appeal of stability selection lies in its explicit control of error rates on all selected variables. Empirically, by setting a high frequency threshold for variable selection (e.g., a 90% frequency indicates the variable selected in 90% of subsamples), the method ensures that only consistently important predictors across subsamples are selected, which in turn controls the rate of false discoveries. Another key parameter in stability selection is the sub-sampling proportion, which determines the size of the subsamples. A common choice is to set this proportion to about 0.5, meaning that each subsample is about half the

size of the original dataset. Accordingly, in terms of computational cost, stability selection is approximately 3 times more expensive than 10 fold cross validation when fitting LASSO models[19].

In addition to stability selection, the knockoff methods have also been proposed to control the error rates of selecting false variables[21;50]. Wu et al.[51] have proposed the strategy to generate pseudo variables as the linear combinations of the original predictors. Simply put, they randomly permute the rows (w.r.t to samples) of original predictors X to X^* by following standard normal distribution, $N(0,1)$, under independent and identical distributional assumption. By further reducing the false discover rate, they have developed a procedure by replacing pseudo matrix X^* using $(I - H_X)X^*$, where $H_X = X(X^TX)^{-1}X^T$. By subtracting Hat Matrix H_X from identity matrix I, the pseudo matrix X^* reduced the risks for selecting real important variables.

In further development, Barber et al.[21] have developed a variable selection technique, named as 'knockoffs', by using high-dimensional linear regression that aims to control the false discovery rate (FDR). The main idea is to create pseudo variables, knockoffs, for each original variable, and then perform regression on the augmented dataset consisting of the original and knockoff variables. By comparing the importance of each original variable to its knockoff copy in order to determine the subset of important variables. To construct knockoffs, Barber et al.[21] have first generated a covariance matrix based on the original variables, $\Sigma = \frac{1}{n}X^\top X$, and then construct the knockoff covariance matrix with the same marginal distributions with flipping signs of pairwise correlations. In the knockoff covariance matrix $\tilde{\Sigma}$, the elements are defined as follows: $\tilde{\Sigma}_{ij} = \Sigma_{ij}$ when $i = j$ and $\tilde{\Sigma}_{ij} = -\Sigma_{ij}$ when $i \neq j$. This construction ensures that the knockoff variables have the same marginal distributions as the original ones up to flipped signs of pairwise correlations. Finally, it proceeds by sampling from a multivariate Gaussian distribution with the new covariance matrix to generate knockoff variables. Generally, the geometrical knockoff variables $\tilde{X}$ have been created to pair with original ones denoted by X, such as the matrix $(X, \tilde{X})$. This procedure does not require the variables, covariates, to be random or their distribution to be known, which makes it more applicable in practice. However, the procedure assumes a linear relationship between

the response and predictors. In addition, the procedure requires the number of observations to be larger than the number of variables/predictors. These deficiencies made limitation for applying the procedure on high dimensional data when number of observations are less than the number of predictions.

Candès et al.[22] have proposed an updated procedure to construct the knockoffs. The construction involves an approach to generate knockoff variables that mimic the distribution of the original variables while maintaining a certain level of independence. This construction ensures the exchangeability of the original variables and their knockoff counterparts, making it a suitable for knockoff-based variable selection methods. Given X, the knockoff variables $\tilde{X}$ are generated as follows: (1)draw an independent copy Z from the same distribution as X; (2)compute the conditional expectation $\mathbb{E}[Z|X]$ and the conditional covariance $\text{Var}(Z|X)$; (3) generate $\tilde{X}$ from a multivariate Gaussian distribution with mean $\mathbb{E}[Z|X]$ and covariance $\text{Var}(Z|X)$. The joint distribution of $(X, \tilde{X})$ is exchangeable by construction, which means that swapping X and $\tilde{X}$ does not change the joint distribution. This property is important for the validity of the updated knockoffs procedure as it guarantees that the knockoff variables are as desirable as the original copy in terms of the distribution. The knockoff variables are then used in conjunction with a suitable test statistic to perform variable selection, with the guarantee that the proportion of false discoveries among the selected variables is well controlled.

1.3 The Proposed Methods

Although stability selection can facilitate the identification of more stable and reproducible findings, by far, it has been merely applied to non-robust variable selection methods[44], which is not resistant to the heterogeneous cancer genomics data with outlying observations. Such a deficiency has motivated us to propose the stability selection-assisted robust variable selection method. Specifically, we adopt the LAD-LASSO as the baseline robust penalization approach, and conduct shrinkage estimation using LAD-LASSO across multiple subsamples by following the strategy of stability selection. Through extensive simulation studies, we

have demonstrated the advantage of the proposed method over multiple competing alternatives. In addition, analyses of the skin cutaneous melanoma (SKCM) and human breast cancer datasets from The Cancer Genome Atlas (TCGA) have demonstrated that the stability selection-assisted LAD–LASSO yields better prediction performance and findings with important biological implications.

Chapter 2

Robust Variable Selection with Stability Selection

2.1 Introduction

Cancer, a major global health challenge, has seen stabilized incidence rates and significant advancements in survival rates due to the development of genomic techniques. These advancements are largely attributed to a deeper understanding of prognostic markers, more efficient techniques for early-stage cancer detection, and more effective diagnostic and therapeutic strategies. Central to this understanding is gene feature selection, which plays a crucial role in deciphering high-dimensional genomic data and contributes to advancements in cancer research, prevention, early detection, and treatment. According to the latest US report, approximately 19 million new cases and 600,000 related deaths are expected in 2022. However, the past decade has witnessed a notable upturn: a 68% 5-year survival rate for adults and 85% for children[52]. These achievements could be attributed to the plethora of feature selection techniques that have been developed in recent years, propelled by the rapid advancements in genomic research[1;2].

Given the high-dimensional nature of data encountered in genomic research, datasets often exhibit heavy-tailed distributions. As an example of analyzed data in this study, we

used the (log-transformed) Breslow's depth from The Cancer Genome Atlas (TCGA) Skin Cutaneous Melanoma (SKCM) data. Figure 2.1 show the both boxplot and histogram of the residuals data with heavy-tailed and non-normal trends. The Shapiro–Wilk test indicates deviation from normality with a p-value < 0.01. This observation arises frequently due to the heterogeneity of complex diseases including cancer. Furthermore, in cancer research, cost constraints often prevents rigorous patient recruitment, leading to data contamination from mixed disease subtypes or unreliable data recordings[53;54].

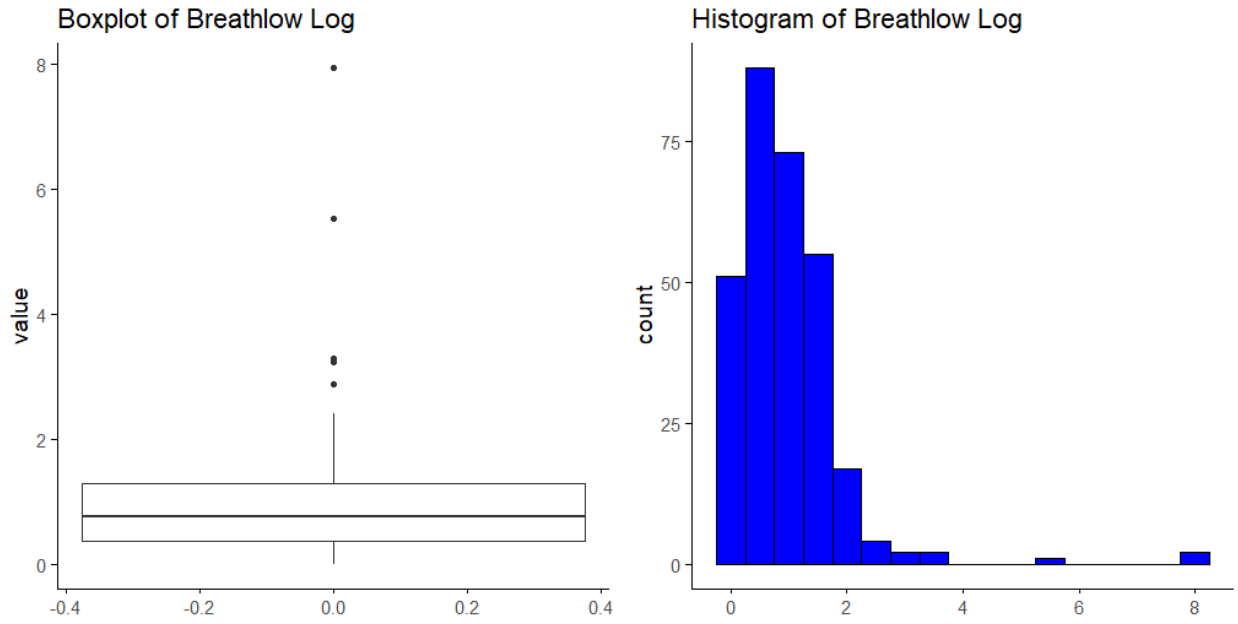

Figure 2.1: *Residual distribution for log-transformed breslow depth in TCGA's SKCM Study*

Considering the challenges posed by the high dimensions and heavy-tailed nature of data and potential contamination, a range of robust feature selection techniques has been developed in recent years, driven by the rapid advancements in genomic research[1]. These regularization methods adopt robust loss functions that differ from non-robust least squares loss or Gaussian likelihood-based procedures, offering the flexibility to account for data irregularities and model specifications. For example, quantile regression based high-dimensional methods have been extensively developed to analyze multi-omics data with skewed disease phenotypes[32;33;43;55;56]. In addition, rank based variable selection methods have also been

proposed, where non-smooth rank loss functions effectively account for heterogeneity in both continuous and survival outcomes in biomedical research[37;57–59]. With robust regularization, the penalty functions, such as the L1-norm penalization corresponding to LASSO, that are widely developed for conducting sparse estimation in non-robust variable selection methods are still crucial for feature selection and controlling model complexity in the presence of long-tailed distributions in the phenotypic traits. Combining robust loss functions with the LASSO penalty, as seen in methods like least absolute deviation LASSO (LAD-LASSO)[13], enhances resistance to outliers and data contamination in high-dimensional data. Nonetheless, challenges persist in accurately selecting relevant features, with high false positive rates often resulting from the difficulty in choosing a reasonable tuning parameter to impose appropriate amount of penalty on regression coefficients.

For regularization methods, statistical inference plays a critical role to quantify uncertainty of shrinkage estimates[44]. For example, Meinshausen and Bühlmann[19] have shown that graphical LASSO may lead to unstable sparse graphical structures. To remediate the shortcoming in the analysis, a common practice is to report multiple graphs across the solution path[17]. Meinshausen and Bühlmann have further developed uncertainty quantification procedures to stabilize the graphical LASSO estimates, which are accompanied by rigorous error controls on false positive signals[19]. Technically, stability selection procedure fits regularization models, such as the (graphical) LASSO on multiple subsamples of the original data, and yield the stable findings if the predictors corresponds to the selection frequency above certain cutoff which is associated with the error control on selected variables[19;20;60]. In literature, the knockoff methods have also been developed to explicitly control the False Discovery Rate (FDR), which have enjoyed wide applications in genomics studies[21;22;50].

Despite the success of stability selection, a major disadvantage is that by far, it has been mainly integrated with non-robust variable selection methods. Therefore, in the presence of heterogeneous cancer outcomes, stability selection assisted penalization still result in inaccurate findings because of the non-robustness nature of the method. To fill the knowledge gap, we develop a novel and robust gene selection method that effectively integrates the least absolute deviation (LAD) LASSO within the stability selection framework. The proposed

method significantly distinguishes itself from existing methodologies. Firstly, in contrast to the original stability selection framework[19], our method employs the least absolute deviation loss function, which is particularly well-suited for handling heavy-tailed distributions and data contamination in the phenotypic traits. With the L1 penalty to promote sparsity in estimation, the LAD–LASSO offers a tailored solution for robust analysis in high-dimensional data, despite the absence of a universally superior robust loss function[1]. Secondly, compared to the LAD-LASSO which is well known for producing a large number of false positive findings[13], our method shows superior performance in identification and estimation by significantly reducing the number of false positive findings. The proposed method is not sensitive to the choice of tuning parameters, which is consistent with the pattern of stability selection with non-robust regulation methods. Third, we adopt an efficient model fitting algorithm based on the coordinate descent framework, providing a computational advantage over many existing robust variable selection methods which often exhibit high computational costs in complex data and model scenarios[33;34;61].

Overall, integrating robust penalization within the stability selection framework for the identification of important cancer genomic features has not been reported in published studies. To demonstrate the utility of the proposed method, we compare it with the alternatives through extensive simulation studies under a variety of heavy-tailed errors in both homogeneous and heterogeneous models settings. In addition, we also apply all methods under comparison to the TCGA cancer genomics data, the skin cutaneous melanoma (SKCM) and human breast cancer tumor (HBT) datasets. The proposed method has better prediction performance and identifies findings with important biological implications.

We organize the succeeding sections as follows. In the Methods section, we formulate the newly developed method. A comprehensive analysis of all methods under comparison are presented in the Simulation section and Real Data Analysis section. In the Discussion section, we interpret our findings and highlight the benefits of our results for the study of gene selections and potential future research directions. Finally, we conclude with a summary of our contributions and emphasize the significance of our study in the context of gene selection research.

2.2 Data and Model Settings

In a typical genomic dataset with n observations ($i = 1, ..., n$), let $(y_i, \boldsymbol{x_i}, \boldsymbol{C}_i)$ be independent and identically distributed random vectors, where we denote y_i as phenotypic measurement for the ith observation, and the p-dimensional vector $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^\top$ as omics measurements, such as the gene expressions, for the ith observation. In addition, the $(k+1)$-dimensional vectors $\boldsymbol{C_i} = (C_{i0}, C_{i1}, \ldots, C_{ik})^\top$ denote intercept and clinical covariates. The $C_{i0} = 1$ and the other k components of $\boldsymbol{C}_i$ correspond to clinical covariates. Consider the following model:

$$y_i = \sum_{t=0}^{k} \alpha_t C_{it} + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \tag{2.1}$$

where α_t's represents the regression coefficients for intercept and clinical covariates, and β_j's denotes the coefficients for the jth genetic factor, respectively. The ϵ_i's are random errors. We define $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, ..., \alpha_k)^\top$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^\top$, respectively. The preeminent methods for selecting effect genes involves identifying significant predictors of the response variable, which allows researchers to distinguish the genes that have a meaningful impact on the outcome. Accrossing the entire genome, millions of genes are contained, but only small amount of genes could affect gene expression. This condition manifests the sparse property of gene data. When analyzing millions of genotypes, the traditional approach can result in increased false discoveries and may not adhere to the normal distribution assumption. Consequently, there's a pressing need for a novel method to accurately identify instances where $\beta_j \neq 0$.

The least absolute shrinkage and selection operator (LASSO)[7] has emerged as a powerful regularization method to perform penalized estimation by shrinking regression coefficients corresponding to the omics features towards 0. LASSO and other regularization methods work by adding a penalty term in the objective function to fit constrained least square model. However, LASSO is well known to yield a large number of false positives due to its biased nature[27]. This can inadvertently lead to the underestimation and shrinking of true non-zero

coefficients. Stability selection has been proposed to quantify uncertainty of shrinkage estimates. It improves the reliability of variable selection by incorporating data variability, thus aiding in the differentiation between true positives and false positives, and ensuring that true non-zero coefficients are retained.[19;20]. Meanwhile, robust regression techniques have been developed to provide more accurate coefficient estimates in the presence of outliers, complementing non-robust penalization methods (such as LASSO) under heterogeneous data[13;14;61].

2.2.1 The LAD–LASSO

Given the heterogeneity nature of complex diseases, outlying observations and long-tailed distributions have been widely encountered in genomics studies. Therefore, non-robust variable selection methods such as LASSO, SCAD and MCP will lead to biased estimation results and false findings[1;62]. To incorporate robustness in variable selection, we first consider the LAD–LASSO[13] as follows:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{i=1}^{n} |y_i - \sum_{t=0}^{k} \alpha_t C_{it} - \sum_{j=1}^{p} \beta_j x_{ij}| + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{2.2}$$

The main difference of objective functions between LAD–LASSO and LASSO, is the replacement of least square loss in LASSO by least absolute deviation loss in LAD-LASSO. The LAD loss function quantifies prediction error using L1 norm, thus down-tuning residuals with respect to outliers compared to the least square loss. Therefore, LAD-LASSO is less sensitive to outliers in contrast to non-robust variable selection methods such as LASSO.

In high-dimensional bioinformatics studies, identifying a subset of important genes can be very challenging due to the curse of dimensionality. In the "large p, small n" setting, shrinkage estimates can be very unstable especially when accommodating complicated sparsity structures[19], as a very small perturbation in selecting tuning parameter λ can result in significant changes in the selected predictors and associated sparse structures. Meinshausen and Bühlmann have developed the stability selection procedure to be coupled with regularization methods, which not only stabilizes the shrinkage estimates but also provides error control on expected number of false positives.

However, the stability selection assisted LASSO is still a non-robust variable selection method and thus vulnerable to long-tailed distribution and outliers in the phenotypic trait. To marry the advantage in robust estimation and the stability selection for penalized variable selection in high-dimensional cancer genomics data, we have developed the stability selection with LAD-LASSO, as described in the next section.

2.2.2 Stability Selection with LAD-LASSO

We describe the detailed procedures of stability selection with LAD–LASSO below. A graphical representation is provided in Figure 2.2. The original dataset is denoted as $(y_i, \boldsymbol{C_i}, \boldsymbol{x_i})$, $i = 1, 2, \ldots, n$.

Firstly, create a subsample by selecting a fraction of q rows from the original dataset randomly without replacement, where q ranges from 0 to 1. The dimensions of original dataset are $n \times (k + 2 + p)$, where y_i represents the response, $\boldsymbol{C_i}$ denotes the $(k + 1)$–dimensional vector of clinical covariates including the intercept, and $\boldsymbol{x_i}$ are the p-dimensional measurements of omics predictors on the i-th subject. This step creates a subsample of dataset with dimension of nq by $k + 2 + p$. It represents the proportion of the original dataset in our subsample which results in a smaller set of size nq. This step is repeated, for example m times, resulting m different subsample datasets. Each dataset is assigned with an identifier, l, ranging from 1 to m, to represent all m different datasets, which is denoted as $(\boldsymbol{y}, \boldsymbol{C}, \boldsymbol{x})^{(l)}$.

Secondly, apply LAD-LASSO to each subsample. From each subsample, we identify the sparse coefficient vector $\hat{\boldsymbol{\beta}}^{(l)}$ using LAD-LASSO. In each subsample, if a predictor is chosen, we mark it as 1; if not, it's marked as 0. We sum these marks across all subsamples to determine the overall selection frequency for each predictor. For example, the frequency for selecting j-th predictor across all m subsamples can be calculated as $\sum_{l=1}^{m} I(\hat{\beta}_j^{(l)} \neq 0)$. Note that clinical covariates including intercept, represented by $\hat{\boldsymbol{\alpha}}^{(l)}$, are not subject to selection. We define a set of frequencies for all the predictors, $S = \{f_j = \sum_{l=1}^{m} I(\hat{\beta}_j^{(l)} \neq 0), j \in 1, 2, \ldots, p\}$, that is equal to the total selection frequency of each predictor across

all m subsamples. Each frequency of predictor in S has a value ranging between 0 and m. Alternatively, each element of S denotes how often a particular predictor was non-zero in the m different subsamples. A predictor with a high frequency in S is considered more important.

Finally, we choose a threshold, θ, for S. The threshold is a fraction, ranging from 0 to 1, that serves as a criterion to determine the importance of predictors. A predictor is considered important and included in the final set, $W(\theta)$, if its selected frequency across all subsamples surpasses this threshold θ. This average frequency for selecting a predictor, for example the jth predictor, can be calculated as $\frac{\sum_{l=1}^{m} I(\hat{\beta}_j^{(l)} \neq 0)}{m}$. If the average of frequency of a predictor in the final set exceeds the threshold θ, it will be retained. Otherwise, it will be discarded from the final model. This procedure ensures that the important predictors are identified and selected for further analysis. The detailed selection procedure is described below:

$$W(\theta) = \{x_j : j \in 1, 2, \ldots, p \text{ such that } \frac{\sum_{l=1}^{m} I(\hat{\beta}_j^{(l)} \neq 0)}{m} \geq \theta\}, \tag{2.3}$$

where: the $W(\theta)$ is the set of selected predictors which surpass a given threshold θ. Here the p is the total number of predictors; m is the number of subsamples; the $I(\cdot)$ is the indicator function; the $\hat{\beta}_j^{(l)}$ is the j-th coefficient vector estimated from the l-th subsample of the data.

In general, the stability selection algorithm can be described by repeatedly subsampling the dataset, performing feature selection on each subsample and then deciding features based on the frequency of their selection across all subsample[19;20]. The hypothesis of this method is that the feature selection methods can consistently identify those prominent features in subsamples. The accuracy of our approach is gauged by how frequently the estimates emerged across subsamples. In subsequent part, our proposed stability selection employs the LAD-LASSO method, detailed in following computational algorithm.

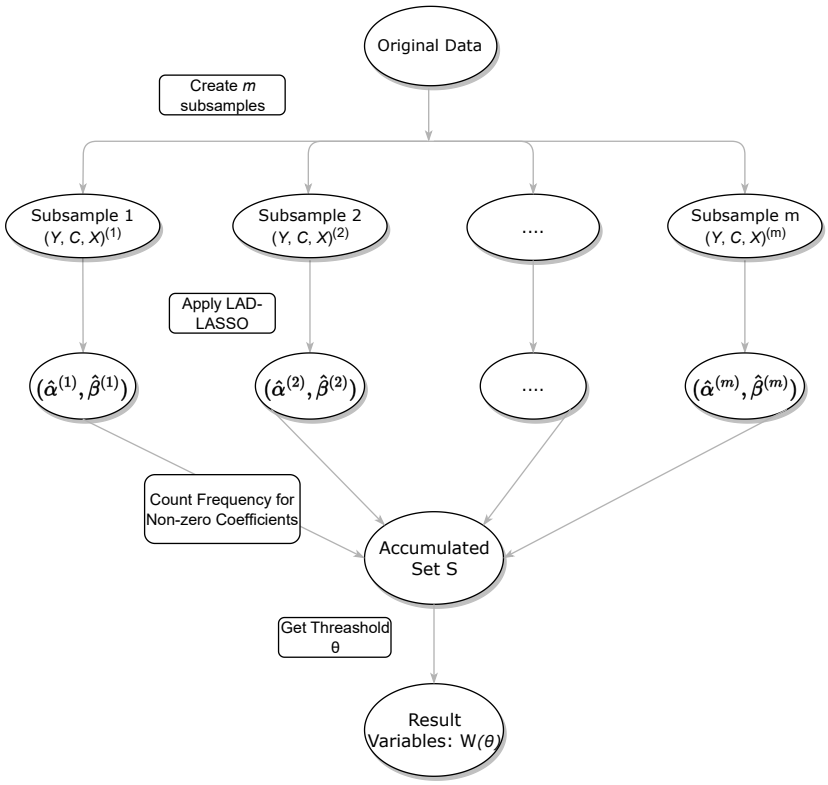

Figure 2.2: *Procedures of LAD-LASSO with stability selection algorithm*

2.2.3 Computation algorithm for LAD-LASSO

The algorithm for fitting LAD-LASSO is based on the greedy coordinated descent update with weighted median regression, which has been developed in literature for non-differentiable loss functions such as the LAD loss and quantile check loss[33;34;36;61]. The loss function of LAD-LASSO in our model can be written as $g(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} |y_i - \sum_{t=0}^{k} \alpha_t C_{it} - \sum_{j=1}^{p} \beta_j x_{ij}| + \lambda \sum_{j=1}^{p} |\beta_j|$. The function can be further re-expressed as:

$$\begin{aligned} g(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \sum_{i=1}^{n} |x_{ij}| \left| \frac{y_i - \sum_{t=0}^{k} \alpha_t C_{it}}{x_{ij}} - \sum_{j=1}^{p} \beta_j \right| + \lambda \sum_{j=1}^{p} |\beta_j| \\ &= \sum_{i=1}^{n} |x_{ij}| \left| \frac{y_i - \sum_{t=0}^{k} \alpha_t C_{it} - \sum_{s \neq j}^{p} \beta_s x_{is}}{x_{ij}} - \beta_j \right| + \lambda \sum_{j=1}^{p} |\beta_j|, \end{aligned} \tag{2.4}$$

.

where coefficient vector $\boldsymbol{\alpha}$ represent the coefficients corresponding to the intercept and clinical covariates. We first update α by weighted median regression. Then, update β by

finding the weighted median of the loss function below when α holds fixed:

$$\beta_j^{(d)} = \text{argmin}_{\beta_j} \frac{1}{n+1} \sum_{i=1}^{n+1} w_{ij} u_{ij}, \tag{2.5}$$

where w_{ij} and u_{ij} are defined below. The solver of the above loss function is actually the weighted median of the errors. the dth coefficient is updated as u_{ij}. Where:

$$u_{ij}^{(d)} = \begin{cases} \frac{y_i - \sum_{t=0}^{k} \alpha_t C_{it} - \sum_{s \neq j} \beta_s^{(d-1)} x_{is}}{x_{ij}} - \beta_j & i = 1, 2, \ldots, n \\ 0 - \beta_j & i = n+1. \end{cases} \tag{2.6}$$

And:

$$w_{ij}^{(d)} = \begin{cases} |x_{ij}| & i = 1, 2, \ldots, n \\ \lambda & i = n+1. \end{cases} \tag{2.7}$$

The procedure will continue to iterate and update α and β until the estimates converge. We apply the 'regnet' R package to implement above weighted median LAD-LASSO algorithm[34;35]. The tuning parameters are determined by 5-folds cross-validation in following simulation study and real data analysis.

In comparison to the conventional LAD-LASSO estimator, our proposed weighted coordinate descent algorithm offers a more scalable and efficient solution, especially for large datasets. The conventional approach often attempts to address its computational challenges by transforming the LAD-LASSO problem into a pure LAD problem using an augmented dataset and solving in linear programming[13]. This involves expanding the original dataset $\{(y_i, \boldsymbol{C_i}, \boldsymbol{x_i})\}_{i=1}^{n}$ with dimensions $n \times (1 + k + 1 + p)$ to $(n + p) \times (1 + k + 1 + p)$, denoted as $\{(y_i^*, \boldsymbol{C_i^*}, \boldsymbol{x_i^*})\}_{i=1}^{n+p}$. The first n rows mirror the original. In the additional rows, y_i^* is set to zero, $\boldsymbol{C_i^*}$ is also set to zero ensuring they're unaffected by the LASSO penalty, and $\boldsymbol{x_i^*}$ is structured as vectors of length p containing a singular λ, embodying the LASSO regularization. The objective function for the augmented dataset is to minimize $\sum_{i=1}^{n+p} \left| Y_i^* - \sum_{t=0}^{k} \alpha_t C_{it}^* - \sum_{j=1}^{p} \beta_j X_{ij}^* \right|$. However, the procedure would cost too much in computation.

By applying the stability selection algorithm to LAD-LASSO, the accuracy and robustness of a model can be improved by selecting important predictors across subsample sets. To evaluate the performance of the stability selection of LAD-LASSO, we assess combinations of threshold θ and subsample size q under different distributions in following section. In addition, subsampling method is not limited to being used only for the stability selection algorithm. It can also be used individually to obtain model features or combined with other tuning parameter selection methods such as cross-validation, validation, Bayesian information criteria, etc. The methods can help to further improve the accuracy and robustness of the model.

Table 2.1: *Stability Selection Algorithm*

Stability Selection Algorithm

Draw a size of fraction q from original data without replacement randomly:
Repeat this process for m times independently to formulate m subsamples.
For each subsample of the data
Initialize $d = 0$, $\alpha^{(0)}$ and $\beta^{(0)}$;
Repeat
Update $\alpha^{(d+1)}$ component-wise using weighted median regression;
for $j = 1, 2, \ldots, p$
Compute u_j and w_j via equations (2.6) and (2.7);
Update $\beta_j^{(d+1)}$ using the weighted median in equation (2.5);
Increment j by 1;
end for
Increment d by 1;
until convergence.

End For
Count the frequency of predictors that have non-zero coefficients across the m subsamples. Choose a threshold θ. Select final predictors via equation (2.3).

Remark: Within our adapted framework of the stability selection algorithm, the parameters fraction q and threshold θ are important. The fraction q delineates the subsample size, which directly influences the variance of our results. Simultaneously, the threshold θ sets a decisive criterion for predictor inclusion, based on their consistent presence across subsamples. While the optimal settings for these parameters can vary, their thoughtful selection is important to ensure the robustness and accuracy of the results.

2.3 Simulation

We have conducted a comprehensive evaluation of multiple variable selection methods in simulation study. Our proposed method is the stability selection assisted LAD–LASSO, termed as LADL–SS. Other alternatives such as LAD–LASSO (LADL)[13] and LASSO[7] are considered. Variants of LASSO, that are the LASSO with stability selection, LASSO–SS[19] and LASSO with permutation–assisted tuning, LP–SS[50] are also included for comparison. Moreover, we have extended LP–SS to its robust counterpart, LAD with permutation–assisted tuning, or LADLP–SS, which has been proposed for the first time and adopted in simulation. All methods under comparison are summarized in Table A.1.

For those methods integrating stability selection, we set a selection threshold of $\theta = 90\%$, with a sample size m=80. This indicates that signal predictors were chosen if they appeared in more than 90% of selections, or more than 72 times. In our evaluations, we have compared L_1 penalty within LAD regression to L_1 penalty within least square regression. We have also contrasted the efficacy of the stability selection algorithm using cross-validation tuning versus the permutation-assisted tuning procedure. To assess the robustness of our proposed method, LADL-SS, we have compared it to LADL, LASSO and previously mentioned variants across normal and different heavy-tailed distributions.

For a comprehensive evaluation of our proposed method's effectiveness, we implement the data generation process across two divergent covariance structures: namely, the First Order Autoregressive Model AR(1), and the data formulated from the Skin Cutaneous Melanoma (SKCM) dataset. In the context of the AR(1) covariance structure, the analysis extends to encompass both homogeneous and heterogeneous variance models, thus facilitating a robust methodological assessment. We first generate a random matrix X containing n observations of p predictors from a multivariate normal distribution $MVN(0, \Sigma)$. The covariance matrix Σ is a symmetric positive definite matrix M of size $p \times p$, where each element $M_{[i,j]}$ is assigned the value $0.5^{|i-j|}$. Note that this matrix X does not include any non-genetic covariates. Then, we randomly select k signal predictors from the total of p predictors to be the affected predictors. In this study, we set k to 20. Finally, we generate the responses by using the

following linear homogeneous error model: $y_i = 1 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i$ and heterogeneous error model: $y_i = 1 + \sum_{j=1}^{k} \beta_j x_{ij} + (1 + x_{ij})\epsilon_i$. Here, for each affected predictor, we set $\beta_j =$ 0.3 or -0.3 with even probability. We generate two sets of data: $n = 500, p = 1000$ or $n = 1000, p = 2000$ for homogeneous error model and one set of data: $n = 1000, p = 2000$ for heterogeneous error model. The affected predictors β are equal to 0.3 or -0.3 with equal probability, respectively. We also consider five different error distributions for ϵ_i: N(0, 1)(Error 1),10%N(0,4) + 90%N(0,1) (Error 2), LogNormal(0,1) (Error 3), 90%N(0,1) + 10%Cauchy(0, 1) (Error 4), t-distribution with 2 degrees of freedom ($t(2)$) (Error 5). Except Error 1 is normal distribution, Error 2 to Error 5 are heavy tailed distributions.

To evaluate the performance of our proposed method on real-world datasets, we construct our matrix X using a multivariate normal distribution with a marginal mean of 0. The covariance matrix is directly generated from the actual SKCM dataset with dimensions $n = 500, p = 1000$ and $n = 1000, p = 2000$. In this dataset, our objective is to assess the performance of our methods under various covariance structures. All other settings are mirrored those used in before where we applied a linear homogeneous error model with five distinct error distributions for ϵ_i.

To enhance the validity of our results, we evaluate the performance of each method based on their identification accuracy across 100 different datasets under each distribution. Then, we considere four performance metrics which are true positive (TP), the number of correctly identified real predictors; false positive (FP), the number of incorrectly identified false predictors; F1 score, the harmonic mean of precision and recall, which measures the balance between the two metrics; Matthews Correlation Coefficient (MCC), a correlation coefficient between the predicted and true binary classifications, which takes into account true and false positives and negatives.

As delineated in Table 2.2, with dimensions $n = 1000, p = 2000$, the performance of the LADL-SS method consistently surpasses that of the other five L1 penalty methods, and this is particularly evident under heavy-tailed distributions. Notably, when we consider the standard normal Error1 distribution, the F1 and MCC metrics exhibit negligible disparities between LADL-SS, LADLP-SS, LASSO-SS, and LP-SS, despite the slightly superior perfor-

mance of LS-SS. However, these four methods significantly outperform the baseline LADL and LASSO methods within the context of normal distributions, thereby validating the utility of the stability selection algorithm in enhancing performance—a conclusion that aligns with earlier research on LASSO[19].

In the case of heavy-tailed distributions, ranging from Error2 to Error5, methods incorporating stability selection algorithms (namely, LADL-SS, LADLP-SS, LASSO-SS, and LP-SS) consistently surpass those that do not (namely, LADL and LASSO) for each respective regression. The performance of all methods, with or without the stability selection algorithm, diminishes under heavy-tailed distributions, with the sole exception of LADL-SS. For instance, in the Error2 scenario, the declining F1 scores and MCC of LP-SS, LADLP-SS, and LASSO-SS denote an overall drop in performance. On the other hand, LADL-SS displays a minor increase in both its F1 score (0.96(sd 0.03)) and MCC (0.96(sd 0.03)), thereby underscoring its robust performance. In the Error3 distribution, the LAD regression methods, notably LADL-SS and LADLP-SS, outperform other methods, evidenced by their higher F1 scores of 0.97 (sd 0.03) and 0.86 (sd 0.09), respectively, and MCC values of 0.97 (sd 0.03) and 0.87 (sd 0.08), respectively. In comparison, the LS regression methods, LASSO-SS and LP-SS, report lower F1 scores of 0.76 (sd 0.12) and 0.66 (sd 0.17), and MCC values of 0.77 (sd 0.11) and 0.7 (sd 0.14), respectively. LADL-SS upholds the best performance, overshadowing LADLP-SS.

In the Error4 scenario, which combines a normal distribution with a Cauchy distribution, LADL-SS and LADLP-SS significantly outperform LASSO-SS and LP-SS. This performance superiority is exemplified by their higher F1 scores of 0.96(sd 0.03) and 0.69(sd 0.13), respectively, and MCC values of 0.96(sd 0.03) and 0.72(sd 0.1), respectively. In contrast, LASSO-SS and LP-SS yield F1 scores of only 0.3(sd 0.36) and 0.26(sd 0.32) as well as MCC values of 0.3(sd 0.37) and 0.3(sd 0.32), respectively. Even under such conditions, methods incorporating stability selection algorithms retain superior F1 scores compared to LADL 0.30(sd 0.10) and LASSO 0.26(sd 0.2), and MCC compared to LADL 0.40(sd 0.11) and LASSO 0.25(sd 0.2). The Error5 distribution mirrors the pattern observed under Error4, with LAD regression methods (LADL-SS and LADLP-SS) eclipsing LS regression methods (LASSO-SS

and LP-SS). Here, LADL-SS outshines all, achieving the highest F1 scores of 0.94(sd 0.05) and MCC of 0.94(sd 0.05). This overall pattern suggests that LAD regressions exhibit more effectiveness than LS regressions in L1 penalty models under heavy-tailed distributions. In summary, LAD regression showcases a higher level of robustness than LS regression within L1 penalty models, and this robustness is further enhanced by the inclusion of the stability selection algorithm, particularly under heavy-tailed distributions. Our proposed LADL-SS method consistently achieves high F1 scores and MCC across both normal and heavy-tailed distributions, confirming its superior and reliable performance.

Table 2.2: *Results of Identifications for Simulated Data Under Homogeneous Error: Dimension* (n, p) *= (1000,2000).*

		LADL–SS	LADL	LADLP-SS	LASSO-SS	LASSO	LP-SS
Error 1	TP	19.90(0.20)	20.00(0.00)	17.40(2.30)	19.90(0.20)	20.00(0.00)	19.90(0.20)
	FP	3.10(1.70)	198.40(102.60)	0.60(0.80)	1.50(1.00)	76.50(21.00)	0.70(1.20)
	F1	0.93(0.04)	0.19(0.07)	0.91(0.06)	0.96(0.02)	0.35(0.06)	0.98(0.03)
	MCC	0.93(0.03)	0.30(0.07)	0.91(0.07)	0.96(0.02)	0.45(0.06)	0.98(0.03)
Error 2	TP	19.90(0.30)	19.90(0.20)	14.10(3.00)	19.70 (0.60)	19.90(0.20)	17.30(1.90)
	FP	1.60(1.20)	137.90(50.80)	0.60(0.70)	8.40(3.60)	82.20(23.70)	1.20(1.10)
	F1	0.96(0.03)	0.25(0.09)	0.8(0.11)	0.82(0.06)	0.34(0.07)	0.90(0.05)
	MCC	0.96(0.03)	0.36(0.08)	0.82(0.1)	0.83(0.05)	0.44(0.06)	0.90(0.05)
Error 3	TP	19.90(0.40)	19.70(0.60)	15.70(2.80)	15.10(3.70)	19.80(1.60)	10.70(3.90)
	FP	1(0.90)	84.50(33.60)	0.30(0.50)	3.90(3.50)	69.20(25.50)	0.60(0.80)
	F1	0.97(0.03)	0.34(0.09)	0.86(0.09)	0.76(0.12)	0.37(0.07)	0.66(0.17)
	MCC	0.97(0.03)	0.44(0.07)	0.87(0.08)	0.77(0.11)	0.46(0.06)	0.70(0.14)
Error 4	TP	19.60(0.60)	18.90(1.40)	10.80(2.80)	5.70(7.30)	9.20 (8.20)	4.00(5.40)
	FP	1.40(1.10)	87.80(36.20)	0.20(0.40)	1.70(2.70)	33.40(37.20)	0.20(0.40)
	F1	0.96(0.03)	0.30(0.10)	0.69(0.13)	0.30(0.36)	0.26(0.20)	0.26(0.32)
	MCC	0.96(0.03)	0.40(0.11)	0.72(0.10)	0.30(0.37)	0.25(0.20)	0.30(0.32)
Error 5	TP	18.70(1.20)	19.90(0.30)	10.30(2.70)	7.80(4.70)	14.20(4.90)	4.30(3.00)
	FP	1.30(1.40)	90.10(29.90)	0.30(0.60)	2.10(1.70)	40.90(19.90)	0.50(0.70)
	F1	0.94(0.05)	0.32(0.08)	0.66(0.12)	0.47(0.25)	0.37(0.12)	0.32(0.20)
	MCC	0.94(0.05)	0.43(0.07)	0.70(0.10)	0.50(0.25)	0.42(0.13)	0.40(0.20)

In Table 2.3, with dimensions $n = 500, p = 1000$, the overall performance pattern of each method resembles that of Table 2.2. LADL-SS method consistently outperforms than

Table 2.3: *Results of Identifications for Simulated Data Under Homogeneous Error: Dimension* (n, p) *= (500,1000).*

		LADL–SS	LADL	LADLP-SS	LASSO-SS	LASSO	LP-SS
Error 1	TP	19.80(0.40)	19.60(0.71)	5.90(2.40)	19.80(0.40)	20.00 (0.00)	16.10(2.60)
	FP	3.60(2.00)	17.70(5.00)	0.10(0.30)	2.90(1.90)	73.60(18.40)	0.40(0.60)
	F1	0.92(0.04)	0.68(0.06)	0.44(0.14)	0.93(0.03)	0.36(0.06)	0.88(0.08)
	MCC	0.92(0.04)	0.71(0.05)	0.52(0.11)	0.93(0.03)	0.45(0.05)	0.88(0.07)
Error 2	TP	15.60(2.40)	10.80(2.70)	5.90(2.40)	13.80(2.50)	190(1.10)	7.60(3.00)
	FP	1.70(1.30)	2.80(1.60)	0.70(0.90)	2.80(2.10)	66.80(22.90)	1.00(1.50)
	F1	0.83(0.08)	0.64(0.11)	0.43(0.14)	0.75(0.08)	0.37(0.07)	0.51(0.16)
	MCC	0.83(0.07)	0.65(0.11)	0.5(0.12)	0.75(0.08)	0.45(0.06)	0.57(0.14)
Error 3	TP	13.50(3.40)	18.30(1.50)	3.90(2.30)	5.90(4.10)	14.30(5.40)	3.00(2.30)
	FP	0.50(0.70)	18.80(4.50)	0.20(0.40)	1.60(1.70)	50.80(33.10)	0.20(0.40)
	F1	0.78(0.12)	0.64(0.06)	0.31(0.17)	0.39(0.23)	0.35(0.11)	0.24(0.16)
	MCC	0.80(0.11)	0.67(0.06)	0.39(0.19)	0.43(0.23)	0.40(0.10)	0.34(0.17)
Error 4	TP	18.60 (1.10)	18.70 (1.20)	6.70 (2.90)	3.40 (6.10)	6.80 (7.30)	2.40 (4.60)
	FP	2.40(1.80)	19.70(4.00)	0.40(0.80)	1.10(2.20)	23.00(28.70)	0.40(0.60)
	F1	0.91(0.05)	0.61(0.06)	0.48(0.20)	0.19(0.29)	0.19(0.16)	0.16(0.26)
	MCC	0.91(0.05)	0.65(0.05)	0.54(0.13)	0.21(0.30)	0.22(0.18)	0.20(0.27)
Error 5	TP	14.00(2.90)	14.00(2.10)	4.70(2.60)	4.30(3.80)	11.00(5.80)	2.50(2.00)
	FP	1.60(1.60)	13.70(3.00)	0.50(0.60)	1.10(1.40)	32.70(22.20)	0.50(0.60)
	F1	0.78(0.11)	0.58(0.08)	0.36(0.16)	0.29(0.25)	0.32(0.14)	0.21(0.15)
	MCC	0.79(0.10)	0.61(0.09)	0.44(0.14)	0.32(0.26)	0.35(0.15)	0.28(0.18)

LADLP-SS, LASSO–SS and LP-SS and LAD regressions are better than LS regressions especially in heavy-tailed distributions. Comparing the two tables, we notice a decrease in model performance as sample size and dimensions are reduced. Nevertheless, methods utilizing LAD regression and the stability selection algorithm exhibit some degree of resistance to this effect. Notably, our proposed LADL-SS method demonstrates strong resistance. In Error1 the F1 scores and MCC remain 0.92(sd 0.04) and 0.92(sd 0.04) respectively. In Error2 the F1 scores and MCC are 0.83(sd 0.08) and 0.83(sd 0.07). In Error3 the F1 scores and MCC are 0.78(sd 0.12) and 0.8(sd 0.11). In Error4 the F1 scores and MCC are 0.91(sd 0.05) and 0.91(sd 0.05); and in Error5 the F1 scores and MCC are 0.78(sd 0.11) and 0.79(sd 0.1).

The performance profiles for various methods across different error distributions are vi-

sually presented in the Figure 2.3 and 2.4. For the Error1, as depicted in the Firgure 2.3, LADL-SS demonstrates competitive performance, often aligning closely with LADLP-SS, LASSO-SS and LP-SS in terms of F1 and MCC. However, as we transition to more complex error distributions, the both figures distinctly capture the edge LADL-SS possesses over other methods. In particular, under conditions with increased noise or heavy-tailed errors, LADL-SS consistently maintains its robustness across different metrics. Methods without stability selection often don't perform as well in many testing conditions. Looking closely at the plots, LADL-SS stands out. When compared to LP-SS and LADLP-SS, the strength of LADL-SS is clear, often giving better or similar results. While certain error distributions can be challenging for many methods, including LADL-SS, across the range of situations shown in the plots, LADL-SS consistently proves to be a reliable and effective choice.

Moreover, for the data generation from SKCM real scenarios in Appendix of Table A.2 and Table A.3 perform similar pattern of results in comparing to Table 2.2 and Table 2.3. In real data set, the 100 replicates with five distributions indicate 100 different covariance matrix data structure with five distributions. It gives more difficulties in correct feature selection. Considering the LADL-SS method, it shows a consistently high TP, F1 score, MCC, and low FP across all types of errors. It is notable that LADL-SS method consistently performs better in terms of both F1 Score and MCC, compared to the other methods, especially under heavy-tailed distribution. For instance, under Error3, LADL-SS scores 0.79(sd 0.06) on the F1 and 0.78(sd 0.06) on the MCC, both of which are relatively high scores. In contrast, other methods, such as LADLP-SS and LP-SS, show poor performance, as evidenced by low F1 scores, for instance, 0.05(sd 0.06) and 0.05(sd 0.06), and low MCC scores, for instance, 0.25(sd 0.13) and 0.27(sd 0.14) under Error3 repectively. Furthermore, the performance of methods with stability selection algorithm remains comparatively stable and efficient in Error1 than the methods without. LADL-SS scores 0.86(sd 0.05) for F1 and LASSO-SS scores 0.84(0.04). But our proposed method LADL-SS maintains a good balance between identifying correct positive results (TP) and not producing too many false positives (FP) across all distributions. The F1 Scores and MCCs across all error types also stay high, demonstrating good performance in terms of precision, recall, and the quality of binary

classifications. The performance of LASSO-SS decreased in other types of distributions especially under heavy tailed distribution. For example, in normal and cauchy mixture distribution Error4, the LASSO-SS only scores F1 0.38(sd 0.29). This results support that the LADL-SS method is not only effective but also reliable under various types of errors and data structures.

By far, under the heterogeneous error structure with dimensions $n = 500, p = 1000$ in Appendix of Table A.4, methods still perform similar pattern of results to compare with Table 2.2 and Table 2.3. LADL-SS consistently emerges as a leading method in 100 replicates. For the Error1, LADL-SS achieves a TP of 19.98(0.16) and an FP of 1.10(1.19), outperforming other methods like LADL, which, despite a perfect TP, has a notably higher FP of 88.80(30.14). Under the Error2, LADL-SS maintains its edge with a TP of 17.10(1.89), whereas methods such as LP-SS lag behind with a TP of just 1.85(1.18). In the context of the Error3, LADL-SS records a TP of 19.55(0.76), surpassing others like LASSO-SS, which only manages a TP of 7.05(4.22). For the Error4, LADL-SS continues its dominance by posting a TP of 19.80(0.52), clearly ahead of methods like LASSO, which yields a TP of only 6.60(8.17). Lastly, in the fifth error distribution, LADL-SS secures a TP of 18.90(0.97), proving its consistency, while others, such as LP-SS, fall behind with a TP of just 2.15(2.28). In summation, under heterogeneous error distributions, LADL-SS consistently showcases superior performance across the board, often outperforming or being markedly more consistent than the competing methods, as supported by the presented data.

We address a sensitivity analysis of the LADL-SS model concerning its configurable parameters. In practice, under the sparse hypothesis, the LADL-SS model exhibits insensitivity to parameter selection within a reasonable range. Notably, as sample size and dimensions increase, it becomes generally easier to select signals compared to when they decrease. We've examined two methodologies to fine-tune the LADL-SS model, the performance of which has been thoroughly recorded in Appendices under Tables A.5, A.6, A.7, A.8, and Table A.9. The results highlighted in boldface indicate the optimal performance across various settings. Firstly, the adjustment of the threshold (θ) was explored. An intentional decrement in the threshold implies the inclusion of a greater number of predictors in the predictor

variable set $S(\theta)$. This adjustment strategy can prove advantageous in scenarios dealing with datasets characterized by smaller sizes and dimensions or when the selection of gene variables is considerably minimal. For instance, referring to Table A.5, under the lognormal distribution Error3, both LADL-SS and LASSO-SS exhibited superior F1 scores of 0.84 (sd 0.08) and 0.47 (sd 0.22) respectively at the 80% threshold, compared to F1 scores of 0.78 (sd 0.12) and 0.39 (sd 0.23) respectively at the 90% threshold. The second method of LADL-SS adjustment entails the reduction of the subsample size during each selection phase within the stability selection algorithm. This strategy can be particularly beneficial for handling datasets with larger sizes and dimensions or in cases where an overwhelmingly large number of gene variables is selected. For instance, as evidenced in Table A.8, our proposed method, LADL-SS with 80% subsample size and 90% threshold, demonstrated lower F1 scores compared to those shown in Table A.6 LADL-SS with 70% subsample size and 90% threshold across all distributions under identical data dimensions. The results from heterogeneous error simulation in Table A.9 also support this view. The similar pattern for LADL-SS with 70% subsample size and 90% threshold shows better performance than other tuning across all 5 distributions.

The results above demonstrate that our proposed approach, LADL-SS, which combines the LAD with the stability selection algorithm for L_1 penalty, outperforms other methods in terms of F1 and MCC scores. The consistently strong performance of LADL-SS across various distributions and dataset sizes highlights the method's extensive applicability. Further investigation into the effects of adjusting the threshold θ and subsample size reveals additional potential applications for dealing with unknown distributions. In summary, the robust performance of LADL-SS underscores its wide-ranging utility in real-world scenarios.

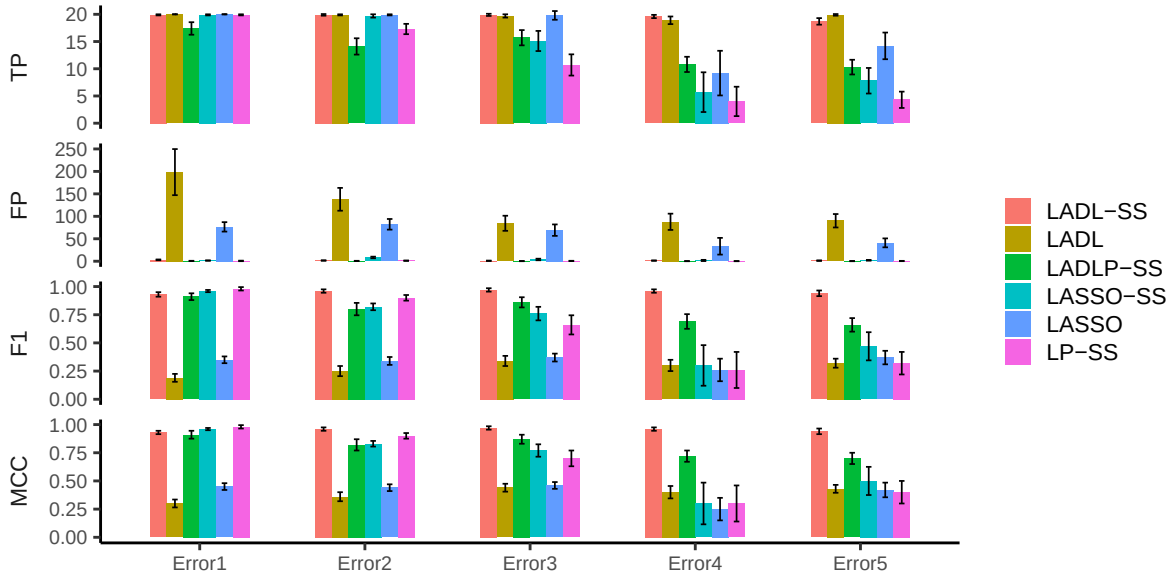

Figure 2.3: *Performance for simulated gene expression data with n=1000, p=2000 based on 100 replicates.*

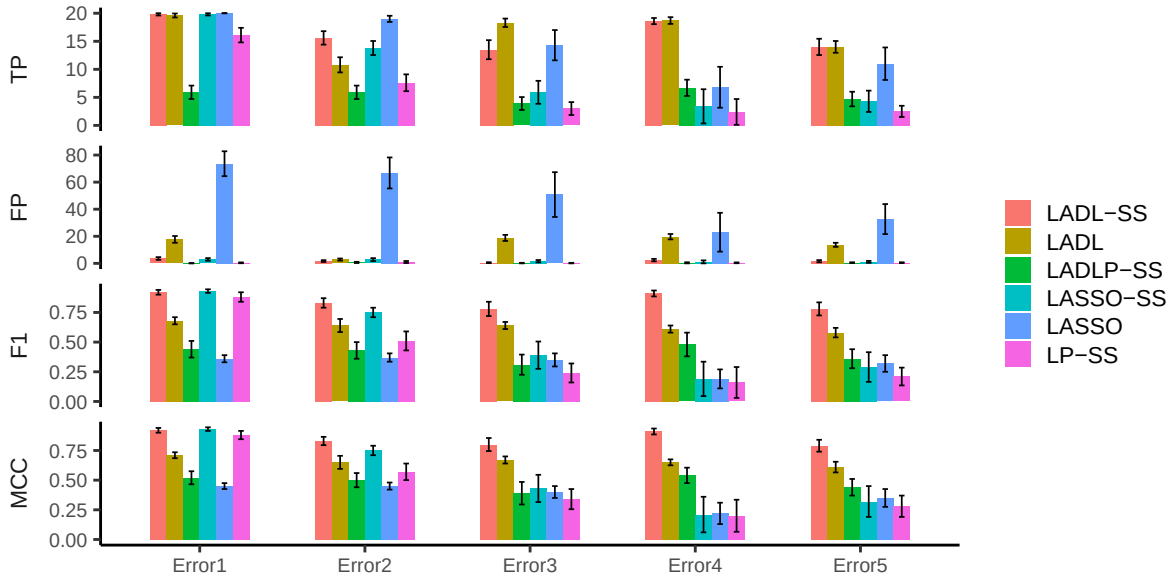

Figure 2.4: *Performance for simulated gene expression data with n=500, p=1000 based on 100 replicates.*

2.4 Real Data Analysis

2.4.1 TCGA skin cutaneous melanoma data

The Cancer Genome Atlas (TCGA), backed by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), is a combined initiative that provides an array of high-quality multi-omics data. In this study, we apply similar setting which used by[14]. The data of SKCM downloaded from the cBio Cancer Genomics Portal[63]. The aim of our investigation is to pinpoint genes that manifest genetic main effect on Breslow's thickness which is a significant prognostic variable for SKCM[64]. We took log-transformed Breslow's depth as the response variable and screened data by matching phenotypes and genotypes. Finally, we got data with 294 subjects and 20,531 gene expressions variables with age and gender as clinical variables. We adopt the marginal linear model as the pre-screening method before downstream analysis. After pre-screening, we took the 600 genes from least p-value ranking for further study.

Our proposed approach LADL–SS identifies 17 genes effects. The detailed estimation results are available from Table B.1 in the Appendix. One of the gene we identified is SEL1L3 (encoding Sel-1-like family member 3), which was reported as one of the four genes within a Tumor Immune-Relevant signature that significantly associates with overall survival in melanoma patients. This signature can potentially inform prognosis, risk assessment, and prediction of response to ipilimumab, an anti-CTLA4 antibody used in melanoma treatment. Thus, SEL1L3 might be implicated in the immune response and the progression of cutaneous melanoma, suggesting its potential as a predictive biomarker and therapeutic target[65]. The gene IER5 (Immediate early response 5) was reported to be one the gene consistently upregulated in melanoma, indicative of a potential role in chemoresistance and recurrence, which might serve a crucial role in the resilience of melanoma to treatment and its propensity for relapse, implying its possible function as a target for more effective therapeutic strategies[66]. The gene FMNL2 was identified in our proposed method. It was reported to be an influential gene in melanoma progression and invasiveness. High expression of FMNL2 correlate

with adverse patient outcomes, while its suppression impacts melanoma cell morphology and migration, suggesting potential therapeutic implications in melanoma management[67].

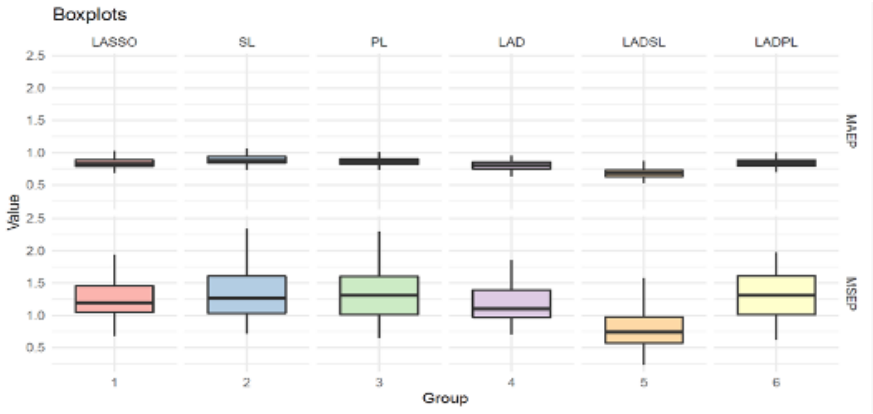

Figure 2.5: *Prediction performance of proposed method and alternative methods for SKCM. The boxplot represents the distribution of 100 replicates of each method for predicted mean absolute errors(Top) and predicted mean square errors(Bottom)*

The SKCM data was rigorously analyzed not only through our proposed LADL-SS method but also via various alternative approaches. An overlap of the effective genes identified by these methods is presented in Table 2.4, while the coefficients of the genes are detailed in the appendix Table B.1. To further contrast the predictive performance of LADL-SS with the five alternative methods, we executed a multi-split testing procedure. This procedure involved random segregation of 294 subjects into a training set (n=200) and a testing set (n=94). Each method was applied to the training set, which allowed us to derive a reduced model. This reduced model was then employed on the testing set to yield predicted responses. The discrepancies between the responses and predicted responses were then quantified to calculate the prediction mean absolute errors (PMAEs) and prediction mean square errors (PMSEs) for each method. The graphical representation of these results is provided in Figure 2.5.

Our proposed method, LADL-SS, demonstrated superior predictive performance compared to the other methods, with PMAE and PMSE values of 0.68 (sd 0.08) and 0.79 (sd 0.29), respectively. In contrast, the alternative methods yielded higher errors: LADLP-SS (PMAE: 0.84, sd 0.07; PMSE: 1.30, sd 0.35), LASSO-SS (PMAE: 0.89, sd 0.07; PMSE: 1.34,

sd 0.39), LP-SS (PMAE: 0.84, sd 0.07; PMSE: 1.30, sd 0.35), LADL (PMAE: 0.80, sd 0.07; PMSE: 1.19, sd 0.29), and LASSO (PMAE: 0.84, sd 0.08; PMSE: 1.26, sd 0.32). Thus, LADL-SS consistently outperforms other methods in terms of prediction accuracy.

Table 2.4: *The numbers of main Gene effects identified by different approaches and their overlaps.*

SKCM	Gene effects					
	LADL-SS	LADL	LADLP-SS	LASSO-SS	LASSO	LP-SS
LADL-SS	17	6	5	6	9	4
LADL		28	15	0	5	0
LADLP-SS			21	1	5	1
LASSO-SS				16	12	9
LASSO					32	15
LP-SS						15
HBT	Gene effects					
	LADL-SS	LADL	LADLP-SS	LASSO-SS	LASSO	LP-SS
LADL-SS	31	14	25	16	18	19
LADL		32	26	8	18	15
LADLP-SS			47	15	23	22
LASSO-SS				33	15	22
LASSO					35	29
LP-SS						43

2.4.2 TCGA human breast tumors data

Breast cancer, one of the most prevalent cancers worldwide, is a heterogeneous disease that can be categorized into three primary therapeutic groups: Estrogen Receptor (ER) positive, HER2/ERBB2 amplified, and Triple Negative Breast Cancers (TNBC), also known as Basal-like breast cancers[68]. Each group has distinct characteristics and treatment approaches. ER positive breast cancers are the most common and diverse, with genomic tests available to predict patient outcomes when receiving endocrine therapy. HER2/ERBB2 amplified breast cancers have seen significant clinical success due to the effectiveness of therapies targeting HER2/ERBB2. TNBCs, on the other hand, currently only have chemotherapy as a treatment option. Notably, TNBCs have an increased incidence in patients with germline BRCA1 mutations or of African ancestry[69]. The BRCA1 gene is particularly important in the context

of breast cancer because mutations in this gene are associated with a higher risk of developing breast cancer, specifically the TNBC subtype. This connection suggests that BRCA1 could be a potential target for therapeutic interventions. Most molecular studies of breast cancer have primarily focused on mRNA expression profiling or DNA copy number analysis, and more recently, massively parallel sequencing. These studies have established that breast cancers encompass several distinct disease entities, often referred to as the intrinsic subtypes of breast cancer[70]. Therefore, we identified BRAC1 as the responses variable and other SNPs with patients' age as covariates to apply SNPs analysis. The age effects has been been suggested to be associated with breast tumors[70].

Our proposed methods have identified EIF1B (Eukaryotic translation initiation factor 1B) as a gene associated with proteasome inhibition, a promising therapeutic strategy against breast cancer. The upregulation of EIF1B expression, among other translation initiation factors, is significantly enhanced by proteasome inhibition, thus highlighting its potential role in breast cancer therapeutics[71]. Furthermore, our methodology has detected the pivotal roles of NBR1 and NBR2, with a particular focus on NBR2. These genes have been reported to influence the development of breast cancer by potentially leading to decreased BRCA1 expression[72]. Moreover, the COPS6 gene, part of the ubiquitin protein ligase complex, was significantly impacted in PT12 breast cancer tumors treated with endocrine therapy using EWD as per Ingenuity analysis. This alteration in expression may influence estrogen signaling pathways within these tumors, suggesting a potential link between COPS6 activity and breast cancer progression or therapy response. However, a comprehensive understanding of this relationship and its therapeutic implications warrants further research[73].

The dataset was subjected to rigorous analysis using various methodologies, which notably included our proposed LADL-SS approach. The coefficients of potentially effective genes identified by each method have been detailed in an appendix Table B.2. We additionally provide an overlapping comparison of these effective genes identified by each method, illustrated in Table 2.4. In order to evaluate the predictive performance of each method, we implemented a strategy of multiple random splits of the data into training and testing sets. This facilitated a comparison of prediction mean square errors (PMSE) and prediction mean

absolute errors (PMAE). The corresponding results have been depicted in Figure 2.6.

LADL-SS demonstrated a commendable performance, registering a PMAE of 0.24 (sd 0.03) and a PMSE of 0.29 (sd 0.05). These results surpass those attained by the alternative methods: LADL (PMAE: 0.33, sd 0.03; PMSE: 0.37, sd 0.05), LADLP-SS (PMAE: 0.34, sd 0.04; PMSE: 0.37, sd 0.05), LASSO (PMAE: 0.37, sd 0.04; PMSE: 0.39, sd 0.05), LASSO-SS (PMAE: 0.40, sd 0.03; PMSE: 0.40, sd 0.04), and LP-SS (PMAE: 0.37, sd 0.03; PMSE: 0.39, sd 0.05). These findings support the superior performance of LADL-SS in both PMSE and PMAE criteria.

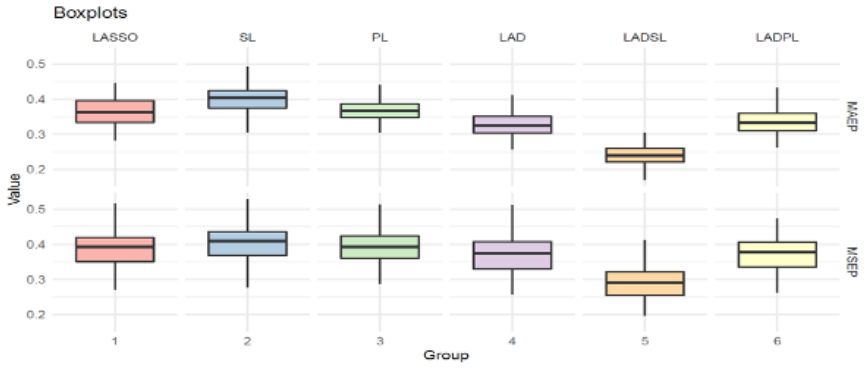

Figure 2.6: *Prediction performance of proposed method and alternative methods for HBT. The boxplot represents the distribution of 100 replicates of each method for predicted mean absolute errors(Top) and predicted mean square errors(Bottom)*

Chapter 3

Discussion

The high-throughput technologies has brought forward an abundance of multi-omics data that demand efficient and reliable variable selection methods. In this study, we have developed a novel method to integrate the least absolute deviation (LAD) LASSO with stability selection in order to prioritize important genes in cancer genomics studies. The proposed method can perform efficient and stable robust variable selection, which significantly advances from existing stability selection based non-robust penalization methods, especially in the presence of outliers and heavy-tailed distributions in the phenotypic traits.

The stability selection assisted LAD–LASSO can be extended in the following aspects. First, since LAD–LASSO cannot account for structured sparsity such as strong correlations among omics features, the network structure[34;74] and the sparse group/bi-level structure[12;75], it is potentially applicable if the LAD–LASSO is replaced by robust regularization methods incorporating tailored sparsity patterns. Second, the proposed framework can be generalized to other types of phenotypes beyond continuous ones, such as the survival outcomes. Nevertheless, due to censoring in survival analysis, a direct implementation of resampling step with robust regularization in stability selection is not trivial. Similarly, the proposed framework can also be extended to longitudinal analysis[39;41]. We will postpone the investigations to the near future. Last but not least, we focus on estimation in this study. The lack of inferential procedures with frequentist robust variable selection methods has partially

motivated the examinations on its Bayesian counterpart [14]. It is tantalizing to develop uncertainty quantification procedures when stability selection is coupled with robust penalization methods.

Bibliography

[1] Cen Wu and Shuangge Ma. A selective review of robust variable selection with applications in bioinformatics. *Briefings in bioinformatics*, 16(5):873–883, 2015.

[2] Cen Wu, Fei Zhou, Jie Ren, Xiaoxi Li, Yu Jiang, and Shuangge Ma. A selective review of multi-level omics data integration using variable selection. *High-throughput*, 8(1):4, 2019.

[3] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

[4] David Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991.

[5] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[6] Christopher M Bishop. Pattern recognition and machine learning. *Springer google schola*, 2:5–43, 2006.

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

[8] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[9] George Casella, Malay Ghosh, Jeff Gill, and Minjung Kyung. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5:369 – 411, 2010.

[10] Yu Jiang, Yuan Huang, Yinhao Du, Yinjun Zhao, Jie Ren, Shuangge Ma, and Cen Wu. Identification of prognostic genes and pathways in lung adenocarcinoma using a bayesian approach. *Cancer informatics*, 16:1176935116684825, 2017.

[11] Cen Wu and Yuehua Cui. A novel method for identifying nonlinear gene–environment interactions in case–control association studies. *Human genetics*, 132:1413–1425, 2013.

[12] Fei Zhou, Jie Ren, Xi Lu, Shuangge Ma, and Cen Wu. Gene–environment interaction: A variable selection perspective. *Epistasis: Methods and Protocols*, pages 191–223, 2021.

[13] Hansheng Wang, Guodong Li, and Guohua Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25(3):347–355, 2007.

[14] Jie Ren, Fei Zhou, Xiaoxi Li, Shuangge Ma, Yu Jiang, and Cen Wu. Robust bayesian variable selection for gene–environment interactions. *Biometrics*, 79(2):684–694, 2023.

[15] Xi Lu, Kun Fan, Jie Ren, and Cen Wu. Identifying gene–environment interactions with robust marginal bayesian variable selection. *Frontiers in Genetics*, 12:667074, 2021.

[16] Fei Zhou, Jie Ren, Shuangge Ma, and Cen Wu. The bayesian regularized quantile varying coefficient model. *Computational Statistics & Data Analysis*, 187:107808, 2023.

[17] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[18] Rahul Mazumder and Trevor Hastie. The graphical lasso: New insights and alternatives. *Electronic journal of statistics*, 6:2125, 2012.

[19] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473, 2010.

[20] Rajen D Shah and Richard J Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80, 2013.

[21] Rina Foyge Barber and Emmanuel Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[22] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 80(3):551–577, 2018.

[23] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[24] Gene H Golub, Per Christian Hansen, and Dianne P O'Leary. Tikhonov regularization and total least squares. *SIAM journal on matrix analysis and applications*, 21(1):185–194, 1999.

[25] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

[26] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

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

[28] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[29] Congrui Yi and Jian Huang. Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics*, 26(3):547–557, 2017.

[30] Sophie Lambert-Lacroix and Laurent Zwald. Robust regression through the Huber's

criterion and adaptive lasso penalty. *Electronic Journal of Statistics*, 5(none):1015 – 1053, 2011.

[31] David Kepplinger. Robust variable selection and estimation via adaptive elastic net s-estimators for linear regression. *Computational Statistics & Data Analysis*, 183:107730, 2023.

[32] Youjuan Li and Ji Zhu. L 1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17(1):163–185, 2008.

[33] Bo Peng and Lan Wang. An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics*, 24(3):676–694, 2015.

[34] Jie Ren, Yinhao Du, Shaoyu Li, Shuangge Ma, Yu Jiang, and Cen Wu. Robust network-based regularization and variable selection for high-dimensional genomic data in cancer prognosis. *Genetic epidemiology*, 43(3):276–291, 2019.

[35] Jie Ren, Luann C Jung, Yinhao Du, Cen Wu, Yu Jiang, and Junhao Liu. R package 'regnet'. 2022.

[36] Cen Wu, Qingzhao Zhang, Yu Jiang, and Shuangge Ma. Robust network-based analysis of the associations between (epi) genetic measurements. *Journal of multivariate analysis*, 168:119–130, 2018.

[37] Cen Wu, Xingjie Shi, Yuehua Cui, and Shuangge Ma. A penalized robust semiparametric approach for gene–environment interactions. *Statistics in medicine*, 34(30):4016–4030, 2015.

[38] Cen Wu, Yu Jiang, Jie Ren, Yuehua Cui, and Shuangge Ma. Dissecting gene-environment interactions: A penalized robust approach accounting for hierarchical structures. *Statistics in medicine*, 37(3):437–456, 2018.

[39] Fei Zhou, Jie Ren, Gengxin Li, Yu Jiang, Xiaoxi Li, Weiqun Wang, and Cen Wu. Penalized variable selection for lipid–environment interactions in a longitudinal lipidomics study. *Genes*, 10(12):1002, 2019.

[40] Fei Zhou, Jie Ren, Yuwen Liu, Xiaoxi Li, Weiqun Wang, and Cen Wu. Interep: An r package for high-dimensional interaction analysis of the repeated measurement data. *Genes*, 13(3):544, 2022.

[41] Fei Zhou, Xi Lu, Jie Ren, Kun Fan, Shuangge Ma, and Cen Wu. Sparse group variable selection for gene–environment interactions in the longitudinal study. *Genetic epidemiology*, 46(5-6):317–340, 2022.

[42] Fei Zhou, Jie Ren, and Cen Wu. Springer: An r package for bi-level variable selection of high-dimensional longitudinal data. *Frontiers in Genetics*, 14:1088223, 2023.

[43] Yichao Wu and Yufeng Liu. Variable selection in quantile regression. *Statistica Sinica*, 19(2):801–817, 2009.

[44] Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen. High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statistical science*, pages 533–558, 2015.

[45] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413–468, 2014.

[46] Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[47] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[48] Arindam Chatterjee and Soumendra N Lahiri. Rates of convergence of the adaptive lasso

estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics*, 41(3):1232–1259, 2013.

[49] Hanzhong Liu and Bin Yu. Asymptotic properties of lasso+ mls and lasso+ ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics*, 7:3124–3169, 2013.

[50] Songshan Yang, Jiawei Wen, Scott T Eckert, Yaqun Wang, Dajiang J Liu, Rongling Wu, Runze Li, and Xiang Zhan. Prioritizing genetic variants in gwas with lasso using permutation-assisted tuning. *Bioinformatics*, 36(12):3811–3817, 2020.

[51] Yujun Wu, Dennis D Boos, and Leonard A Stefanski. Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association*, 102 (477):235–243, 2007.

[52] Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, and Ahmedin Jemal. Cancer statistics, 2022. *CA: a cancer journal for clinicians*, 72(1):7–33, 2022.

[53] Rasika Rampatige, Saman Gamage, Sharika Peiris, and Alan D Lopez. Assessing the reliability of causes of death reported by the vital registration system in sri lanka: medical records review in colombo. *Health Information Management Journal*, 42(3): 20–28, 2013.

[54] Katja Fall, Fredrik Strömberg, Johan Rosell, Ove Andrèn, Eberhard Varenhorst, and South-East Region Prostate Cancer Group. Reliability of death certificates in prostate cancer patients. *Scandinavian journal of urology and nephrology*, 42(4):352–357, 2008.

[55] Alexandre Belloni and Victor Chernozhukov. l1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

[56] Jianqing Fan, Yingying Fan, and Emre Barut. Adaptive robust variable selection. *Annals of statistics*, 42(1):324–351, 2014.

[57] Brent A Johnson and Limin Peng. Rank-based variable selection. *Journal of Nonparametric Statistics*, 20(3):241–252, 2008.

[58] Lan Wang and Runze Li. Weighted wilcoxon-type smoothly clipped absolute deviation method. *Biometrics*, 65(2):564–571, 2009.

[59] T Cai, Jikun Huang, and L Tian. Regularized estimation for the accelerated failure time model. *Biometrics*, 65(2):394–404, 2009.

[60] Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe Vert. Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology*, 6(1):1–17, 2012.

[61] Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008.

[62] Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

[63] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, 2(5):401–404, 2012.

[64] Ashfaq A Marghoob, Karen Koenig, Flavia V Bittencourt, Alfred W Kopf, and Robert S Bart. Breslow thickness and clark level in melanoma: support for including level in pathology reports and in american joint committee on cancer staging. *Cancer*, 88(3): 589–595, 2000.

[65] Ying Mei, Mei-Ju May Chen, Han Liang, and Li Ma. A four-gene signature predicts survival and anti-ctla4 immunotherapeutic responses based on immune classification of melanoma. *Communications Biology*, 4(1):383, 2021.

[66] Jasper Wouters, Marguerite Stas, Olivier Govaere, Kathleen Van den Eynde, Hugo Vankelecom, and Joost J van den Oord. Gene expression changes in melanoma metastases in response to high-dose chemotherapy during isolated limb perfusion. *Pigment cell & melanoma research*, 25(4):454–465, 2012.

[67] Maria Gardberg, Vanina D Heuser, Ilkka Koskivuo, Mari Koivisto, and Olli Carpén. Fmnl2/fmnl3 formins are linked with oncogenic pathways and predict melanoma outcome. *The Journal of Pathology: Clinical Research*, 2(1):41–52, 2016.

[68] Charles M Perou. Molecular stratification of triple-negative breast cancers. *The oncologist*, 16(S1):61–70, 2011.

[69] Lisa A Carey, Charles M Perou, Chad A Livasy, Lynn G Dressler, David Cowan, Kathleen Conway, Gamze Karaca, Melissa A Troester, Chiu Kit Tse, Sharon Edmiston, et al. Race, breast cancer subtypes, and survival in the carolina breast cancer study. *Jama*, 295(21):2492–2502, 2006.

[70] Brigham & Women's Hospital & Harvard Medical School Chin Lynda 9 11 Park Peter J. 12 Kucherlapati Raju 13, Genome data analysis: Baylor College of Medicine Creighton Chad J. 22 23 Donehower Lawrence A. 22 23 24 25, Institute for Systems Biology Reynolds Sheila 31 Kreisberg Richard B. 31 Bernard Brady 31 Bressler Ryan 31 Erkkila Timo 32 Lin Jake 31 Thorsson Vesteinn 31 Zhang Wei 33 Shmulevich Ilya 31, et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[71] H Karimi Kinyamu, Jennifer B Collins, Sherry F Grissom, Pratibha B Hebbar, and Trevor K Archer. Genome wide transcriptional profiling in breast cancer cells reveals distinct changes in hormone receptor target genes and chromatin modifying enzymes after proteasome inhibition. *Molecular Carcinogenesis: Published in cooperation with the University of Texas MD Anderson Cancer Center*, 47(11):845–885, 2008.

[72] Christopher R Mueller and Calvin D Roskelley. Regulation of brca1 expression and its relationship to sporadic breast cancer. *Breast Cancer Research*, 5:1–8, 2002.

[73] Peter Kabos, Jessica Finlay-Schultz, Chunling Li, Enos Kline, Christina Finlayson, Joshua Wisell, Christopher A Manuel, Susan M Edgerton, J Chuck Harrell, Anthony Elias, et al. Patient-derived luminal breast cancer xenografts retain hormone receptor heterogeneity and help define unique estrogen-dependent gene signatures. *Breast cancer research and treatment*, 135:415–432, 2012.

[74] Jie Ren, Tao He, Ye Li, Sai Liu, Yinhao Du, Yu Jiang, and Cen Wu. Network-based regularization for high dimensional snp data in the case–control study of type 2 diabetes. *BMC genetics*, 18:1–12, 2017.

[75] Cen Wu and Yuehua Cui. Boosting signals in gene-based association studies via efficient snp selection. *Briefings in bioinformatics*, 15(2):279–291, 2014.

Appendix A

Appendices for Chapter 2

A.1 Summary of methods.

Table A.1: *Summary of the proposed and alternative methods.*

		Methods	Reference
Robust	**LADL-SS**	Robust LAD LASSO with stability selection	proposed for the first time
	LADL	Robust LAD LASSO	Wang et al. (2007) [13]
	LADLP-SS	Robust LAD LASSO with assisted tuning	proposed for the first time
Non-robust	**LASSO-SS**	LASSO with stability selection	Meinshausen et al. (2010) [19]
	LASSO	Original LASSO	Tibshirani. (1996) [7]
	LP-SS	LASSO with assisted tuning	Yang et al. (2020) [50]

Note: The models in the references are modified to be applicable to this study settings.

A.2 Additional simulation results

A.2.1 Real Data Simulation

Table A.2: *Results of Identifications for Simulation from Real Data Under Homogeneous Error: Dimension (n, p) = (500,1000). mean(sd) of TP, FP, F1 and MCC based on 100 replicates.*

		LADL–SS	LADL	LADLP-SS	LASSO-SS	LASSO	LP-SS
Error 1	TP	19.2(0.42)	6.33(2.39)	6.75(1.13)	19.23(0.27)	19.90(0.11)	19.11(0.46)
	FP	6.65(2.17)	79.20(5.37)	6.18(3.34)	8.03(3.03)	87.13(14.65)	18.28(6.25)
	F1	0.86(0.05)	0.08(0.03)	0.42(0.06)	0.84(0.04)	0.32(0.03)	0.69(0.07)
	MCC	0.87(0.04)	0.08(0.04)	0.42(0.06)	0.85(0.04)	0.43(0.03)	0.72(0.05)
Error 2	TP	18.90(0.90)	7.28(2.75)	7.45(2.74)	16.85(1.83)	19.60(0.67)	11.03(2.78)
	FP	6.75(2.91)	166.65(17.18)	6.15(3.47)	5.75(2.75))	64.53(12.22)	9.25(4.33)
	F1	0.83(0.06)	0.08(0.03)	0.44(0.13)	0.79(0.08)	0.37(0.05)	0.54(0.09)
	MCC	0.83(0.06)	0.07(0.05)	0.44(0.13)	0.79(0.08)	0.46(0.04)	0.54(0.09)
Error 3	TP	16.32(2.15)	6.28(2.79)	0.8(0.79)	6.45(3.56)	14.1(3.72)	3.78(2.31)
	FP	5.62(2.16)	138.70(9.28)	8.90(4.75)	2.71(1.68)	53.90(14.77)	4.83(2.96)
	F1	0.79(0.06)	0.08(0.03)	0.05(0.05)	0.42(0.13)	0.32(0.06)	0.25(0.13)
	MCC	0.78(0.06)	0.07(0.06)	0.05(0.06)	0.44(0.12)	0.36(0.08)	0.27(0.14)
Error 4	TP	16.15(2.07)	8.53(2.44)	6.70 (2.9)	6.85(6.26)	11.38(7.60)	4.90(4.18)
	FP	7.88(2.84)	154.98(25.85)	7.5(2.89)	3.2(3.34)	45.2(32.16)	5.13(4.51)
	F1	0.73(0.08)	0.04(0.04)	0.47(0.09)	0.38(0.29)	0.25(0.13)	0.28(0.21)
	MCC	0.73(0.08)	0.01(0.07)	0.47(0.09)	0.39(0.29)	0.28(0.16)	0.29(0.20)
Error 5	TP	16.58(2.27)	6.20(2.91)	5.38(1.89)	7.85(4.37)	13.9(5.21)	4.75(3.16)
	FP	6.53(2.42)	146.23(12.86)	4.78(2.90)	3.28(2.69)	53.68(23.01)	4.93(3.23)
	F1	0.77(0.08)	0.07(0.03)	0.35(0.10)	0.47(0.22)	0.30(0.10)	0.30(0.17)
	MCC	0.77(0.08)	0.06(0.06)	0.37(0.10)	0.49(0.21)	0.35(0.12)	0.31(0.17)

Table A.3: *Results of Identifications for Simulation from Real Data Under Homogeneous Error Distributions: Dimension (n, p) = (1000,2000). mean(sd) of TP, FP, F1 and MCC based on 100 replicates.*

		LADL–SS	LADL	LADLP-SS	LASSO-SS	LASSO	LP-SS
Error 1	TP	19.80(0.52)	4.33(2.79)	1.75(1.25)	19.93(0.17)	20(0)	19.80(0.39)
	FP	5.65(2.17)	39.10(3.37)	28.18(5.35)	7.05(2.93)	82.21(15.72)	19.38(6.34)
	F1	0.87(0.05)	0.14(0.09)	0.07(0.05)	0.85(0.05)	0.33(0.04)	0.67(0.07)
	MCC	0.88(0.04)	0.13(0.09)	0.06(0.05)	0.86(0.05)	0.45(0.05)	0.71(0.06)
Error 2	TP	18.90 (1.28)	3.45 (2.75)	1.70 (1.22)	16.90 (1.81)	19.60 (0.67)	14.22 (2.48)
	FP	6.9 (3.36)	39.45 (2.84)	25.08 (6.08)	8.00 (3.45)	83.4 (15.26)	15.75 (5.08)
	F1	0.83 (0.08)	0.11 (0.09)	0.07 (0.05)	0.76 (0.09)	0.32 (0.04)	0.57 (0.09)
	MCC	0.83 (0.08)	0.10 (0.10)	0.06 (0.05)	0.76 (0.09)	0.42 (0.04)	0.58 (0.09)
Error 3	TP	19.1 (1.06)	3.68 (2.67)	1.60 (1.17)	11.28 (3.76)	17.58 (2.31)	8.70 (3.93)
	FP	8.18 (3.75)	57.90 (3.20)	27.13 (6.30)	6.33 (2.81)	73.7 (16.33)	12.85 (4.52)
	F1	0.81 (0.08)	0.09 (0.06)	0.07 (0.05)	0.59 (0.14)	0.32 (0.05)	0.40 (0.14)
	MCC	0.82 (0.07)	0.09 (0.08)	0.06 (0.05)	0.59 (0.14)	0.40 (0.06)	0.40 (0.15)
Error 4	TP	19.20 (1.07)	2.3 (2.6)	1.83 (1.38)	4.83 (5.67)	8.65 (8.21)	3.55 (4.58)
	FP	6.93 (2.35)	38.8 (2.89)	25.08 (7.67)	2.4 (3.05)	38.75 (37.26)	5.8 (6.31)
	F1	0.83 (0.05)	0.08 (0.09)	0.08 (0.06)	0.27 (0.29)	0.18 (0.14)	0.18 (0.20)
	MCC	0.84 (0.05)	0.07 (0.09)	0.07 (0.06)	0.29 (0.29)	0.22 (0.18)	0.19 (0.19)
Error 5	TP	17.00 (1.87)	2.83 (1.99)	1.1 (1.03)	5.35 (3.38)	12.13 (4.79)	3.75 (2.37)
	FP	6.45 (2.32)	28.25 (2.73)	158.48 (6.21)	2.95 (2.11)	53.68 (21.27)	7.03 (4.47)
	F1	0.78 (0.08)	0.11 (0.08)	0.05 (0.05)	0.35 (0.19)	0.28 (0.08)	0.23 (0.13)
	MCC	0.78 (0.08)	0.10 (0.08)	0.05 (0.05)	0.39 (0.18)	0.32 (0.11)	0.24 (0.13)

A.2.2 Autoregressive Structure under Heterogeneous Error

Table A.4: *Results of Identifications for Simulation Under Heterogeneous Error Distributions: Dimension (n, p) = (1000,2000). mean(sd) of TP, FP, F1 and MCC based on 100 replicates.*

		LADL-SS	LADL	LADLP-SS	LASSO-SS	LASSO	LP-SS
Error 1	TP	19.98(0.16)	20.00(0.00)	13.85(6.38)	19.95(0.32)	20.00(0.00)	18.73(1.36)
	FP	1.10(1.19)	88.80(30.14)	2.95(5.94)	9.85(3.07)	74.60(16.55)	1.00(1.22)
	F1	0.97(0.03)	0.33(0.08)	0.75(0.33)	0.80(0.05)	0.36(0.05)	0.94(0.04)
	MCC	0.97(0.03)	0.43(0.07)	0.75(0.33)	0.82(0.04)	0.46(0.04)	0.94(0.04)
Error 2	TP	17.10(1.89)	19.90(0.31)	7.60(2.11)	3.35(3.17)	10.25(4.47)	1.85(1.18)
	FP	0.89(0.07)	0.36(0.05)	0.54(0.11)	0.23(0.20)	0.33(0.10)	0.16(0.10)
	F1	0.89(0.07)	0.46(0.04)	0.60(0.09)	0.27(0.22)	0.37(0.08)	0.27(0.12)
	MCC	0.96(0.03)	0.36(0.08)	0.82(0.1)	0.83(0.05)	0.44(0.06)	0.9(0.05)
Error 3	TP	19.55(0.76)	19.15(0.93)	14.55(2.42)	7.05(4.22)	14.05(3.33)	3.95(2.48)
	FP	1.10(1.21)	2.50(1.10)	0.40(0.60)	2.40(2.80)	46.15(17.18)	0.70(0.86)
	F1	0.96(0.03)	0.92(0.03)	0.83(0.08)	0.45(0.16)	0.36(0.07)	0.31(0.15)
	MCC	0.96(0.03)	0.92(0.03)	0.84(0.07)	0.50(0.13)	0.40(0.08)	0.39(0.13)
Error 4	TP	19.80(0.52)	20.00(0.00)	14.55(2.46)	4.15(6.58)	6.60(8.17)	2.75(4.42)
	FP	1.40(1.57)	85.90(32.93)	0.50(0.76)	1.55(3.09)	25.25(36.79)	0.55(0.69)
	F1	0.96(0.04)	0.34(0.09)	0.83(0.08)	0.22(0.31)	0.16(0.16)	0.19(0.27)
	MCC	0.96(0.04)	0.44(0.08)	0.84(0.08)	0.24(0.32)	0.19(0.19)	0.22(0.28)
Error 5	TP	18.90(0.97)	20.00(0.00)	10.05(3.35)	3.60(4.10)	8.10(6.10)	2.15(2.28)
	FP	0.60(0.82)	80.60(38.76)	0.20(0.41)	1.45(1.79)	32.95(31.26)	0.40(0.368)
	F1	0.96(0.03)	0.36(0.11)	0.65(0.14)	0.24(0.25)	0.23(0.14)	0.17(0.17)
	MCC	0.96(0.03)	0.46(0.09)	0.69(0.12)	0.27(0.26)	0.26(0.15)	0.24(0.20)

A.2.3 Sensitivity analysis results

In this section, the following tables record results demonstrate the sensitivity of ratio q for creating subsamples and the threashold θ. Specifically, we considered: 1) two different ratio for which are $q = 70\%$ and 80%; 2) three different thresholds of $\theta = 50\%$, 80%, and 90%. Table A.5, A.6, A.7, A.8 show the homogeneous data structure while Table A.9 presents data for heterogeneous structure. Bolded entries in the tables denote the best performing values.

Table A.5: *Sensitivity analysis using subsamples with ratio q=70% and cutoff θ = (50%, 80%, 90%) under dimension (n, p) = (500,1000) with homogeneous data structure. mean(sd) of TP, FP, F1 and MCC based on 100 replicates.*

		LADL-SS			LASSO-SS		
		50%	80%	90%	50%	80%	90%
Error 1	TP	**19.90(0.20)**	19.90(0.30)	19.80(0.40)	20.00(0.00)	19.90(0.20)	19.80(0.40)
	FP	80.10 (16.70)	11.40 (3.70)	**3.60 (2.00)**	42.80 (5.50)	4.70 (2.00)	1.10 (1.00)
	F1	0.34 (0.05)	0.78 (0.06)	**0.92 (0.04)**	0.49 (0.03)	0.90 (0.04)	0.97 (0.03)
	MCC	0.43 (0.04)	0.79 (0.05)	**0.92 (0.04)**	0.55 (0.03)	0.90 (0.04)	0.97 (0.03)
Error 2	TP	**19.40 (0.80)**	17.70 (1.90)	15.60 (2.40)	18.60 (1.30)	15.80 (2.10)	13.80 (2.50)
	FP	42.10 (15.80)	5.10 (3.00)	**1.70 (1.30)**	49.80 (18.20)	8.70 (4.80)	2.80 (2.10)
	F1	0.49 (0.10)	0.83 (0.07)	**0.83 (0.08)**	0.44 (0.08)	0.71 (0.07)	0.75 (0.08)
	MCC	0.56 (0.08)	0.83 (0.07)	**0.83 (0.07)**	0.50 (0.06)	0.71 (0.07)	0.75 (0.08)
Error 3	TP	**19.10 (1.00)**	16.30 (2.50)	13.50 (3.40)	13.00 (3.80)	8.30 (4.40)	5.90 (4.10)
	FP	24.92 (10.10)	2.31 (1.61)	**0.50 (0.70)**	28.8 (10.21)	4.52 (2.90)	1.6 (1.70)
	F1	0.61 (0.09)	**0.84 (0.08)**	0.78 (0.12)	0.42 (0.11)	0.47 (0.22)	0.39 (0.23)
	MCC	0.65 (0.07)	**0.84 (0.08)**	0.80 (0.11)	0.44 (0.1)	0.48 (0.22)	0.43 (0.23)
Error 4	TP	**19.81 (0.38)**	19.31 (1.01)	18.60 (1.12)	7.80 (6.60)	4.40 (6.32)	3.40 (6.10)
	FP	41.00 (13.31)	6.90 (3.50)	**2.40 (1.80)**	29.30 (26.11)	3.40 (5.21)	1.10 (2.20)
	F1	0.50 (0.08)	0.84 (0.06)	**0.91 (0.05)**	0.23 (0.15)	0.22 (0.26)	0.19 (0.29)
	MCC	0.56 (0.06)	0.85 (0.05)	**0.91 (0.05)**	0.25 (0.16)	0.23 (0.27)	0.21 (0.30)
Error 5	TP	**18.70 (1.40)**	16.10 (2.61)	14.00 (2.91)	11.10 (4.50)	6.21 (4.22)	4.33 (3.82)
	FP	36.21 (13.32)	5.00 (3.51)	**1.60 (1.60)**	28.11 (11.90)	3.91 (3.20)	1.10 (1.40)
	F1	0.51 (0.07)	**0.78 (0.09)**	0.78 (0.11)	0.36 (0.12)	0.37 (0.22)	0.29 (0.25)
	MCC	0.56 (0.06)	0.78 (0.09)	**0.79 (0.10)**	0.37 (0.12)	0.38 (0.22)	0.32 (0.26)

Table A.6: *Sensitivity analysis using subsamples with ratio q=70% and cutoff θ = (50%, 80%, 90%) under dimension (n, p) = (1000,2000) with homogeneous data structure. mean(sd) of TP, FP, F1 and MCC based on 100 replicates.*

		LADL-SS			LASSO-SS		
		50%	80%	90%	50%	80%	90%
Error 1	TP	**20.00 (0.00)**	**20.00 (0.00)**	19.90 (0.20)	20.00 (0.00)	20.00 (0.00)	19.90 (0.20)
	FP	63.50 (12.10)	9.31 (3.80)	**3.12 (1.73)**	60.61 (8.24)	6.10 (2.61)	1.50 (1.00)
	F1	0.39 (0.05)	0.82 (0.06)	**0.93 (0.04)**	0.40 (0.03)	0.87 (0.05)	0.96 (0.02)
	MCC	0.49 (0.04)	0.83 (0.05)	**0.93 (0.03)**	0.49 (0.03)	0.88 (0.04)	0.96 (0.02)
Error 2	TP	**19.90 (0.20)**	19.90 (0.21)	19.90 (0.30)	19.91 (0.22)	19.90 (0.31)	19.71 (0.63)
	FP	35.50 (9.21)	4.20 (2.00)	**1.61 (1.21)**	138.11(19.61)	25.2 (6.61)	8.41 (3.62)
	F1	0.54 (0.06)	0.90 (0.04)	**0.96 (0.03)**	0.23 (0.03)	0.62 (0.07)	0.82 (0.06)
	MCC	0.60 (0.05)	0.91 (0.04)	**0.96 (0.03)**	0.34 (0.03)	0.66 (0.05)	0.83 (0.05)
Error 3	TP	**19.90 (0.16)**	19.90 (0.30)	19.90 (0.40)	18.70 (1.60)	16.70 (2.90)	15.10 (3.70)
	FP	21.10 (5.00)	3.00 (1.70)	**1.00 (0.90)**	68.40 (36.50)	11.90 (8.30)	3.90 (3.50)
	F1	0.66 (0.05)	0.93 (0.04)	**0.97 (0.03)**	0.38 (0.09)	0.70 (0.10)	0.76 (0.12)
	MCC	0.70 (0.04)	0.93 (0.03)	**0.97 (0.03)**	0.46 (0.08)	0.71 (0.09)	0.77 (0.11)
Error 4	TP	**19.90 (0.20)**	19.90 (0.30)	19.60 (0.60)	9.80 (7.30)	7.00 (7.70)	5.70 (7.30)
	FP	24.10 (7.30)	3.80 (2.10)	**1.40 (1.10)**	43.40 (45.40)	6.10 (8.40)	1.70 (2.70)
	F1	0.63 (0.07)	0.91 (0.04)	**0.96 (0.03)**	0.23 (0.15)	0.30 (0.30)	0.30 (0.36)
	MCC	0.67 (0.06)	0.91 (0.04)	**0.96 (0.03)**	0.28 (0.17)	0.31 (0.31)	0.30 (0.37)
Error 5	TP	**19.80 (0.50)**	19.40 (0.81)	18.70 (1.20)	14.70 (3.70)	10.30 (4.60)	7.80 (4.70)
	FP	25.70 (7.60)	3.40 (1.90)	**1.30 (1.40)**	46.40 (14.80)	6.70 (3.30)	2.10 (1.70)
	F1	0.61 (0.07)	0.91 (0.05)	**0.94 (0.05)**	0.37 (0.09)	0.53 (0.21)	0.47 (0.25)
	MCC	0.66 (0.06)	0.91 (0.04)	**0.94 (0.05)**	0.42 (0.09)	0.53 (0.21)	0.50 (0.25)

Table A.7: *Sensitivity analysis using subsamples with ratio q=80% and cutoff θ = (50%, 80%, 90%) under dimension (n, p) = (500,1000) with homogeneous data structure. mean(sd) of TP, FP, F1 and MCC based on 100 replicates.*

		LADL-SS			LASSO-SS		
		50%	80%	90%	50%	80%	90%
Error 1	TP	**20.00 (0.00)**	19.90 (0.20)	19.90 (0.30)	20.00 (0.00)	20.00 (0.00)	19.90 (0.20)
	FP	100.00(18.50)	23.80 (6.00)	**10.20 (3.40)**	102.20(10.70)	31.90 (6.00)	15.70 (3.50)
	F1	0.29 (0.04)	0.63 (0.06)	**0.80 (0.05)**	0.28 (0.06)	0.56 (0.05)	0.72 (0.05)
	MCC	0.39 (0.04)	0.67 (0.05)	**0.81 (0.05)**	0.38 (0.05)	0.61 (0.04)	0.74 (0.04)
Error 2	TP	**19.70 (0.60)**	18.80 (1.10)	17.80 (1.80)	19.10 (1.00)	17.60 (1.80)	16.30 (1.90)
	FP	62.80 (18.80)	13.40 (5.20)	**5.30 (2.50)**	92.80 (27.90)	25.70 (11.60)	11.80 (6.10)
	F1	0.40 (0.08)	0.73 (0.08)	**0.83 (0.07)**	0.30 (0.07)	0.57 (0.09)	0.69 (0.07)
	MCC	0.48 (0.07)	0.74 (0.07)	**0.83 (0.07)**	0.39 (0.06)	0.60 (0.07)	0.69 (0.06)
Error 3	TP	**19.60 (0.70)**	18.10 (1.80)	16.90 (2.10)	15.10 (3.90)	11.60 (5.20)	9.80 (4.90)
	FP	38.60 (14.60)	6.70 (3.60)	**2.50 (1.70)**	48.90 (18.20)	12.70 (7.10)	5.80 (3.90)
	F1	0.52 (0.09)	0.81 (0.07)	**0.86 (0.07)**	0.36 (0.09)	0.49 (0.21)	0.52 (0.22)
	MCC	0.57 (0.07)	0.81 (0.07)	**0.86 (0.07)**	0.41 (0.10)	0.49 (0.21)	0.52 (0.22)
Error 4	TP	**19.90 (0.30)**	19.60 (0.70)	19.30 (1.10)	8.30 (6.80)	6.10 (6.70)	4.70 (6.60)
	FP	53.80 (16.20)	12.90 (5.10)	**5.80 (3.40)**	54.00 (42.80)	12.50 (15.02)	4.40 (6.90)
	F1	0.44 (0.07)	0.75 (0.07)	**0.86 (0.06)**	0.16 (0.11)	0.23 (0.2)	0.21 (0.26)
	MCC	0.51 (0.06)	0.77 (0.06)	**0.86 (0.06)**	0.18 (0.14)	0.24 (0.22)	0.22 (0.26)
Error 5	TP	**19.10 (1.10)**	17.70 (2.00)	16.40 (2.50)	12.30 (4.40)	8.60 (4.20)	6.40 (4.60)
	FP	51.30 (15.30)	12.60 (5.10)	**5.60 (2.90)**	47.70 (16.30)	10.60 (6.90)	4.50 (3.80)
	F1	0.43 (0.06)	0.71 (0.08)	**0.78 (0.09)**	0.30 (0.10)	0.42 (0.18)	0.36 (0.25)
	MCC	0.50 (0.05)	0.72 (0.07)	**0.78 (0.09)**	0.33 (0.12)	0.42 (0.17)	0.36 (0.25)

Table A.8: *Sensitivity analysis using subsamples with ratio q=80% and cutoff θ = (50%, 80%, 90%) under dimension (n, p) = (1000,2000) with homogeneous data structure. mean(sd) of TP, FP, F1 and MCC based on 100 replicates.*

		LADL-SS			LASSO-SS		
		50%	80%	90%	50%	80%	90%
Error 1	TP	**20.00 (0.00)**	20.00 (0.00)	19.90 (0.20)	20.00 (0.00)	20.00 (0.00)	20.00 (0.00)
	FP	79.30 (12.70)	19.50 (4.20)	**8.20 (2.80)**	109.30(22.40)	26.80 (12.60)	11.40 (7.10)
	F1	0.34 (0.04)	0.68 (0.05)	**0.83 (0.05)**	0.27 (0.04)	0.62 (0.11)	0.79 (0.11)
	MCC	0.44 (0.03)	0.71 (0.04)	**0.84 (0.04)**	0.39 (0.04)	0.67 (0.08)	0.81 (0.09)
Error 2	TP	**19.90 (0.20)**	19.90 (0.20)	19.90 (0.20)	19.90 (0.20)	19.90 (0.30)	19.90 (0.40)
	FP	53.40 (13.40)	11.20 (3.00)	**4.50 (2.10)**	210.00(17.50)	63.50 (10.60)	29.40 (6.50)
	F1	0.44 (0.06)	0.78 (0.05)	**0.90 (0.04)**	0.16 (0.01)	0.39 (0.04)	0.58 (0.05)
	MCC	0.52 (0.05)	0.80 (0.04)	**0.90 (0.04)**	0.28 (0.01)	0.48 (0.04)	0.63 (0.04)
Error 3	TP	**19.90 (0.20)**	19.90 (0.20)	19.90 (0.30)	19.20 (1.30)	18.00 (2.30)	17.30 (2.70)
	FP	30.50 (8.30)	6.30 (3.30)	**2.80 (1.90)**	125.10(62.80)	32.50 (19.40)	15.10 (10.10)
	F1	0.57 (0.06)	0.87 (0.06)	**0.94 (0.04)**	0.27 (0.09)	0.54 (0.11)	0.67 (0.11)
	MCC	0.63 (0.05)	0.87 (0.05)	**0.94 (0.04)**	0.37 (.08)	0.58 (0.09)	0.69 (0.1)
Error 4	TP	**19.90 (0.20)**	19.90 (0.20)	19.90 (0.30)	10.80 (7.70)	8.20 (8.00)	7.80 (7.80)
	FP	35.6 (10.10)	8.10 (3.50)	**3.80 (2.00)**	80.00 (71.70)	18.70 (22.50)	7.50 (9.60)
	F1	0.54 (0.07)	0.83 (0.06)	**0.91 (0.04)**	0.17 (0.12)	0.25 (0.24)	0.33 (0.31)
	MCC	0.60 (0.06)	0.84 (0.05)	**0.92 (0.04)**	0.22 (0.15)	0.26 (0.25)	0.33 (0.32)
Error 5	TP	**19.90 (0.50)**	19.70 (0.60)	19.40 (0.80)	15.70 (3.00)	12.70 (4.60)	11.20 (4.50)
	FP	39.30 (9.40)	7.80 (2.50)	**3.40 (1.80)**	76.50 (24.40)	20.00 (8.40)	8.60 (4.10)
	F1	0.51 (0.06)	0.83 (0.04)	**0.91 (0.04)**	0.29 (0.07)	0.46 (0.18)	0.54 (0.2)
	MCC	0.58 (0.05)	0.84 (0.04)	**0.91 (0.04)**	0.36 (0.08)	0.48 (0.18)	0.54 (0.20)

Table A.9: *Sensitivity analysis using subsamples with ratio q=70% and cutoff θ = (50%, 80%, 90%) under dimension (n, p) = (1000,2000) with heterogeneous data structure. mean(sd) of TP, FP, F1 and MCC based on 100 replicates.*

		LADL-SS			LASSO-SS		
		50%	80%	90%	50%	80%	90%
Error 1	TP	**20.00(0.00)**	19.98(0.16)	19.98(0.16)	20.00(0.00)	19.98(0.16)	19.95(0.32)
	FP	28.53(8.90)	3.95(2.15)	**1.10(1.19)**	133.28(12.54)	26.80(5.86)	9.85(3.07)
	F1	0.59(0.07)	0.91(0.04)	**0.97(0.03)**	0.23(0.02)	0.60(0.05)	0.80(0.05)
	MCC	0.65(0.06)	0.91(0.04)	**0.97(0.03)**	0.35(0.02)	0.65(0.04)	0.82(0.04)
Error 2	TP	**19.75(0.44)**	18.60(1.50)	17.10(1.89)	12.10(3.06)	6.50(3.85)	3.35(3.17)
	FP	22.85(8.71)	3.55(2.67)	**1.15(1.18)**	44.90(16.25)	5.40(3.90)	1.55(1.82)
	F1	0.64(0.09)	0.88(0.07)	**0.89(0.07)**	0.32(0.06)	0.38(0.21)	0.23(0.20)
	MCC	0.68(0.07)	**0.89(0.07)**	**0.89(0.07)**	0.35(0.07)	0.38(0.20)	0.27(0.22)
Error 3	TP	**19.9(0.31)**	19.75(0.44)	19.55(0.76)	13.85(3.84)	9.45(3.87)	7.05(4.22)
	FP	18.35(5.12)	3.25(1.86)	**1.10(1.21)**	48.20(26.94)	7.00(6.27)	2.40(2.80)
	F1	0.69(0.07)	0.92(0.04)	**0.96(0.03)**	0.35(0.08)	0.50(0.13)	0.45(0.16)
	MCC	0.72(0.06)	0.92(0.04)	**0.96(0.03)**	0.40(0.08)	0.52(0.12)	0.50(0.13)
Error 4	TP	**19.95(0.22)**	19.80(0.52)	19.80(0.52)	7.35(7.75)	5.05(6.93)	4.15(6.58)
	FP	23.35(7.97)	3.15(2.32)	**1.40(1.57)**	37.65(42.09)	5.00(8.10)	1.55(3.09)
	F1	0.64(0.08)	0.92(0.05)	**0.96(0.04)**	0.16(0.15)	0.23(0.27)	0.22(0.31)
	MCC	0.68(0.07)	0.93(0.05)	**0.96(0.04)**	0.19(0.18)	0.25(0.28)	0.24(0.32)
Error 5	TP	**19.85(0.37)**	19.30(0.92)	18.90(0.97)	9.70(5.05)	5.50(4.61)	3.60(4.10)
	FP	19.85(6.93)	2.10(1.83)	**0.60(0.82)**	43.05(18.65)	5.25(4.23)	1.45(1.79)
	F1	0.67(0.07)	0.93(0.05)	**0.96(0.03)**	0.25(0.12)	0.31(0.23)	0.24(0.25)
	MCC	0.71(0.06)	0.93(0.05)	**0.96(0.03)**	0.28(0.14)	0.32(0.23)	0.27(0.26)

Appendix B

Estimations for Real Data Analysis

B.0.1 Skin Cutaneous Melanoma

Table B.1: Analysis of the SKCM data using LADL-SS.

Gene*	LASSO	LASSO-SS	LP-SS	LADL	LADL-SS	LADLP-SS
ALDH4A1				0.013		
ANTXR2				-0.018		
ATAD2B				0.165		
BOLA1		-0.117				
BTNL9				-0.148		-0.174
C13orf21	0.100	0.142	0.163		0.171	
C14orf61	-0.025			-0.101	-0.143	-0.127
C1ORF216	0.126					
CCDC142					-0.036	-0.071
CDKN1A	0.044	0.062	0.053			
COQ9		-0.117			-0.318	
DKFZP586C1324	0.117	0.131	0.142		0.103	

Continued on the next page

Table B.1: Continued from the previous page.

Gene*	LASSO	LASSO-SS	LP-SS	LADL	LADL-SS	LADLP-SS
D SE1L1				0.142		
DNLC2B	-0.186	-0.171	-0.170			
DOCK11						0.014
DSTYK				0.129		
EBF2	-0.145		-0.270			
FAM222A-AS1	0.100					
FAM49A				0.010		-0.023
FLJ13621				-0.051		
FLJ20373	-0.084					
FMNL2	-0.043			-0.168	-0.161	
H2A	-0.177					
HBD	-0.077		-0.131			
HPE2		0.073				
IER5	0.111	0.082			0.027	
INPP5K				0.159	-0.003	0.113
JMS				-0.126		-0.147
KBF2	-0.053	-0.126	-0.134			-0.219
KCNG4	-0.065					
KCTD16	-0.095		-0.018			
LBR						-0.040
LINC00847	-0.132					
LOC100240735	0.035					
LOC100506644	-0.032			-0.127		-0.045

Continued on the next page

Table B.1: Continued from the previous page.

Gene*	LASSO	LASSO-SS	LP-SS	LADL	LADL-SS	LADLP-SS
LOC105371602	-0.031		-0.084			
LOC115722				0.098	0.032	0.077
LOC116242				0.029		0.119
LOC148709	0.101	0.149				
LOC90374					0.087	
LOC91492		0.049				
LOC92224				0.027		
MYO1C						0.000
NEIL2				0.048		
PARD6G				0.018	-0.039	0.044
PER1		-0.182				
PITP	0.151	0.135	0.122		0.146	
QRSL1	-0.026					
RABIF				-0.064		
RBFA					0.029	
RYBP	-0.027	-0.161	-0.095			
SDPR	-0.045			-0.106	-0.148	-0.112
SEL1L3				-0.182	-0.094	
SEMA3A	-0.135	-0.307	-0.147		-0.369	
SHF		0.085				
SLC8A1	-0.037			0.041		0.007
SMPX	-0.070		-0.092			
SP3				0.136		0.157

Continued on the next page

Table B.1: Continued from the previous page.

Gene*	LASSO	LASSO-SS	LP-SS	LADL	LADL-SS	LADLP-SS
STAG2				0.058		0.009
TDRG1	-0.210	-0.271				
TFR2	0.097				0.099	
TMEM159	0.134	0.160	0.177			
TTC28-AS1	-0.062		-0.359			
ZEB1				-0.125		-0.091
ZFHX3				-0.084		
ZNF41				0.119		0.145
ZNF586				-0.067		-0.034

B.0.2 Human Breast Tumor

Table B.2: Analysis of the Human Breast Tumor data using LADL-SS.

Gene*	LASSO	LASSO-SS	LP-SS	LADL	LADL-SS	LADLP-SS
AARSD1						-0.014
ANKRD13B			0.010			
AP3M1		0.055				
ARID5B		0.037				
ASPM						0.004
BECN1			0.050			
BUB1B				0.007		
C10orf76	0.076	0.078	0.071			
C17orf53	-0.003		-0.017	0.003	0.018	-0.025
C1orf135						0.019
CASC5						0.027
CBX3						0.008
CCDC43				0.021		-0.001
CCDC56	0.089		0.036	0.113	0.019	0.040
CDC25C	-0.045			0.019		0.017
CDC6	0.015		-0.012			
CENPK	0.057		0.034	0.030	0.019	-0.007
CENPM	0.044		0.017	0.022	0.022	0.037
CENPQ	0.050	0.040	0.052	0.043	0.068	0.054
CEP152					0.021	
CHTF18		0.057	0.021			0.055

Continued on the next page

Table B.2: Continued from the previous page.

Gene*	LASSO	LASSO-SS	LP-SS	LADL	LADL-SS	LADLP-SS
CIT	0.041		-0.001			
CMTM5	-0.064	-0.064	-0.067			
COPS6					0.045	0.042
COQ5	0.032		0.026			
DARS2			0.020			
DDX52		0.042	0.034			
DTL	0.059	0.083	0.058	0.058	0.079	0.064
EFTUD2				-0.034		-0.040
EIF1B	-0.080	-0.091	-0.072		-0.105	
FST		0.054				
GCN5L2				-0.002	0.018	-0.020
GINS1	0.045		0.012			
HCN3				0.049		0.025
HECA			-0.039			
HISPPD2A			0.048			
HMGN2		0.078			0.047	
HMMR				0.037		0.036
KHDRBS1	0.093	0.064	0.071		0.068	0.069
KIAA0101	-0.002			0.043	0.012	0.008
KLHL13	-0.054	-0.050	-0.057		-0.071	-0.088
KPNB1					0.022	0.054
LIG1						0.005
LMNB1					0.024	0.031

Continued on the next page

Table B.2: Continued from the previous page.

Gene*	LASSO	LASSO-SS	LP-SS	LADL	LADL-SS	LADLP-SS
LSM12	0.024		-0.008	0.079	-0.014	0.023
MCM6		0.067	0.088	0.073	0.038	0.038
MFGE8	-0.011					
MKI67						-0.011
MND1		0.053				
NBR1		0.071			0.059	0.049
NBR2	0.216	0.186	0.203	0.260	0.207	0.233
NCAPG2				-0.057		
NFATC1	-0.047	-0.052	-0.043			
NKAPL				-0.019		0.021
NLE1			0.014			
NMT1				0.012	-0.001	0.014
NUDT1					0.046	
OSR1		-0.058			-0.088	
PAICS		-0.049				
PCDHB16						-0.027
PLEKHH3					0.021	0.023
PLSCR4		0.055				
PSME3	0.044	0.059	0.023	0.051	0.052	0.046
RAB11FIP4		0.026				
RDM1	-0.014			0.005		-0.011
RPL27		0.116	0.113		0.096	0.081
RRM1				-0.016		

Continued on the next page

Table B.2: Continued from the previous page.

Gene*	LASSO	LASSO-SS	LP-SS	LADL	LADL-SS	LADLP-SS
SLC25A22	0.016		0.028			
SPAG5	0.043		0.008			
SPINK5L3	0.043	0.045	0.032		0.054	0.060
SUZ12				0.019		
TIMELESS	-0.047			0.022		-0.009
TMPRSS4		0.039	0.043			
TNFRSF11B		-0.049				
TOP2A	0.045	0.063	0.050	0.060		0.045
TRAIP	0.040		-0.006	0.038		0.000
TSPYL2		-0.055	-0.024			
TUBA1B	0.039	0.032	0.054	0.064	0.049	0.060
UHRF1	-0.004			0.013		0.002
VDAC1	0.030		0.015			
VPS25	0.050		-0.004		0.033	0.077
WDR51A			0.021	0.045		0.028
WDR76				-0.045		
XRCC2	0.067	0.100	0.057		0.054	0.029
ZNF189		-0.069	-0.038			