ABSTRACT

PENG, HUIMIN. Selection and Inference for High-Dimensional Regression with Applications in Biomedical Research. (Under the direction of Wenbin Lu and Xinge Jessie Jeng.)

High-dimensional inference of penalized regression has been a challenging problem. In high-dimensional setting, penalized regression method can be applied to conduct variable selection and parameter estimation simultaneously. In high-dimensional inference, we should perform hypothesis test and compute confidence interval accounting for randomness in the selected model. Desparsifying lasso approach and covariance test are among the most important methods to construct confidence interval and to perform hypothesis testing in high-dimensional penalized regression.

Our first project is focused upon the statistical inference for high-dimensional regression based upon lasso solution path. Under null model, covariance test statistics can be transformed into Q statistics distributed as ordered i.i.d. p values under certain conditions. Previously Benjamini-Hochberg procedure has been applied to Q statistics but it shows high false positive proportion in our simulation for high-dimensional setting. For Benjamini-Hochberg procedure to achieve FDR control, it is assumed that perfect separation holds with high probability and that true variable intensity exceeds a minimum. In order to achieve adaptivity in performance, we propose a model size estimation procedure based upon Q statistics. We obtain an upper and a lower bound of the estimated model size under less assumptions and prove that our procedure has selection consistency when true variables are of sufficiently strong intensity. We study the performance of our procedure in comparison with other methods in simulation and apply our procedure to the eQTL dataset concerning Down Syndrome.

Our second project studies high-dimensional inference for personalized treatment deci-
sion. Precision medicine aims to design optimal treatment for each individual according to their specific conditions. Some recent development in statistical inference for personalized treatment focuses on penalized variable selection, where a large number of patients’ covariates are considered in a regression model. Personalized treatment decision is described through interactions between treatment and covariates. Although a subset of interaction terms can be retained by existing variable selection methods, there is a lack of method and theory on confidence interval and hypothesis testing on the interaction coefficients. We propose a bias-corrected estimator based on lasso solution for the interaction coefficients. We derive the limiting distribution of the bias-corrected estimator when baseline function of the regression model is unknown and possibly misspecified. Confidence intervals and p values are calculated using the limiting distribution to make personalized treatment decision. We illustrate the finite sample performance of the proposed method in simulation and apply the method to STAR*D study for major depression disorder.
Selection and Inference for High-Dimensional Regression with Applications in Biomedical Research

by

Huimin Peng

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Statistics

Raleigh, North Carolina

2017

APPROVED BY:

Jung-Ying Tzeng
Howard Bondell

Wenbin Lu
Co-chair of Advisory Committee
Xinge Jessie Jeng
Co-chair of Advisory Committee
DEDICATION

To my parents, my advisors and my friends.
BIOGRAPHY

I was from Jiangling county, Jingzhou city in Hubei province, China. I completed my high school education in Jingzhou city in 2008. I completed my bachelor’s degree of hydraulic engineering and economics from Tsinghua University in 2012. I entered the statistics PhD program at NC State University in 2012. In 2014, I received my en-route master’s degree of mathematical statistics.
ACKNOWLEDGEMENTS

I would like to thank my committee and my advisors for their help.
TABLE OF CONTENTS

| LIST OF TABLES | vii |
| LIST OF FIGURES | ix |

Chapter 1 Introduction ......................................................... 1
  1.1 Background ................................................................. 1
  1.2 Penalized regression method .......................................... 4
  1.3 Inference in high-dimensional regression ...................... 10
    1.3.1 Desparsifying lasso approach .................................. 11
    1.3.2 Covariance test statistic ...................................... 13
  1.4 Precision medicine ...................................................... 14
  1.5 Applications ............................................................... 22
    1.5.1 Description of eQTL dataset .................................... 22
    1.5.2 Description of STAR*D study .................................... 24
  1.6 Project overview ......................................................... 25

Chapter 2 Statistical inference for high-dimensional regression: Post-lasso solution path inference ......................................................... 28
  2.1 Introduction ................................................................. 29
    2.1.1 Covariance test statistic and Q statistic .................... 30
    2.1.2 Previous method based upon Q statistic ...................... 33
  2.2 QVS procedure: Model size estimation .......................... 35
  2.3 Statistical property of QVS procedure ............................ 37
    2.3.1 Upper bound for QVS estimated model size .................. 37
    2.3.2 Lower bound for QVS estimated model size ................. 44
  2.4 Selection consistency of QVS procedure .......................... 48
  2.5 Simulation ................................................................. 50
    2.5.1 Simulation setting .................................................. 50
    2.5.2 Upper and lower bound of QVS estimated model size ........ 53
    2.5.3 Comparison of QVS performance with other method .......... 53
  2.6 Application to eQTL dataset concerning Down Syndrome .......... 63

Chapter 3 High-dimensional inference for personalized treatment decision ......................................................... 66
  3.1 Introduction ................................................................. 66
  3.2 Method and Theory ......................................................... 69
    3.2.1 Desparsified Lasso estimator with misspecified baseline function .... 69
    3.2.2 Limiting distribution of the desparsified Lasso estimator .......... 71
    3.2.3 Interval estimation and hypothesis testing ..................... 85
    3.2.4 Some special case of \( \hat{\mu}_\pi \) ................................. 86
LIST OF TABLES

Table 2.1  Frequency of QVS estimated model size located between lower and
upper bound when true variable intensity $\beta^* = 0.3$ and sample size
$n = 200$. Dimension $p = 2,000$ or $p = 10,000$. Number of true variables
$s$ is 10, 20, 30 or 40. TPP is true positive proportion. Standard errors
are in parenthesis. $m_0$ is the entry right before first noise variable.
$m_1$ is the entry of last true variable. $F(m_0 < \hat{G} \leq m_1)$ is frequency of
QVS estimated model size located between upper and lower bound.
$F(m_0 < \hat{G})$ is frequency of QVS estimated model size greater than
lower bound. $F(\hat{G} \leq m_1)$ is frequency of QVS estimated model size
less than upper bound. ................................................................. 53

Table 2.2  A table of comparison between QVS and popular methods when sam-
ple size $n = 200$ and dimension $p = 10,000$. Covariance matrix of
design matrix follows AR structure. Number of true variables $s$ = 10, 20
or 40. True variable intensity $\beta^* = 0.3$ or 0.6. $\hat{G}$ is QVS estimated model
size. TPP is true positive proportion. $s^*$ is number of true variables
covered in lasso solution path. Standard errors are in parenthesis. . . . . 54

Table 2.3  Table of covariance test statistics, p values and Q statistics along lasso
solution path for eQTL data concerning Down Syndrome. There are
three populations: European population, Chinese and Japanese pop-
ulation and population in Yoruba. For each population, covariance
test statistics, p values and Q statistics are displayed. P values are from
covariance test. ................................................................. 64

Table 3.1  Coverage of confidence interval in (Eq. 3.55) based upon limiting dis-
tribution of desparsified estimator and mean absolute bias of despar-
sified estimator in (Eq. 3.5) when $p = 300$ and $s = 5$. True variable
intensity is 1 or 1.5. Case 1 is simulated from linear baseline model.
Case 2 is simulated from quadratic baseline model. Case 3 is simulated
from trigonometric baseline model. Standard errors are in parenthesis. 92

Table 3.2  Coverage of confidence interval in (Eq. 3.55) based upon limiting dis-
tribution of desparsified estimator and mean absolute bias of despar-
sified estimator in (Eq. 3.5) when $p = 300$ and $s = 10$. True variable
intensity is 1 or 1.5. Case 1 is simulated from linear baseline model.
Case 2 is simulated from quadratic baseline model. Case 3 is simulated
from trigonometric baseline model. Standard errors are in parenthesis. 92
Table 3.3 Coverage of confidence interval in (Eq. 3.55) based upon limiting distribution of desparsified estimator and mean absolute bias of desparsified estimator in (Eq. 3.5) when dimension $p = 100$ and number of true variables $s = 5$. True variable intensity is 1 or 1.5. Case 1 is simulated from linear baseline model. Case 2 is simulated from quadratic baseline model. Case 3 is simulated from trigonometric baseline model. Standard errors are in parenthesis.

Table 3.4 Coverage of confidence interval in (Eq. 3.55) based upon limiting distribution of desparsified estimator and mean absolute bias of desparsified estimator in (Eq. 3.5) when dimension $p = 100$ and number of true variables $s = 10$. True variable intensity is 1 or 1.5. Case 1 is simulated from linear baseline model. Case 2 is simulated from quadratic baseline model. Case 3 is simulated from trigonometric baseline model. Standard errors are in parenthesis.

Table 3.5 Coverage of confidence interval in (Eq. 3.55) based upon limiting distribution of desparsified estimator and mean absolute bias of desparsified estimator in (Eq. 3.5) when dimension $p = 200$ and number of true variables $s = 5$. True variable intensity is 1 or 1.5. Case 1 is simulated from linear baseline model. Case 2 is simulated from quadratic baseline model. Case 3 is simulated from trigonometric baseline model. Standard errors are in parenthesis.

Table 3.6 Coverage of confidence interval in (Eq. 3.55) based upon limiting distribution of desparsified estimator and mean absolute bias of desparsified estimator in (Eq. 3.5) when dimension $p = 200$ and number of true variables $s = 10$. True variable intensity is 1 or 1.5. Case 1 is simulated from linear baseline model. Case 2 is simulated from quadratic baseline model. Case 3 is simulated from trigonometric baseline model. Standard errors are in parenthesis.

Table 3.7 Table of p values for significant interaction between treatment and patients’ covariates. P values are from the limiting distribution of desparsified lasso estimator. Variables are treatment interactions with the variables. The meaning of these variables are as follows. Qccur_r_rate: QIDS-C score changing rates. URNONE: no symptoms in patients’ urination category. NVTRM: Tremors. IMPWR: indicating whether patients thought they have special powers. hWL: hRS Weight loss. DSMTD: Recurrent thoughts of death, recurrent suicidal ideation, or suicide attempt. hMNIN: hRS Middle insomnia. EMSTU: Did you worry a lot that you might do something to make people think that you were stupid or foolish? [18] P values are computed using limiting distribution of desparsified lasso estimator.
LIST OF FIGURES

Figure 1.1 An example of lasso solution path. Horizontal axis is $\|\xi\|_1$ and also is $\phi$. Vertical axis is $\hat{\xi}_1, \cdots, \hat{\xi}_5$ respectively. There are 5 variables. Light blue, green, dark purple, red and black variables are added to the selected set sequentially. ................................................................. 7

Figure 2.1 An example of $m_0$ and $m_1$ in lasso solution path. Variables are ordered along lasso solution path. $m_0$ is the entry right before first noise variable. $m_1$ is the entry of last true variable. ................................. 35

Figure 2.2 Comparison of QVS with popular methods when sample size $n = 200$ and dimension $p = 100$. Number of true variables $s = 5, 10$ or 20. Covariance matrix of design matrix follows AR structure. Vertical axis is true positive proportion (TPP), false positive proportion (FPP), g-measure and false discovery proportion (FDP) respectively. Horizontal axis is true variable intensity. ................................. 56

Figure 2.3 Comparison of QVS with popular methods when sample size $n = 200$ and dimension $p = 400$. Number of true variables $s = 10, 20$ or 40. Covariance matrix of design matrix follows AR structure. Vertical axis is true positive proportion (TPP), false positive proportion (FPP), g-measure and false discovery proportion (FDP) respectively. Horizontal axis is true variable intensity. ................................. 58

Figure 2.4 Comparison of QVS with popular methods when sample size $n = 200$ and dimension $p = 2,000$. Number of true variables $s = 10, 20$ or 40. Covariance matrix of design matrix follows AR structure. Vertical axis is true positive proportion (TPP), false positive proportion (FPP), g-measure and false discovery proportion (FDP) respectively. Horizontal axis is true variable intensity. ................................. 59

Figure 2.5 Comparison of QVS with popular methods when sample size $n = 200$ and dimension $p = 10,000$. Number of true variables $s = 10, 20$ or 40. Covariance matrix of design matrix follows AR structure. Vertical axis is true positive proportion (TPP), false positive proportion (FPP), g-measure and false discovery proportion (FDP) respectively. Horizontal axis is true variable intensity. ................................. 62

Figure 3.1 The 95% confidence interval of selected interaction in Table 3.7. The horizontal axis is the index of variables in Table 3.7. The vertical axis is the desparsified lasso estimate. ................................. 98
Figure 3.2  Boxplot of average IPWE from bootstrap samples. We compare the estimated optimal personalized treatment strategy with the strategy of assigning only SER and the strategy of assigning only BUP. Horizontal axis is average clinical outcome. In vertical axis, 'optimal' is estimated optimal personalized treatment regime. 'all.BUP' is the strategy of assigning only BUP. 'all.SER' is the strategy of assigning only SER.

Figure 3.3  Boxplot of average IPWE from cross validation testing samples. We compare the estimated optimal personalized treatment strategy with the strategy of assigning only SER and the strategy of assigning only BUP. Horizontal axis is average clinical outcome. In vertical axis, 'optimal' is estimated optimal personalized treatment regime. 'all.BUP' is the strategy of assigning only BUP. 'all.SER' is the strategy of assigning only SER.
1.1 Background

Big data analysis is a challenging problem and has been researched exclusively in literature. Data dimension refers to the number of variants or features in the dataset. With the growing amount of data, the dimension increases and is usually greater than the sample size. High-dimensional data analysis has been a popular research topic recently. Many challenges are present when analyzing high-dimensional data. Curse of dimensionality is due to the
fact that statistical methods for the low dimensional setting are not directly applicable to the high-dimensional setting [23, 28, 29, 94]. For example, ordinary least squares estimator in the linear regression is not readily applicable when the dimension exceeds the sample size. We need to study the covariance matrix and precision matrix estimation in order to compute the estimator in high-dimensional setting. Despite the curse of dimensionality, many statistical methods have been proposed to solve problems in high-dimensional data analysis, for example the outliner detection [1], classification [96], visualization [52], feature selection [94] and support vector method [49].

For high-dimensional data analysis, variable selection is conducted to reduce data dimension [90]. We seek to identify the most important features that concern our problem of interest. In high-dimensional dataset, too many noise variables or correlated features are contained in the dataset which makes it difficult to select true variables. Usually in data pre-processing, we remove noise variables or highly correlated features as much as possible [53]. Our goal is to identify all the important features and to exclude all the noise ones from the selected model [67]. Whether we can achieve this goal or not depends upon the intensity of true variables. Only when the true variables have strong intensity, will we be able to select all the true variables without including any noise variable.

Variable selection has been researched exclusively in literature [44]. Popular variable selection methods include Mallows’ \( C_p \) measure [60], information criteria [2], cross validation, random forest [38], decision tree [44] and support-vector-machine based methods [70]. For different subsets of variables, information criteria can be applied to find the better selected model. For example, in Wang [88], it proposes using forward selection combined with information criteria to perform variable selection. The procedure in Wang [88] has
screening property asymptotically in high-dimensional setting.

However, enumerating all the subsets of variables to find the best one is computationally expensive when data dimension is high [27]. Penalized regression adds specific penalty functions to the likelihood. In high-dimensional setting, specific penalty functions produce sparse estimator of the unknown coefficient parameter. Penalized regression method allows us to conduct variable selection and parameter estimation simultaneously in high-dimensional data. Penalized regression method can be used for joint modeling and the effect of variants can be analyzed [45].

Penalized regression methods have been researched thoroughly in literature. Efficient algorithms have been developed for penalized regression to achieve fast computation. Convergence rates of parameter estimates and selection consistency have been shown for penalized regression. Penalized regression is one of the most important variable selection methods in high-dimensional setting. Lasso, SCAD and MCP are the major penalized regression approaches.

Among the challenges of applying penalized regression method in high-dimensional setting, the interpretation of selected model from penalized regression is of importance. We should provide confidence intervals of the unknown parameters. We should conduct hypothesis test of parameters after the variable selection is conducted. The randomness in variable selection will influence the confidence interval and hypothesis test of the unknown coefficients. The randomness of selected model is from the randomness in the observed data that should be accounted for [25].

High-dimensional inference of lasso penalized regression has been researched thoroughly in literature. In high-dimensional setting, desparsifying lasso method and covariance...
test are the most important methods proposed for lasso penalized regression to compute confidence intervals and to conduct hypothesis tests. The major research is in the following literature work Zhang and Zhang [100], Clemmensen et al. [20], Lockhart et al. [57], Van De Geer et al. [86], Taylor et al. [81], Buhlmann and Van De Geer [11], Dezeure et al. [21], Buhlmann [13], Efron [25], Tibshirani et al. [84], the importance of which will be elaborated in the following sections. In our first project, we conduct variable selection based upon covariance test. In our second project, desparsifying lasso approach is extended to the penalized regression in precision medicine. In Van De Geer et al. [86], desparsifying lasso approach is proposed for linear regression and it is also extended to generalized regression.

1.2 Penalized regression method

Penalized regression methods are widely applied in high-dimensional data. The penalized methods simultaneously select the variables and estimate the unknown coefficients. Estimators from penalized methods have bias. Consider the linear model:

\[ y = X \xi_0 + \epsilon, \quad (1.1) \]

where \( E(\epsilon|X) = 0 \). For the linear regression with the \( n \times p \) design matrix \( X \) and the response \( y \), the penalized regression coefficient estimator can be written as

\[ \hat{\xi} = \operatorname{argmin} ||y - X \xi||_2^2 + \rho_\phi(\xi), \quad (1.2) \]
where $\rho_\phi(\xi)$ is the penalty function and $\phi > 0$ is the tuning parameter. The penalty functions may take several different functional forms. Based upon the different convexity property of the penalty function, penalized regression can be categorized into different classes of methods such as Lasso [82], MCP [99] and SCAD [26]. The estimator of ridge regression does not have sparsity property. Ridge regression is not applicable to variable selection in high-dimensional setting. The penalty function should be singular at the origin for the coefficient estimate to be sparse [26]. When the penalty function takes the form of

$$
\rho_\phi(\xi) = \phi \|\xi\|_1,
$$

(1.3)

it is lasso penalized regression. Lasso is the least squares regression with constrained L1 norm of coefficients such that

$$
\|\xi\|_1 \leq \phi.
$$

(1.4)

Unlike ridge regression, the estimator from lasso penalized regression has sparsity property and can be used for variable selection in high-dimensional setting. The lasso coefficient estimate $\hat{\xi}$ can be computed fast through the LARS (least angle regression) procedure proposed in Efron et al. [24]. It has the geometric explanation that is to identify the variable which has the greatest correlation with the response. Then the response is projected to the variable and the residual is the new response. Another approach to obtain lasso solution is the shooting method proposed in Fu [36]. Shooting method is proposed for bridge regression which includes both ridge and lasso regression. Shooting method uses the modified Newton-Raphson (MNR) algorithm.
One of the advantages of lasso penalized regression is that it conducts variable selection and parameter estimation simultaneously. Contrary to enumerating all variables subsets to make variable selection which is computationally expensive in high-dimensional setting, penalized regression provides a sequence of variables to be added to the selected set. Solution path is the plot of parameter estimates $\hat{\xi}_1, \ldots, \hat{\xi}_m$ versus the tuning parameter $\phi$, where $m$ is the length of solution path. An example of solution path is in figure Fig. 1.1. We can see that solution path is composed of piecewise linear lines with knots. Each line represents the coefficient estimate for one variable as the tuning parameter $\phi$ decreases. When a coefficient estimate turns from zero to be nonzero, it means that the variable is added to the selected set. The bounds of lasso coefficient are given in Tibshirani [83]. The bounds are applicable under any general design matrix. The signs of lasso coefficients estimates stay the same for $\hat{\xi}$ as tuning parameter decreases, as shown in lemma 11 of Tibshirani [83].

The knots in solution path are marked by the entry of new variables into the selected set. If the design matrix $X$ satisfies the positive cone condition, there is no removal of variables along the lasso solution path [57]. The number of knots in Lars solution path is $\min(n, p)$ [24].

The lasso solution can be represented in solution path as in figure Fig. 1.1. The limiting distribution of lasso solution as $n, p \to \infty$ has been researched exclusively in literature. In Knight and Fu [51], the asymptotic distribution of the parameter estimate $\hat{\xi}$ is proposed for the low dimensional setting under the assumed scale of the tuning parameter. The asymptotic distribution is represented as the minimizer of a random function, which contains true parameter.
Figure 1.1 An example of lasso solution path. Horizontal axis is $\|\xi\|_1$ and also is $\phi$. Vertical axis is $\hat{\xi}_1, \ldots, \hat{\xi}_5$ respectively. There are 5 variables. Light blue, green, dark purple, red and black variables are added to the selected set sequentially.
Another important property of lasso penalized regression is selection consistency. The selection consistency is that lasso procedure contains all the important variables in the selected model as long as the irrepresentable conditions proposed in [103] hold. It is a condition imposed upon the covariance matrix of design matrix. It requires that noise variables are 'irrepresentable' by true variables. It also requires that error terms should have finite moments. Lasso has selection consistency properties for both fixed \( p \) setting and large \( p \) growing with \( n \) setting.

Lasso has disadvantages as well. Lasso solution has bias. For the penalty function with singularity at the origin, small coefficients shrink to zero. The lasso estimator of large coefficients also has bias. Desparsified lasso [86] has been proposed to correct the bias. Adaptive lasso estimator [104] does not have bias. In lasso solution path, true variables with strong intensities are ranked first. But true variables with weak intensities are ranked later and are mixed with noise variables. Another disadvantage of lasso is that it is not applicable when the dimension is too high compared to the sample size. Note that lasso does not necessarily have unique solution path. Lasso solution path is unique if the design matrix is in the general position [83].

Another important class of penalized regression method is SCAD [26]. SCAD is the abbreviation of smoothly clipped absolute deviation penalty. It uses the nonconcave penalty in penalized regression. Large coefficient estimator from SCAD does not have bias approximately. For small coefficients, SCAD shrinks them to be zero as well. SCAD has the derivative of the penalty function to be

\[
\rho'_\phi(\xi_j) = \phi \left\{ I(\xi_j \leq \phi) + \frac{(a \phi - \xi_j) I(\xi_j > \phi)}{(a-1)\phi} \right\},
\]

(1.5)
for some \( a > 2 \) and \( \xi_j \geq 0 \). SCAD estimator is explicitly represented as \([26]\)

\[
\xi_j = \begin{cases} 
\text{sgn}(z)|z| - \phi, & \text{when } |z| \leq 2\phi, \\
\{(a-1)z - \text{sgn}(z)a\phi\}/(a-2), & \text{when } 2\phi < |z| \leq a\phi, \\
z, & \text{when } |z| > a\phi,
\end{cases} 
\tag{1.6}
\]

where \( z \) is the ordinary least squares estimator of the coefficient. Based upon Bayesian analysis to minimize Bayesian risk, \( a \) is taken to be 3.7. For sufficiently strong intensity where \( |z| > a\phi \), SCAD estimator is equal to ordinary least squares estimator which is unbiased. In simulation of Fan and Li \([26]\), it shows that for sparse situation with large error variance, lasso performs better in terms of estimation. For sparse situation with small error variance, SCAD is better. In Fan and Li \([26]\), it discusses SCAD mainly for the low dimensional setting. Later in Fan and Peng \([29]\), Xie and Huang \([93]\) and Fan and Lv \([28]\), they extend SCAD for high-dimensional setting. SCAD estimators have similar properties and is applicable for variable selection and parameter estimation in high-dimensional setting.

Another important penalized regression method is MCP. MCP is the abbreviation of minimax concave penalized likelihood estimation. It uses a nonconvex (concave) penalty. The penalty function of MCP is

\[
\rho_{\phi}(\xi) = \phi \int_0^{|\xi|} (1 - t/(r\phi))_+ dt, 
\tag{1.7}
\]

where \( r > 0 \) is a regularization parameter of the concavity of penalty function. As \( r \) increases from 0 to \( \infty \), the concavity of penalty function increases and the estimators have
more bias. When $r = \infty$, it corresponds to lasso penalized regression. In Zhang [99], a penalized linear unbiased selection (PLUS) algorithm has been developed to compute the coefficient estimator from MCP penalized regression in a fast and efficient way. When the regularization parameter $r$ increases, the PLUS algorithm obtains the solution faster. MCP penalized regression also has the selection consistency property and it does not require the irrepresentable condition in lasso. The conditions for selection consistency in MCP are less strict than those in lasso. In Fan et al. [30], it establishes the strong oracle property of MCP with a modification to the concave penalty.

We have reviewed Lasso, SCAD and MCP penalized regression methods. These penalty functions may be combined to create other penalized regression. For example, the combination of lasso and MCP penalty is discussed in Fan and Lv [32]. An R package ncvreg has been developed to implement the penalized regression with SCAD or MCP penalty.

### 1.3 Inference in high-dimensional regression

In linear regression, we use least squares method to find the estimators. We conduct simple or composite hypothesis tests of the unknown coefficients. We also obtain confidence intervals. In penalized regression, variable selection is conducted when we fit the model to obtain the parameter estimates. When we perform the hypothesis tests or find the confidence intervals, we need to consider the randomness in the selected model [8, 9, 33, 54, 55, 80, 102].

High-dimensional inference has been developed for several statistical methods, such as the least absolute deviation regression [7], the quantile regression [6], the model averaging [39] and the instrumental variables [19]. My research focuses upon high-dimensional
inference of lasso penalized regression methods.

There are mainly three methods for the high-dimensional inference of lasso: the test based upon desparsified lasso estimate \cite{86, 100}, covariance test \cite{57} and exact test for lasso \cite{84}. The high-dimensional adaptive inference \cite{81} and the uniform high-dimensional inference \cite{5} have been discussed and applied in literature as well. High-dimensional inference methods of lasso include the test based upon bias correction and the test based upon the solution path. Desparsified lasso estimate is based upon bias correction of lasso solution. The solution path based tests, such as the sequential test and the exact test are conditional tests that provide the inference conditional upon the selected model. My research focus is upon the desparsifying lasso approach \cite{86} and the solution path based covariance test \cite{57}.

### 1.3.1 Desparsifying lasso approach

Desparsified lasso estimator is the bias-corrected lasso solution. The estimators from lasso have bias, which is from the tuning parameter in the penalty function \cite{51}. The bias correction of lasso estimator should be considered. After bias correction, the asymptotic distribution of the bias-corrected estimator can be derived. Desparsified lasso estimator in Van De Geer et al. \cite{86} is based upon the bias correction to lasso estimator.

The bias in lasso estimator can be corrected by either of the two methods introduced in Van De Geer et al. \cite{86} and Zhang and Zhang \cite{100}. In Van De Geer et al. \cite{86}, the derivation is based upon the KKT condition of penalized regression. Using KKT condition, it creates a bias-corrected lasso estimator. In Zhang and Zhang \cite{100}, it performs the following procedure for every dimension in the design matrix. First it projects both the response and the
dimension upon all the other dimensions using nodewise lasso regression. After projection, it performs the partial least squares of the response upon the dimension to obtain the bias-corrected lasso estimator.

Desparsified lasso estimator can be extended to other penalized regression methods by using the KKT conditions. In Van De Geer et al. [86], it extends desparsified lasso estimator to the generalized linear models. In Buhlmann and Van De Geer [12], the method in Zhang and Zhang [100] is applied to derive the asymptotic distribution of desparsified lasso estimator when the linear model is misspecified.

After we obtain desparsified lasso estimator, we have a set of p values for the variables. Based upon the p values, we can make variable selection. For variables in the selected set, p values may be greater than the pre-specified significance level. For variables excluded from the selected set, p values may be less than the significance level. In addition, based upon desparsified lasso estimator in Van De Geer et al. [86], we can obtain the confidence intervals of the unknown parameters [16]. In simulations of Van De Geer et al. [86], they show that the confidence intervals have good coverage properties.

In the desparsifying lasso procedure [86], the performance of desparsified lasso estimator depends upon two external methods: error variance estimator and ‘relaxed inverse’ i.e. the precision matrix [86]. The residual sum of squares of the selected model can be used to estimate the error variance [22]. Scaled lasso method can be used to estimate the coefficients and error variance simultaneously using gradient descent [79]. In Reid et al. [71], a variance estimation method for lasso is proposed for the high-dimensional setting.

We need to find a stable estimation method of covariance matrix and precision matrix since design matrix does not have full column rank in high-dimensional setting [15]. Co-
variance matrix may be estimated with the thresholding methods. The precision matrix can be obtained from the nodewise inverse [35, 62], the tuning insensitive estimation [56] or the network methods [68]. There are various approaches to get the stable and consistent estimators of the covariance and precision matrix in literature. As a result, though the desparsified lasso estimate depends upon the error variance estimate and the precision matrix estimate, these two problems have been well solved in literature. The desparsified lasso estimate is widely applied in high-dimensional problems. In Waldorp [87], it applies the desparsifying lasso method to test for the graph difference.

1.3.2 Covariance test statistic

Covariance test is constructed by considering the reduction in covariance between residual and response along the lasso solution path. In stepwise test, we calculate the reduction in covariance between the residual and the response to see whether adding a new variable is necessary. For adding a noise variable, the reduction in the residual covariance with response is not significant. For lasso penalized regression, solution path is piecewise linear with every knot corresponding to adding a new variable. At every knot of lasso solution path, the reduction in the residual covariance with response can be calculated to see whether adding a new variable at this knot is significant or not.

Covariance test in Lockhart et al. [57] is a solution-path-based stepwise test using the reduction in the residual covariance with response. The null hypothesis at each knot is that the selected model up to this knot has contained all the true variables. Assume that design matrix is in general position and satisfies positive cone condition, lasso solution path is unique with no deletion of variable at every knot [57]. In this case, the null hypotheses at
all the knots are nested by its nature. In Lockhart et al. [57], it has provided an R package ‘covTest’ that computes the covariance test statistics and the corresponding p values.

Under the global null hypothesis that there is no true variable in the statistical model, the sequence of covariance test statistics follow independent exponential distributions asymptotically under orthogonal design [57]. In Fan and Zheng [31], similar covariance tests and asymptotic results are derived for forward selection, SCAD, MCP and other penalized methods.

In G’sell et al. [41], variable selection methods based upon p values from covariance test are researched in detail. For a sequence of covariance test statistics $T_1, T_2, \cdots, T_m$, Q statistic is defined as $q_k = \exp(-\sum_{j=1}^{k} T_j)$, $k = 1, 2, \cdots, m$. Q statistics follow the same distribution as ordered independent standard uniform distributions for orthogonal design under null model. That is, Q statistics have the same asymptotic distribution as ordered i.i.d. p values for orthogonal design under null model. In G’sell et al. [41], tail stop is applied to Q statistics to choose the selected model size as $\min\{k : q_k \leq \alpha k / m\}$. In the simulations, tail stop is shown to have better power than other methods.

1.4 Precision medicine

Doctors assign treatment to patients according to their experiences and find better treatment for everyone. One particular treatment is not necessarily optimal for all the patients due to the heterogeneity in the population. Patients have numerous attributes that may affect the clinical outcome, for example gender, age, previous treatments received and the patients’ genetic variants. A more efficient treatment assignment scheme should prescribe treatment based upon the individual covariates that significantly affect the clinical outcome.
The field that analyzes optimal treatment regime is known as precision medicine or personalized medicine. Precision or Personalized medicine transforms traditional prevention and treatment programs into being more targeted at individual patient. The related informational dataset includes sequencing genomes, biobanks and electronic medical records (EMRs). The expectation is to lower medical cost and to improve treatment accuracy by providing individualized medical care [69, 101].

Precision medicine has been developed within the framework of causal inference, which focuses upon unveiling the causal relationship between the treatment and clinical outcome [46, 47, 72, 73]. The definition of clinical outcome is based upon the potential outcome in causal inference [66]. Potential outcome is the clinical outcome that we would observe if a particular treatment option was assigned to the patient. For binary treatment $A \in \{0, 1\}$, the potential outcomes are $Y^*(1)$ and $Y^*(0)$. We cannot observe both $Y^*(1)$ and $Y^*(0)$ for a certain subject since we can only assign one treatment option to each subject. We can observe either $Y^*(1)$ or $Y^*(0)$ but not both. To estimate potential outcome, we need to assume stable-unit-treatment-value ('SUTVA') [74]. In SUTVA, we assume that $Y = Y^*(1)A + Y^*(0)(1 - A)$. We also assume that there is no unmeasured confounder, that is all relevant individual information is observed. The observed individual covariates are in design matrix $X$. The potential outcomes are assumed to be independent of the treatment assignment given observed covariates, that is $\{Y^*(1), Y^*(0)\} \perp A | X$. Under the SUTVA assumption, $E[Y^*(a)] = E_X[E\{Y|A = a, X\}]$. Then $E\{Y|A = a, X\}$ is an unbiased estimator of the expected potential outcome $E[Y^*(a)]$. The treatment regimes is denoted to be $d(\tilde{X})$, which maps the individual covariates $\tilde{X}$ to the treatment regime $a \in \{0, 1\}$. Our
goal is to find the optimal treatment regime $d^{opt}(\tilde{X})$ that maximizes the value function $E[Y^*[d(\tilde{X})]]$. Under the SUTVA assumption, the contrast model of the value function is $E[Y^*(1) - Y^*(0)] = E_x[E\{Y|A = 1, X\} - E\{Y|A = 0, X\}]$. The contrast of value function $E[Y^*(1) - Y^*(0)]$ is positive if $E\{Y|A = 1, X\} > E\{Y|A = 0, X\}$. Since $d(\cdot) \in \{0, 1\}$, under the SUTVA assumption, $E[Y^*(d)] = E_x[E\{Y|A = 1, X\}d(\tilde{X}) + E\{Y|A = 0, X\}\{1 - d(\tilde{X})\}]$. The optimal treatment regime is to assign $d(\tilde{X}) = 1$ if $E\{Y|A = 1, X\} > E\{Y|A = 0, X\}$. That is, the optimal treatment regime is $d^{opt}(\tilde{X}) = I[E\{Y|A = 1, X\} > E\{Y|A = 0, X\}]$ [97].

The treatment regime contains several stages. For each stage, we need to identify the significant personal variants that affect the clinical outcome at that stage. For different stages, the important individual variants may be different as well. Usually the number of individual traits exceeds the number of patients enrolled in the study. To solve this problem, the lasso penalized regression methods can be applied and the corresponding desparsifying lasso procedure is proposed for precision medicine.

Dynamic treatment regime identifies the optimal decision mechanics that brings the best long-term outcome for the individual patient [65]. It utilizes the sequential decision making scheme that depends upon the individual patient’s characteristics [4, 17, 77]. Dynamic treatment regime has two main categories: 'Quality' learning [91] (Q-learning) and advantage learning [65]. Q-learning posits a regression model as a rough approximation for the expected clinical outcome of the individual patient. The expected clinical outcome for each individual is the Q-function $E(Y|X, A)$. We only need to know the contrast of expected clinical outcomes between different treatment options to make the comparison between them. As a result, instead of obtaining expected clinical outcome, it suffices to get only the clinical outcome contrast for different treatment options.
Advantage learning posits regression models for the contrast model of different treatment options and for the treatment assignment scheme (propensity score) in the observational data. In randomized study, propensity score is pre-specified. Advantage learning allows us to compare expected clinical outcomes by estimating only the contrast function. We leave the baseline model to be arbitrary, which allows more flexibility in the statistical model.

1.4.0.1 Variable selection in precision medicine

The optimal treatment decision is based upon the individual covariates. Due to the high dimension of covariates, it is necessary to identify the features that are significantly associated with the clinical outcome. Several methods have been proposed for the variable selection in dynamic treatment regime [37, 42, 43, 48, 58, 78, 98]. In advantage learning, we posit regression model for both the contrast model and the propensity score model. The variable selection should be performed for the interaction between treatment and covariates and for the covariates in the treatment assignment scheme. In Q-learning, we are interested in the variable selection in the approximate regression model for the Q-function.

In Gail and Simon [37], a testing method for qualitative interaction is proposed. Qualitative interaction is present when one treatment is better than the other for a specific group of subjects due to the heterogeneity in the population. There are subsets of subjects indexed by \( i = 1, \cdots, I \). For each subset, there are test statistics \( D_i \) that are independent \( N(\delta_i, \sigma_i^2) \), where \( \delta_i \) is the true difference in treatment efficacy for different subsets of subjects and \( \sigma_i^2 \) is known variance or can be consistently estimated. To test the null hypothesis of no qualitative interaction, a test statistic is proposed: \( H = \sum (D_i - \bar{D})^2 / \sigma_i^2 \), where \( \bar{D} = (\sum D_i / \sigma_i^2) / (\sum 1 / \sigma_i^2) \).
The test statistic $H$ follows central $\chi^2(I - 1)$ under the null hypothesis. The disadvantage of this method is that it suffers from curse of dimensionality.

In Gunter et al. [42], variable-ranking method is studied and two ranking statistics are proposed. There are two quantities of interest: interaction and proportion. The interaction is the magnitude of interaction between covariate and interaction. For $A \in \{0, 1\}$, the interaction is the degree to which $E(Y|X = x, A = 1) - E(Y|X = x, A = 0)$ varies as $x$ varies The proportion is the proportion of patients whose optimal treatment regime changes given knowledge of a variable. Define $d^* = \arg\max_d E(Y|A = d)$. The proportion is $\arg\max_d E(Y|X = x, A = d) \neq d^*$. There are $p$ variables, the ranking of which is of interest. For $j = 1, \cdots, p$, define

$$D_j = \max_{i \leq n} \left( \hat{E}[Y|X_j = x_{ij}, A = d^*] - \hat{E}[Y|X_j = x_{ij}, A \neq d^*] \right)$$

$$- \min_{i \leq n} \left( \hat{E}[Y|X_j = x_{ij}, A = d^*] - \hat{E}[Y|X_j = x_{ij}, A \neq d^*] \right),$$

where $d^* = \arg\max_d E(Y|A = d)$ is optimal treatment regime. Define

$$P_j = \frac{1}{n} \sum_{i=1}^{n} 1 \left\{ \arg\max_d \hat{E}[Y|X_j = x_{ij}, A = d] \neq d^* \right\}.$$

The score proposed for ranking variables is defined as

$$U_j = \left( \frac{D_j - \min_{1 \leq k \leq p} D_k}{\max_{1 \leq k \leq p} D_k - \min_{1 \leq k \leq p} D_k} \right) \left( \frac{P_j - \min_{1 \leq k \leq p} P_k}{\max_{1 \leq k \leq p} P_k - \min_{1 \leq k \leq p} P_k} \right).$$

The variables with large $U_j$ are selected, that is to select variables with both large $D_j$ and $P_j$. 
Another score proposed in Gunter et al. [42] is defined as

\[ S_j = \sum_{i=1}^{n} \left[ \max_d \hat{E}[Y|X_j = x_{ij}, A = d] - \hat{E}[Y|X_j = x_{ij}, A = d^*] \right]. \]

The steps for using the ranking to make variable selection are as follows. They select important variables using lasso-BIC. The variables are ranked using either \( U_j \) or \( S_j \) and only variables with nonzero \( U_j \) or \( S_j \) are included. Then \( K \) nested subsets are created using weighted lasso. The variables selected in lasso-BIC receive weight \( w = 1 \); and other variables in step 2 receive weight \( 0 < w \leq 1 \).

For subset \( k \), define \( d^*_k(x) = \arg\max_d \hat{E}[Y|X = x, A = d] \), \( \hat{V}_k = (1/n) \sum_{i=1}^{n} \hat{E}[Y|X = x_i, A = d^*_k(x_i)] \) and \( AG V_k = (\hat{V}_k - \hat{V}_0)/(\hat{V}_{m^*} - \hat{V}_0)(m^*/k) \), where \( m^* = \arg\max_k \hat{V}_k \) and \( \hat{V}_0 = (1/n) \sum_{i=1}^{n} \hat{E}[Y|A = d^*] \). In the variable selection procedure proposed in Gunter et al. [42], we should select the subset of variables with the highest Adjusted Gain in Value (AGV) criterion. The disadvantage of the variable-ranking method is that it uses marginal screening and it considers one covariate at a time in the variable-ranking method. Another disadvantage is that it lacks theoretical support such as false discovery error rate control or screening property.

The \( L_1 \) penalized regression is applied to make variable selection in Qian and Murphy [69]. They propose a two-step procedure. The first step is to estimate conditional mean response using \( L_1 \) penalized least squares (PLS):

\[
\hat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^J} \left\{ E_n[\Phi(X, A)\theta]^2 + \lambda_n \sum_{j=1}^{J} \hat{\sigma}_j |\theta_j| \right\},
\]

where \( \hat{\sigma}_j = [E_n \Phi_j(X, A)^2]^{1/2} \). We may use cross validation to choose the tuning parameter \( \lambda_n \). The second step is to derive estimated treatment rule from estimated conditional mean.
It derives the error bound for the estimated optimal treatment regime in terms of value function. The disadvantage of this method is that it has no inference for the estimated optimal treatment regime. It does not provide confidence interval or hypothesis test of parameters.

Contrast model is composed of interaction between the treatment and covariates. For penalized regression methods, adaptive lasso can be applied to perform variable selection for interactions between treatment and covariates in precision medicine [58]. They define the loss function in the framework of A learning:

\[
L_n(\beta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} \left[ Y_i - \phi(X_i; \gamma) - \beta^T \tilde{X}_i \{ A_i - \pi(\tilde{X}_i) \} \right]^2.
\]

The interaction coefficients are solved using

\[
\min_{\beta} L_n(\beta, \gamma) + \lambda_n w_j |\beta_j|,
\]

where \( w_j = |\beta_j| \). The BIC is used to choose the tuning parameter \( \lambda_n \). In Lv et al. [58], it provides confidence interval and hypothesis test of the parameter, which is robust to baseline model \( \phi(X_i; \gamma) \) misspecification. The disadvantage of this method is that it studies the fixed \( p \) case.

In genetic data, several variants jointly have an effect upon the clinical outcome given treatment but individually has no effect. In Huang and Fong [48], it proposes a variable combination selection method to identify significant variable combinations rather than individual variables.

Among these methods of variable selection in precision medicine, our concentration
is upon penalized regression in precision medicine for high-dimensional setting. Our research is to use lasso penalty to achieve variable selection and parameter estimation for high-dimensional data, to derive confidence interval and hypothesis test for estimated optimal treatment regime, and to prove that the inference is robust to baseline model misspecification.

1.4.0.2 Penalized method in precision medicine

Advantage learning is proposed in Murphy [65], in which the regret function has the same meaning as the contrast function. For any treatment \(a\), the contrast function \(C(X, a)\) is defined as the difference in clinical outcome between the optimal treatment \(d^*_j\) and \(a\):

\[
C(X, a) = E[Y(d^*_j)|X] - E[Y(a)|X, a],
\]

where \(Y(a)\) is clinical outcome for treatment \(a\). Based upon advantage learning, the conditional mean model of clinical outcome is

\[
E(Y|X, A) = \mu(\tilde{X}) + C(X, A),
\]

which is composed of baseline model \(\mu(\tilde{X})\) and contrast model \(C(X, A)\).

Baseline model \(\mu(\tilde{X})\) does not affect the treatment decision when we make pairwise comparisons between treatment options. Contrast model \(C(X, A)\) contains both the patient's information \(X\) and the treatment assignment \(A\). The posited contrast model contains the treatment main effect and the first-order interactions between the treatment and patients' covariates.
The penalized method in precision medicine based upon advantage learning is

\[
\hat{\beta} = \arg\min_{\beta} \|Y - \hat{\mu}_{\pi}(\tilde{X}) - (A - \pi) \cdot \tilde{X} \beta\|^2_2 + \lambda_n, p \|\beta\|_1, \tag{1.9}
\]

where \(\hat{\mu}_{\pi}(\tilde{X})\) is a baseline model estimator. The posited contrast model \((A - \pi) \cdot \tilde{X} \beta\) is orthogonal to the baseline model \(\mu(\tilde{X})\) given \(X\).

We may not conduct variable selection for baseline model since the comparison between different treatment options only depends upon contrast model. The baseline model can be posited to be arbitrary. The arbitrary specification of the baseline model may affect the estimation of \(\beta\) and the high-dimensional inference of \(\beta\). Our goal is to research the desparsifying lasso approach in the penalized regression in precision medicine. It allows us to obtain confidence intervals of the interaction coefficients and to test both the simple and composite null hypothesis for the interaction in contrast model. The challenge lies in the fact that the baseline model is arbitrary. We need to investigate the impact of the arbitrary specification in the baseline model upon the desparsifying lasso procedure of interaction for penalized regression in precision medicine.

1.5 Applications

1.5.1 Description of eQTL dataset

Gene expression is the protein or RNA produced in the genetic process. Identifying the association between genetic variants and gene expression has been studied exclusively in literature. It is useful in the research of genetic diseases. Down Syndrome is one of the most
common gene-associated disease. People born with Down Syndrome have an extra piece of chromosome 21. Due to the gene irregularity, the IQ of people with Down Syndrome is lower than normal people. There is no medication that treats Down Syndrome. Gene CCT8 is located at the critical region on human chromosome 21 that is associated with Down Syndrome. Our goal is to identify the genetic variants in gene CCT8 that are significantly related to Down Syndrome.

In order to find the genetic variants that are significantly associated with Down Syndrome, we analyze the eQTL dataset [40, 59, 95]. The genetic dataset eQTL (expression quantitative trait loci) is used to identify the significant effect of genetic variants over gene expression [50]. The eQTL dataset contains the genetic variations or locations of different subjects. The genetic variations are recorded based upon large-scale gene experiment. Perturbations of gene are used in finding the associations between gene and gene expression.

The design matrix is the genetic variants in the eQTL dataset. The response is the gene expression which is measured as the quantitative phenotypic traits. The genetic variations associated with the gene expression is found in gene experiments. Genevar (GENe Expression VARiation) is an integrative platform for the visualization and analysis of the eQTL studies [95]. The eQTL dataset of gene CCT8 contains a sample of subjects from four different countries: Europe, Yoruba, Japan and China [10]. There are missing data in the eQTL dataset. The eQTL dataset related to Down Syndrome has also been analyzed in Bradic et al. [10].
1.5.2 Description of STAR*D study

The major depressive disorder (MDD) is a common recurrent and chronic episodic disorder, which has the 4.9%-17.9% of lifetime prevalence. Women and young adults are more prone to suffer from MDD. It is predicted that by the year 2020, MDD is the second most important major causes of disability only next to the ischemic heart disease.

The common treatment of MDD is selective serotonin reuptake inhibitor antidepressant (SSRI). It helps boost the amount of serotonin in the brain that improves mood and develops better sleeping patterns. The patients often exhibit treatment resistance along the process. Identifying the most effective next-step treatment options for each individual patient will significantly improve the clinical results and reduce medical cost in general [76].

STAR*D is a multi-site and multi-stage randomized study for patients with MDD. In level 1 of STAR*D, patients are given citalopram (CIT) which is a selective serotonin reuptake inhibitor antidepressant [75]. Patients who have had unsatisfactory outcome in level 1 receive cognitive therapy at level 2. In level 2, patients receive either sertraline (SER) or bupropion (BUP) [34, 89]. The traditional treatment for depression patients is SSRI. SER is an antidepressant medicine of SSRI that inhibits the serotonin reuptake but does not have an effect upon dopamine or norepinephrine. SER is metabolized exclusively and is excreted by urine. BUP treats depression by inhibiting the dopamine, serotonin and norepinephrine reuptake. BUP is metabolized in the liver and we should be cautious to use BUP in patients with renal or hepatic disease [75].

The dataset in STAR*D contains the treatment assignment, individual characteristics, pretreatment information and individual clinical outcome. In level 2 of STAR*D study, 48% receives BUP and 52% receives SER. The clinical outcome is measured as the clinician-rated,
17-item Hamilton Rating Scale for Depression. The number of patient’s variants is much greater than the number of patients in STAR*D. We apply desparsifying lasso procedure in penalized method for precision medicine in order to estimate the optimal strategy of assigning either SER or BUP to individual patients in level 2.

In order to compare the effectiveness of SER and BUP, we will examine the covariates interaction with treatment and determine whether SER or BUP is better for each individual patient. For penalized method in precision medicine, lasso can be constructed with all the main effects of treatment and patient’s variants and all the interaction between treatment and patient’s variants.

1.6 Project overview

In our first project, we use Q statistics in covariance test to make model selection. Covariance test statistics can be transformed into Q statistics, which follow the same distribution as the ordered i.i.d. p values under null model. Bonferroni procedure and Benjamini-Hochberg procedure have been previously applied to Q statistics. But Bonferroni procedure and Benjamini-Hochberg procedure achieve FDR control only when perfect separation holds with high probability and that the true variable intensity should exceed a minimum. We propose a model size estimation procedure based upon Q statistics. We intend to relax the assumptions of the perfect separation and the minimum intensity requirement. We impose less assumptions and obtain an upper bound and a lower bound of the estimated model size. We prove that our procedure has selection consistency when true variables are of sufficiently strong intensity. In simulation, we compare our method with other testing-based methods which apply Bonferroni or Benjamini-Hochberg procedure to Q statistics. We also compare
our method with other methods such as lasso cross validation and lasso BIC. In simulation, we can see that our method has adaptivity in its performance in high-dimensional settings. In application, we apply our method to the genetic eQTL dataset which is related to the critical region of Down Syndrome. For Chinese and Japanese population, population in Europe and population in Yoruba, we identify different sets of important genetic variants for Down Syndrome.

Desparsifying lasso approach is another important method for hypothesis test in lasso penalized regression. An important application of desparsifying lasso approach is in precision medicine, where optimal personalized treatment decision can be estimated using penalized regression. For comparison between different treatment options, we are only interested in the sign of the contrast function. Baseline model can be posited to be arbitrary. We only need the high-dimensional inference of the coefficients for the interaction between the treatment and covariates. In our second project, desparsifying lasso approach is used to provide the confidence intervals and hypothesis tests of the interaction coefficients in penalized regression to estimate individualized treatment decision. We derive the desparsified lasso estimator for coefficients of interactions between the treatment and the patients’ covariates. We prove that the limiting distribution of the desparsified lasso estimator holds under baseline model misspecification. We select the important interactions with the p values less than 0.05. The optimal individualized treatment decision is estimated using the selected interactions and the desparsified lasso estimates. From our simulation, we can see that the desparsified lasso estimator has smaller mean absolute bias than the lasso estimator. In application, we consider STAR*D study of patients with major depression disorder. Our method is applied to estimate optimal individualized treatment decision.
To evaluate the estimated decision, we take bootstrap samples and compute the average outcome of patients receiving only SER, receiving only BUP and complying with estimated treatment decision. We can see that patients complying with estimated treatment decision have better clinical outcome.
CHAPTER 2

STATISTICAL INFERENCE FOR HIGH-DIMENSIONAL REGRESSION: POST-LASSO SOLUTION PATH INFERENCE
2.1 Introduction

The high-dimensional data which means that $p >> n$, where $p$ is the number of predictors and $n$ is the sample size, has been studied in thorough details in academic research. To reduce the dimension, variable selection is performed. Lasso penalized regression is applied to conduct variable selection and coefficient estimation simultaneously. In the low dimensional setting, inference of lasso has been researched thoroughly in literature. The convergence rate and selection consistency has been established for lasso solution.

However, the inference in usual sense constructs confidence intervals and statistical tests without considering the possible randomness in the model selection. When we make inference on the estimated parameters in the penalized model, if we ignored the model selection, it resulted in over-conservative results. The inference problem accounting for the effect of model selection is necessary [25].

As in Lockhart et al. [57], covariance test is a high-dimensional inference method for lasso in high-dimensional setting. Covariance test is a sequential test along lasso solution path and its limiting distribution is exponential for orthogonal design. They provide the R package 'covTest' which computes the covariance test statistic for lasso solution path in the high-dimensional problem. Covariance test can also be easily extended for other penalized regression methods such as SCAD and MCP solution paths as in Fan and Zheng [31].

Q statistic is developed from covariance test statistic in Lockhart et al. [57]. It is derived from tail stop created in G'sell et al. [41]. Under global null hypothesis, Q statistics follow the same distribution as ordered i.i.d. $p$ values for orthogonal design. Our procedure applies model size estimation to Q statistics.
2.1.1 Covariance test statistic and Q statistic

We mainly consider the setup of the linear regression with uncorrelated and homogeneous error variance structure, where the response vector is $y \in \mathbb{R}^n$ and the predictors matrix is $X \in \mathbb{R}^{n \times p}$:

$$y = X\beta^* + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I),$$  \tag{2.1}

where $\beta^* \in \mathbb{R}^p$ is the true unknown parameter and $\sigma^2$ is the unknown constant variance.

Lasso estimator proposed in Tibshirani [82] can be derived using the least angle regression methodology, as in Efron et al. [24] with minor modifications to the LARS path:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|eta\|_1,$$  \tag{2.2}

where $\lambda \geq 0$ is the tuning parameter. LARS solution path is the plot of the parameter estimate $\hat{\beta}(\lambda)$ with respect to the tuning parameter $\lambda$. The LARS path is piecewise linear and the knots values of $\lambda$ where the slope change occurs are ordered along the path as: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$, where $\lambda_k$, $k = 1, 2, \cdots, m$ is the knot corresponding to the entry of the $k$th variables, and $m$ is the length of LARS path with $m = \min(n - 1, p)$. Along the LARS path, the variables enter only at the knots values of the tuning parameter $\lambda$. When $\lambda = \infty$, parameter estimate $\hat{\beta}$ contains no nonzero elements. Every element of the parameter estimate $\hat{\beta}$ is nonzero when $\lambda = 0$.

The active set is defined to be the set of index for nonzero elements of the parameter estimate $\hat{\beta}$. The index of variables that enter the model at the knot value $\lambda_k$ are denoted
to be $j_k$, $k = 1, 2, \cdots, m$. It is assumed that only one variable enters at the knot value $\lambda_k$ so that $j_k$ represents only one variable. The active set right before the knot value $\lambda_k$ is denoted as $A_k = \{j_1, j_2, \cdots, j_{k-1}\} = \{j : \hat{\beta}_j(\lambda_k^-) \neq 0\}$, $k = 1, 2, \cdots, m$, where $\lambda_k^-$ is the knot value right before the $k$th knot, $\hat{\beta}_j(\lambda_k^-)$ is the coefficient estimate right before $k$th knot and $A_0 = \emptyset$. The true active set is the set of index for all the true variables $\beta^* A^* = \{j : \beta^*_j \neq 0\}$. The size of true active set is $s = |A^*|$, the number of elements in $A^*$. The size of active set right after $\lambda_k$ is $k$.

In Lockhart et al. [57], a covariance test statistic is introduced conditional upon lasso solution path. The covariance test statistic tests for the entry of new variable at the $k$th knot. Under the null hypothesis that $H_{0k} : A^* \subset A_k$, we know that $T_k \sim \exp(1)$ as $n \to \infty$ and $m \to \infty$ under orthogonal design. The null hypothesis $H_{0k}$ implies that the selected model right before the $k$th knot already covers all the true variables. If we reject the null hypothesis, addition of the variable that enters at the $k$th knot is significant. Otherwise, the selected model right before the $k$th knot is sufficient. The null hypothesis $H_{0k}$ is conditional upon the set $A_k$ that is the selected model right before the $k$th knot.

The covariance test statistic is constructed by measuring the reduction in residual covariance each time a new variable is added to the selected model [57]

$$T_k = \left( \langle y, X\hat{\beta}(\lambda_{k+1}) \rangle - \langle y, X\tilde{\beta}_A(\lambda_{k+1}) \rangle \right) / \sigma^2,$$

(2.3)

where the parameter estimates are $\hat{\beta}(\lambda_{k+1}) = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda_{k+1}\|\beta\|_1$, and $\tilde{\beta}_A(\lambda_{k+1}) = \arg\min_{\beta \in \mathbb{R}^{|A|}} \frac{1}{2}\|y - X_A\beta\|_2^2 + \lambda_{k+1}\|\beta_A\|_1$. If we assume that design matrix $X$ has

31
orthogonal columns, the covariance test statistic can be simplified to be [57]

\[ T_k = \lambda_k (\lambda_k - \lambda_{k+1}) / \sigma^2, \quad k = 1, 2, \ldots, m, \quad (2.4) \]

where \( T_m = \lambda_m^2 \).

Assume orthogonal design and that \( \min_{j \in A_0} |\beta^*_j| - \sigma \sqrt{2 \log p} \to \infty \), as \( p \to \infty \), which implies that perfect separation holds with probability converging to 1. (Perfect separation indicates that all the true variables enter before all the noise variable.) Then under the global null hypothesis, we have that for any fixed \( d \),

\[ (T_1, T_2, \ldots, T_d) \xrightarrow{d} (\exp(1), \exp(1/2), \ldots, \exp(1/d)), \quad (2.5) \]

as \( p \to \infty \). The test statistics \( T_1, T_2, \ldots, T_d \) are independent asymptotically. Similarly for the other penalized regression methods, the covariance test statistics can be defined and the asymptotic distributions can be derived.

Since \( T_k \) tests only for the stepwise entry of the new variable at the \( k \)th knot, it is not as powerful in terms of identifying the proper stop point along lasso path, because it only considers a single step at a time. In G’sell et al. [41], the Renyi representation is applied, and we may construct the Q statistics using the covariance test statistics in (Eq. 2.5). Define the Q statistics as

\[ q_k = \exp \left( - \sum_{j=k}^{m} T_j \right). \quad (2.6) \]

Assume that the perfect separation holds. Under the global null hypothesis and the orthog-
onal design, we have that

\[(q_1, q_2, \cdots, q_m) \overset{d}{=} \text{Order statistics of } m \text{ independent Unif}(0, 1) \text{ r.v.'s,} \quad (2.7)\]

as \(m \to \infty\). Notice that the Q statistics \(q_k, k = 1, 2, \cdots, m\) have the same distribution as the ordered i.i.d. p values. The Q statistics jointly consider the covariance test statistics at all the remaining knots. Model selection based upon the Q statistics is more powerful.

### 2.1.2 Previous method based upon Q statistic

Pseudo correlation in design matrix can be a problem since the noise is mixed with true variables along the solution path. With a set covering more true variables, we inevitably includes more noise variables in the set. The benefit of using Q statistics is that it is more powerful than checking stepwise covariance test statistics. Another advantage is that under global null Q statistics have the same distribution of ordered i.i.d. p values. In theory, applying BH procedure to Q statistics controls familywise error rate under global null and perfect separation [41]. When the noise is mixed with true, perfect separation does not hold. As a result, we find the necessity to develop a control over Q statistics that have good properties even when noise is mixed with true.

Desirable properties of variable selection include screening property and familywise error rate control. For solution path where noise is mixed with true, in order to achieve screening property, error rate control is sacrificed. Unless we put assumptions upon the degree of mixture between noise and true, we cannot arrive at conclusions of familywise error control. As a result, for solution path where noise is mixed with true, criteria for
judging variable selection can be from other measures such as $g$-measure and the ROC curve. Theoretically we may provide upper and lower bound of model selection quality.

In the next section, it provides an introduction to our approach of estimating the number of false null hypotheses using Q statistics. For the covariance test statistic, the conditional null hypothesis is that $H_{0k} : A^* \subset A_k$, $k = 1, \cdots, m$. Under the global null hypothesis, $H_{01}, \cdots, H_{0m}$ are all true. For a certain true model size, only a few first null hypotheses are false and the remaining null hypotheses are true. Due to the nested nature of null hypotheses, we can find a proper cutoff value (i.e. estimated model size) where the null hypotheses before it are all false. Our procedure of estimating model size is named 'QVS', where 'Q' stands for Q statistics and 'VS' stands for variable selection.

In the following sections of theoretical study, we find the lower and upper bounds of the selected model size in QVS. The lower and upper bound of QVS meet when there is no mixture between noise and true along the solution path. For the general case where mixture exists, the lower bound is the entry of the first noise and the upper bound is the entry of the last true variable. The bound provides us with the properties of our procedure QVS under general mixture structure in solution path. We also conduct simulations to study the properties of QVS in comparison with other methods such as applying BH over Q statistics, applying fixed threshold, best subset selection with information criterion and lasso cross validation. Among these methods, QVS has the most stable performance across various high-dimensional setting. At the end, QVS is applied to the eQTLs data from the area of genetics in search of significant genetic variants that are associated with Down Syndrome.
2.2 QVS procedure: Model size estimation

In figure Fig. 2.1, it shows a flow of variables that enter the selected model along the solution path. There is a certain degree of mixture of noise and true variables along the solution path. Define $m_1$ to be the location of the last true variable in the solution path. Define $m_0$ as the location of the first noise variable in the solution path. If $m_0 = m_1$, it implies perfect separation, that is all the true variables are before the noise variables along the solution path. Generally, $m_1 \geq m_0$.

![Diagram](image)

* stands for true variable  
| stands for noise variable

**Figure 2.1** An example of $m_0$ and $m_1$ in lasso solution path. Variables are ordered along lasso solution path. $m_0$ is the entry right before first noise variable. $m_1$ is the entry of last true variable.

Define for $0 < t < 1$,

$$U_m(t) = \frac{1}{m} \sum_{i=1}^{m} 1(Z_i \leq t), \quad (2.8)$$

where $Z_1, Z_2, \ldots, Z_m$ are the realizations of i.i.d. Unif(0, 1) random variables. Denote that
$U_1, U_2, \cdots, U_m$ are the increasing order statistics of $Z_1, Z_2, \cdots, Z_m$. Then it is also true that

$$U_m(t) = \frac{1}{m} \sum_{i=1}^{m} 1(U_i \leq t).$$

Define

$$V_m = \sup_{t \in (0,1)} \frac{U_m(t) - t}{\sqrt{t(1-t)}}. \quad (2.9)$$

Define $c_m$ to be the bounding sequence such that: $mc_m$ is monotonically increasing with $m$ and that $P(V_m > c_m) < \alpha \to 0$ as $m \to \infty$ [61, 63].

QVS selects the model to be the set of variables indexed by $\{j_1, \cdots, j_{\hat{G}}\}$. Selected model size $\hat{G}$ in QVS is

$$\hat{G} = m \max_{1 \leq k \leq m/2} \left\{ \frac{k}{m} - q_k - c_m \sqrt{q_k(1-q_k)} \right\}, \quad (2.10)$$

where $q_k$ is the Q statistic and $c_m$ is the bounding sequence. The bounding sequence $c_m$ has the following implications. It controls the Type I error under the global null hypothesis $H_0$, that is $P_{H_0}(\hat{G} > 0) < \alpha \to 0$ as $m \to \infty$. We take $\alpha = 1/\sqrt{\log m}$ to generate the bounding sequence through simulations under the global null. In the next section, $m_0$ and $m_1$ are shown to be the lower and upper bound of the $\hat{G}$. 

36
2.3 Statistical property of QVS procedure

2.3.1 Upper bound for QVS estimated model size

2.3.1.1 Preliminary lemmas

The estimated model size depends solely upon the Q statistics. Since we intend to establish properties about $\hat{G}$, we need to study the distribution of Q statistics for solution path where noise is mixed with true. Q statistics is derived from covariance test statistics, the distribution of which under orthogonal design has been given in Lockhart et al. [57]. Under orthogonal design, we study the asymptotic properties of the Q statistics based upon the asymptotic theory of the covariance test statistic. In the preliminary lemmas, the distribution of knots, covariance test statistics and Q statistics are provided given that the solution path has a mixture of noise and true. The preliminary lemmas provide an insight into the asymptotic distributions of the Q statistics, based upon which we can study the theoretical properties of the QVS method.

Define $H_j = X_j^T Y/\sigma$. Denote $|H_j|$ to be the decreasingly ordered $|H_j|$. Under orthogonal design, it has been shown in Lockhart et al. [57] that $\sigma|H_j|$ are the knots along the solution path. Under orthogonal design, $H_j \sim N(0, 1)$. The knots are ordered i.i.d. normal random variables. Based upon orthogonal design and definition of knots, the covariance test statistic is $T_j = |H_j||(|H_j| - |H_{j+1}|)$. It is a function of knots. The distributions of knots, covariance statistics and Q statistics are given in the lemma 2.3.1, 2.3.2 and 2.3.3. Lemma 2.3.1 provides distribution of knots under orthogonal design, which will be used for deriving distribution of covariance test statistic in lemma 2.3.2. Then lemma 2.3.2 is used to derive the distribution
of Q statistics in lemma 2.3.3 [64, 92].

**Lemma 2.3.1.** Let $Z_1, \ldots, Z_{m-s}$ be $(m - s)$ i.i.d. random variables from normal distribution $N(0, 1)$. Let $|Z_{(1)}|, \ldots, |Z_{(m-s)}|$ be decreasingly ordered $|Z_j|$. Then

$$|H_{(m_1+1)}|, \ldots, |H_{(m)}| \overset{d}{=} |Z_{(m_1-s+1)}|, \ldots, |Z_{(m-s)}|.$$  

The proof is straightforward given the solution path, the definition of $m_1$ and $H_j$. Lemma 2.3.1 says that given the solution path, the joint distribution of $|H_{(m_1+1)}|, \ldots, |H_{(m)}|$ is the same as joint distribution of the last $m - m_1$ order statistics of $m - s$ $N(0, 1)$ i.i.d. random variables.

For instance, $|H_{(m_1+1)}|$ is the $(m_1 + 1)$th order statistic of $s$ $N(\beta, 1)$ and $(m-s)$ $N(0,1)$ independent random variables. Now we remove the $s$ $N(\beta, 1)$ random variables in the sequence, then $|H_{(m_1+1)}|$ is the $(m_1 - s + 1)$th order statistic of $(m-s)$ i.i.d. $N(0, 1)$. Similarly, $H_{(m)}$ is the $(m - s + 1)$th order statistic of $(m-s)$ i.i.d. $N(0, 1)$.

Lemma 2.3.1 provides the asymptotic distribution of the knots which implies the following lemma about the covariance statistic. Using results about extreme value theory, the knots are ordered i.i.d. normal distributions and the covariance statistics are exponential distributions.

**Lemma 2.3.2.** Under orthogonal design,

$$T_{m_1+1}, \ldots, T_m \overset{d}{\to} \exp\left(\frac{1}{m_1 - s + 1}\right), \ldots, \exp\left(\frac{1}{m - s}\right),$$

as $m \to \infty$.  

38
Suppose that \( Z_1, Z_2, \cdots, Z_{m-s} \) are the (decreasing) order statistics of \((m-s)\) i.i.d. \( \mathcal{N}(0,1) \) random variables. Define \( W_1 = Z_1(Z_1 - Z_2), \cdots, W_{m-s-1} = Z_{m-s-1}(Z_{m-s-1} - Z_{m-s}), W_{m-s} = Z_{m-s} Z_{m-s} \). As proved in Lockhart et al. [57], \( W_1, W_2, \cdots, W_{m-s-1} \) are asymptotically jointly independent and have distributions \( \exp(1), \exp(1/2), \cdots, \exp(1/(m-s-1)) \), as \( m \to \infty \). The asymptotic distributions of \( T_{m+1}, \cdots, T_{m-1} \) are the same as those of \( W_{m-s+1}, \cdots, W_{m-s} \).

From lemma 2.3.1, we know that \( T_{m+1} \overset{d}{\to} \exp(1/(m-s+1)) \). The result holds similarly for \( T_{m+2}, \cdots, T_m \).

Now that we have the asymptotic results of the covariance test statistics, we can use the Lemma 2.3.2 to prove the following lemma about the Q statistics using Renyi representation transformation in extreme value theory.

**Lemma 2.3.3.** Let \( U_1, \ldots, U_{m-s} \) be \((m-s)\) i.i.d. random variables following \( \text{Unif}(0,1) \). Let \( U_{(1),m-s}, \ldots, U_{(m-s),m-s} \) be increasingly ordered \( U_j \) so that \( U_{(j),m-s} \sim \text{Beta}(j, m-s+1-j) \). Then

\[
q_{m+1}, \ldots, q_m \overset{d}{\to} \text{joint distribution of } U_{(m-s+1),m-s}, \ldots, U_{(m-s),m-s}.
\]

Define \( K_1 = \exp(-\sum_{j=1}^{m-s} W_j), \cdots, K_{m-s-1} = \exp(-\sum_{j=s}^{m-s} W_j), K_{m-s} = \exp(-\sum_{j=s}^{m-s} W_j) \). As proved in Lockhart et al. [57], \( K_1, K_2, \cdots, K_{m-s-1} \) asymptotically follow the same distributions as the order statistics of \((m-s)\) i.i.d. \( \text{Unif}(0,1) \) random variables. Since \( q_{m+1} = \exp(-\sum_{j=1}^{m} T_j) \), and \( T_{m+1} \) asymptotically follows \( \exp(1/(m_1-s+1)) \), then \( q_{m+1} \) asymptotically follows the same distribution as the \((m_1-s+1)\) order statistics of \((m-s)\) iid \( \text{Unif}(0,1) \), which is \( U_{(m-s+1),m-s} \).

Our procedure QVS has the model size estimated as \( \hat{G} \), which is a function of Q statistics. Lemma 2.3.3 has provided the distribution of Q statistics under orthogonal design as
\( p \to \infty \). We can use results in lemma 2.3.3 to identify the upper and lower bound of our model size estimator \( \hat{G} \).

### 2.3.1.2 Upper bound of QVS estimated model size

The length of lasso solution path from LARS algorithm is \( m = \min(n, p) \). The estimated model size \( \hat{G} \) is less than \( m \). As \( n \to \infty \) and \( p \to \infty \), we have that \( m \to \infty \). The number of true variables is \( s \). Recall that the \( m_0 \) is the entry right before first noise variable and \( m_1 \) is the entry of last true variable. For a general solution path, we can show that our estimated model size \( \hat{G} \) is bounded in the interval \([m_0, m_1]\) in probability.

**Theorem 2.3.1.** Assume that the columns of design matrix \( X \) are orthogonal and that \( s = o(m) \) i.e. \( s = o(n) \) and \( s = o(p) \). Then we have

\[
P(\hat{G} \leq m_1) \to 1,
\]

as \( m \to \infty \).

**Proof of Theorem 2.3.1.**

\( \hat{G} \). From the definitions of bounding sequence in Eq. 2.8 and Eq. 2.9, we have that

\[
P \left( \max_{0 < t < 1} \frac{U_m(t) - t}{\sqrt{t(1-t)}} > c_m \right) = \alpha.
\]  \( \text{(2.11)} \)

Then

\[
P \left( \max_{0 < t < 1} \frac{U_m(t) - t - c_m \sqrt{t(1-t)}}{\sqrt{t(1-t)}} > 0 \right) = \alpha.
\]
Because
\[
\frac{U_m(t) - t - c_m \sqrt{t(1-t)}}{\sqrt{t(1-t)}} > U_m(t) - t - c_m \sqrt{t(1-t)},
\]
for every \( t \in (0, 1) \), then it implies that
\[
P \left( \max_{0 < t < 1} \{ U_m(t) - t - c_m \sqrt{t(1-t)} \} > 0 \right) \leq \alpha.
\]

The definition of estimated model size in Eq. 2.10 is convenient for computation and has an equivalent representation. For any fixed value of \( t \in (0, 1) \), estimated model size in Eq. 2.10 can be rewritten as
\[
\hat{G} = m \max_{0 < t < 1} \{ F_m(t) - t - c_m \sqrt{t(1-t)} \}, \tag{2.12}
\]
where
\[
F_m(t) = \frac{1}{m} \sum_{k=1}^{m} 1(q_k \leq t). \tag{2.13}
\]

We transform the definition of model size estimator in Eq. 2.10 to be 2.12 for convenience of theoretical analysis in \( \hat{G} \). We are to split \( F_m(t) \) to be components made of Q statistics before and after \( m_1 \) since Q statistics before and after \( m_1 \) have different distributions. For notation convenience, define \( \pi_1 = m_1/m \). We have
\[
F_m(t) = \frac{m_1}{m} \frac{1}{m_1} \sum_{j=1}^{m_1} 1(q_j \leq t) + \frac{m - m_1}{m} \frac{1}{m - m_1} \sum_{j=m_1+1}^{m} 1(q_j \leq t),
\]
\[
\leq \pi_1 + (1 - \pi_1) \frac{1}{m - m_1} \sum_{j=m_1+1}^{m} 1(q_j \leq t). \tag{2.14}
\]
Recall that from Lemma 2.3.3, \( q_{m+1}, \ldots, q_m \) converge in distribution to the last \( m-m_1 \) order statistics of \( m-s \) i.i.d. random variables following \( \text{Unif}[0,1] \). Define

\[
U_{m-m_1,m-s}(t) = \frac{1}{m-m_1} \sum_{j=m_1-s+1}^{m-s} 1(q_j \leq t)
\]

where \( U_{j,m-s} \) is the \( j \)th order statistic of the \((m-s)\) i.i.d. \( \text{Unif}(0,1) \) random variables sorted in an increasing order. Then for the Q statistic component in Eq. 2.14, we have that

\[
\frac{1}{m-m_1} \sum_{j=m_1+1}^{m} 1(q_j \leq t) d \approx U_{m-m_1,m-s}(t).
\]

As a result, from Eq. 2.14, we have that \( F_m(t) \leq \pi_1 + (1-\pi_1)U_{m-m_1,m-s}(t) \) in probability.

Notice that our purpose is to prove that estimated model size \( \hat{G} \) is upper bounded by \( m_1 \) in probability. We demonstrate it by proving that the probability of \( \hat{G} \) greater than \( m_1 \) converges to zero. Probability of \( \hat{G} \) greater than \( m_1 \) has the following implications:

\[
P(\hat{G} > m_1) = P(\sup_{0 \leq t < 1} \{F_m(t) - t - c_m \sqrt{t(1-t)}\} > \pi_1),
\]

\[
\leq P(\sup_{0 \leq t < 1} \{\pi_1 + (1-\pi_1)U_{m-m_1,m-s}(t) - t - c_m \sqrt{t(1-t)}\} > \pi_1),
\]

\[
= P(\sup_{0 \leq t < 1} \{(1-\pi_1)U_{m-m_1,m-s}(t) - t - c_m \sqrt{t(1-t)}\} > 0).
\]

Apply triangular inequality, we have that

\[
P(\hat{G} > m_1) \leq P(\sup_{0 \leq t < 1} \{(1-\pi_1)U_{m-m_1,m-s}(t) - U_m(t)\} > 0)
\]

\[
+ P(\sup_{0 \leq t < 1} \{U_m(t) - t - c_m \sqrt{t(1-t)}\} > 0),
\]

\[
\leq P(\sup_{0 \leq t < 1} \{(1-\pi_1)U_{m-m_1,m-s}(t) - U_m(t)\} > 0) + \alpha. \tag{2.15}
\]
Next study the term \((1 - \pi_1)U_{m-m_1,m-s}(t) - U_m(t)\):

\[
(1 - \pi_1)U_{m-m_1,m-s}(t) - U_m(t)
= \frac{m-m_1}{m} \frac{1}{m-m_1} \sum_{j=m_1-s+1}^{m-s} 1(U_{(j),m-s} \leq t) - \frac{1}{m} \sum_{j=1}^{m} 1(U_j \leq t),
= \frac{1}{m} \left[ \sum_{j=1}^{m-s} 1(U_{(j),m-s} \leq t) - \sum_{j=1}^{m} 1(U_j \leq t) - \sum_{j=1}^{m_1-s} 1(U_{(j),m-s} \leq t) \right].
\]

Consider two independent sequences of order statistics, \(U_{1,m-s}, \cdots, U_{(m-s),m-s}\) and \(U_{1,m}, \cdots, U_{(m),m}\) from the two i.i.d. sequences: \(W_{1,m-s}, \cdots, W_{(m-s),m-s}\) and \(W_{1,m}, \cdots, W_{(m),m}\). We have that

\[
\frac{m-m_1}{m} \frac{1}{m-m_1} \sum_{j=1}^{m-s} 1(U_{(j),m-s} \leq t) = \frac{m-s}{m} \frac{1}{m-s} \sum_{j=1}^{m-s} 1(W_{(j),m-s} \leq t) \xrightarrow{p} t,
\]

since \(s = o(m)\) is assumed. We also know that

\[
\frac{1}{m} \sum_{j=1}^{m} 1(U_j \leq t) = \frac{1}{m} \sum_{j=1}^{m} 1(W_{(j),m} \leq t) \xrightarrow{p} t.
\]

As a result, from Eq. 2.16, we have that

\[
(1 - \pi_1)U_{m-m_1,m-s}(t) - U_m(t) = o(1) - \frac{1}{m} \sum_{j=1}^{m_1-s} 1(U_{(j),m-s} \leq t) < 0
\]

in probability. By the definition of bounding sequence, we know that \(a = 0\), as \(m \to \infty\).

From Eq. 2.15, we know that \(P(\hat{G} > m_1) = 0\), as \(m \to \infty\).
2.3.2 Lower bound for QVS estimated model size

Recall that $m_0$ is right before the entry of first noise variable. In this section, we prove that estimated model size $\hat{G}$ is greater than $m_0$ in probability. True variables with sufficiently strong intensity enter the solution path successively from the 1st knot to the $m_0$th knot. This result implies that our procedure QVS at least identifies true variables with strong intensity when the solution path shows mixture of noise and true variables. We assume that minimum true variable intensity exceeds $\sqrt{\log m}$ so that $m_0 > 0$ in probability. Otherwise lower bound for $\hat{G}$ is 0.

**Theorem 2.3.2.** Assume that the columns of design matrix $X$ are orthogonal, minimum true variable intensity $|\beta^*| > \sqrt{\log m}$ and that $s = o(m)$ i.e. $s = o(n)$ and $s = o(p)$. Then we have that

$$P(\hat{G} \geq m_0) \rightarrow 1,$$

as $m \rightarrow \infty$.

**Proof of theorem 2.3.2:**

Estimated model size $\hat{G}$ is composed of Q statistics. The scale of Q statistic $q_{m_0}$ is needed in the proof. First we study the scale of covariance test statistic $T_{m_0+1}$. From the definition of covariance test statistic for orthogonal design in Eq. 2.4, we have that

$$T_{m_0+1} = \lambda_{m_0}^2(\lambda_{m_0+1} - \lambda_{m_0+2}).$$

Recall that $m_0$ corresponds to the knot of entry right before first noise variable. We have that $m_0 + 1$ is a noise variable and $m_0 + 2$ is either noise or true variable. Then $\lambda_{m_0+1}$ is the
largest among \((m - s)\) i.i.d. \(|N(0, 1)|\) corresponding to the first noise variable. For \(\lambda_{m_0+1}\), we know that

\[
\lambda_{m_0+1} \overset{p}{=} \sqrt{2 \log m} - \frac{\log \log m}{2 \sqrt{2 \log m}} \approx \sqrt{2 \log m}
\]  

(2.17)
in probability.

If \(\lambda_{m_0+2}\) is the second largest among \((m - s)\) i.i.d. \(|N(0, 1)|\) corresponding to a noise variable, we have that

\[
\lambda_{m_0+1} - \lambda_{m_0+2} = \Phi^{-1}(1 - 1/(m + 1)) - \Phi^{-1}(1 - 2/(m + 1)) > \sqrt{\log m}
\]  

(2.18)
in probability.

If \(\lambda_{m_0+2}\) is the \((m_0 + 1)\)th largest among \(s\) i.i.d. \(|N(\beta^*, 1)|\) corresponding to a true variable, we have that

\[
\lambda_{m_0+1} - \lambda_{m_0+2} = \beta^* + \Phi^{-1}(1 - 1/(m + 1)) - \Phi^{-1}(1 - (m_0 + 1)/s) > \beta^* > \sqrt{\log m}
\]  

(2.19)
in probability. From Eq. 2.17, Eq. 2.18 and Eq. 2.19, we have that \(T_{m_0+1} > \sqrt{2 \log m}\) in probability. Hence we have that

\[
\exp(-T_{m_0+1}) = o_p\left\{\frac{1}{(m_1 - s) \log m}\right\}.
\]  

(2.20)
The scale of covariance test statistic in Eq. 2.20 is used to derive the scale of Q statistic. By definition of Q statistics in Eq. 2.6, we have that

$$q_{m_0+1} = \exp\left( - \sum_{k=m_0+1}^{m_1} T_k \right) q_{m_1+1}. \quad (2.21)$$

Since we have that

$$\exp\left( - \sum_{k=m_0+1}^{m_1} T_k \right) \leq \exp(-T_{m_0+1}), \quad (2.22)$$

from Eq. 2.20, we know that

$$\exp\left( - \sum_{k=m_0+1}^{m_1} T_k \right) = o_p\left\{ \frac{1}{(m_1-s) \log m} \right\}. \quad (2.23)$$

In addition, we have that $q_{m_1+1} = O_p\{ (m_1-s)/m \}$ from lemma 2.3.1. From Eq. 2.21 and Eq. 2.23, we have that

$$q_{m_0+1} = o_p\left( \frac{1}{m \log m} \right). \quad (2.24)$$

The scale of $q_{m_0+1}$ in Eq. 2.24 is needed for the proof. Notice that we intend to prove that $\hat{G} \geq m_0$ in probability. We demonstrate it by proving that $\hat{G}/m_0 - 1 \geq 0$ in probability. We implicitly take $m_0$ to be positive, that is the solution path starts with a true variable. We assume that the minimum true variable intensity exceeds $\sqrt{\log m}$, which implies that the solution path starts with a true variable in probability.

Recall that in the proof of theorem 2.3.1, we have the representation of estimated model
size \( \hat{G} \) to be in Eq. 2.12. We have the following implications about \( \hat{G}/m_0 - 1 \). From extreme value theory, bounding sequence \( c_m \) has the scale that \( \lim_{m \to \infty} c_m/\sqrt{\log(m)/m} = C < \infty \), where \( C \) is a finite constant. For any \( 0 < t < 1 \), we have that

\[
\frac{\hat{G}}{m_0} - 1 = \sup_{0 < t < 1} \left\{ \frac{1}{m_0} \sum_{k=1}^{m} 1(q_k \leq t) - 1 - \frac{m}{m_0} t - \frac{m}{m_0} \sqrt{\frac{\log m}{m}} \sqrt{t(1-t)} \right\},
\]

\[
\geq \left\{ \frac{1}{m_0} \sum_{k=1}^{m} 1(q_k \leq t) - 1 - \frac{m}{m_0} t - \frac{m}{m_0} \sqrt{\frac{\log m}{m}} \sqrt{t(1-t)} \right\}. 
\tag{2.25}
\]

In Eq. 2.25, conditional upon solution path, take \( t = q_{m_0+1} \). From the scale of \( q_{m_0+1} \) in Eq. 2.24, we have that

\[
\frac{m}{m_0} t^p \to 0, \tag{2.26}
\]

and

\[
\frac{m}{m_0} \sqrt{\frac{\log m}{m}} \sqrt{t(1-t)}^p \to 0. \tag{2.27}
\]

In addition, we have that

\[
\frac{1}{m_0} \sum_{k=1}^{m} 1(q_k \leq t) - 1 = \frac{1}{m_0} \sum_{k=1}^{m_0} 1(q_k \leq t) - 1 + \frac{1}{m_0} \sum_{k=m_0+1}^{m} 1(q_k \leq t)
\]

\[
= \frac{1}{m_0} \sum_{k=m_0+1}^{m} 1(q_k \leq t) \geq 0 \tag{2.28}
\]

in probability.
From Eq. 2.25, Eq. 2.26, Eq. 2.27 and Eq. 2.28, we have that

\[
\frac{\hat{G}}{m_0} - 1 \geq 0
\]  

(2.29)

in probability. That is, \( P(\hat{G} \geq m_0) \rightarrow 1 \).

### 2.4 Selection consistency of QVS procedure

In the previous section, it has been shown that estimated model size is within the interval \([m_0, m_1]\) in probability. The screening property means that selected set of QVS covers all true variables in probability. When the solution path is in perfect separation, \( m_0 = m_1 \) in probability, estimated model size \( \hat{G} = m_1 \), that is our procedure has screening property.

In the following theorem, the assumption that \( m_1 - s = o(m_1 \land m_1^2 / \log m) \) is a restriction upon the degree of mixture in the solution path. The count \( m_1 - s \) is the number of noise variables before the entry of last true variable. In perfect separation, \( m_1 - s = 0 \). The greater \( m_1 - s \) is, the more severe the mixture of noise and true variables is along the solution path.

In this assumption, we require that \( m_1 - s = o(m_1) \) and also \( o(m_1^2 / \log m) \). If \( m_1 = O(\log m) \) or \( m_1 / \log m \to \infty \), we assume that \( m_1 - s = o(m_1) \). If \( m_1 / \log m \to 0 \), we assume that \( m_1 - s = o(m_1^2 / \log m) \). In essence, we assume that \( m_1 - s \) is small compared to \( m_1 \).

**Theorem 2.4.1.** Assume orthogonal design, and that \( s = o(m) \) and that \( m_1 - s = o(m_1 \land m_1^2 / \log m) \). Then we have

\[
P\{\hat{G} > (1 - \epsilon)m_1\} \to 1
\]

for an arbitrarily small constant \( \epsilon > 0 \) as \( m \to \infty \).
We prove that \( \hat{G} > (1 - \epsilon)m_1 \) in probability by proving that \( \frac{\hat{G}}{m_1} - 1 > -\epsilon \) in probability, where \( \epsilon > 0 \) is arbitrarily small. Since \( m_1 \geq s > 0 \), we know that \( m_1 \) is positive. We do not assume any further assumptions. Recall that we rewrite \( \hat{G} \) in Eq. 2.10 to be Eq. 2.12 for convenience of theoretical proof. The bounding sequence \( c_m \) has the scale that \( \lim_{m \to \infty} c_m / \sqrt{\log m / m} = C < \infty \), where \( C \) is a finite constant. For \( \frac{\hat{G}}{m_1} - 1 \), we have the following derivations. For any \( 0 < t < 1 \),

\[
\frac{\hat{G}}{m_1} - 1 = \sup_{0 < t < 1} \left\{ \frac{1}{m_1} \sum_{k=1}^{m} \mathbb{1}(q_k \leq t) - 1 - \frac{m}{m_1} t - \frac{m}{m_1} \sqrt{\frac{\log m}{m}} \sqrt{t(1-t)} \right\},
\]

\[
\geq \left\{ \frac{1}{m_1} \sum_{k=1}^{m} \mathbb{1}(q_k \leq t) - 1 - \frac{m}{m_1} t - \frac{m}{m_1} \sqrt{\frac{\log m}{m}} \sqrt{t(1-t)} \right\}. \tag{2.30}
\]

In Eq. 2.30, conditional upon solution path, we take \( t = q_{m_1+1} \). We already know from lemma (2.3.1) that

\[
q_{m_1} < q_{m_1+1} = O_p \left( \frac{m_1 - s}{m} \right).
\]

We have that

\[
\frac{m}{m_1} t \xrightarrow{p} 0, \tag{2.31}
\]

and that

\[
\frac{m}{m_1} \sqrt{\frac{\log m}{m}} \sqrt{t(1-t)} \xrightarrow{p} 0. \tag{2.32}
\]
In addition, we have that
\[
\frac{1}{m} \sum_{k=1}^{m} 1(q_k \leq t) - 1 = \frac{1}{m} \sum_{k=1}^{m_1} 1(q_k \leq t) - 1 + \frac{1}{m} \sum_{k=m_1+1}^{m} 1(q_k \leq t)
\]
\[
= \frac{1}{m_1} \sum_{k=m_1+1}^{m} 1(q_k \leq t) \geq 0
\]  (2.33)

in probability. From Eq. 2.30, Eq. 2.31, Eq. 2.32 and Eq. 2.33, we have that \( \hat{G} / m_1 - 1 \geq 0 \) in probability.

### 2.5 Simulation

#### 2.5.1 Simulation setting

Our simulation has two purposes. The first is to perform QVS procedure and verify that estimated model size is in the interval \([m_0, m_1]\) in probability. The second is to compare QVS with other variable selection methods, such as Lasso cross validation, lasso with BIC, Bonferroni control applied to Q statistics and Benjamini-Hochberg control applied to Q statistics. We use multiple criteria to make the comparison across a variety of high-dimensional settings.

Design matrix \( X \) is Gaussian random matrix generated from \( N(0, \Sigma) \). Response \( y \) is from \( N(X\beta^*, I) \), where \( \beta^* \) is the true coefficient. The simulation has four settings, \( n = 200 \) and \( p = 100 \), \( n = 200 \) and \( p = 400 \), \( n = 200 \) and \( p = 2,000 \), \( n = 200 \) and \( p = 10,000 \). The nonzero intensity \( \beta^* \) of the true variable is set to be uniform and varies from 0.2 to 1.9. The number of true variables \( s \) has three levels, so that the ratios of the number of true variables to the
length of the solution path vary from 0.05 to 0.2. The true variables are indexed randomly in
the design matrix. Covariance matrix $\Sigma$ has two levels: $\Sigma = I$ and $\Sigma = (0.5|i-j|)_{i=1,\ldots,p; j=1,\ldots,p}$.

Bounding sequence $c_m$ is simulated under global null. Generate $X$ and $y$ from the global
null model. Compute lasso solution path, covariance test statistic and Q statistics. Find $V_m$ in Eq. 2.9 using $V_m = \max_{1 \leq i \leq m/2} (i/m - q_i)/\sqrt{q_i(1-q_i)}$, which is equivalent to Eq. 2.9.
Repeat the four steps above for 1000 times. The output is that $V^1_m, V^2_m, \ldots, V^{1000}_m$. Bounding
sequence $c_m$ is taken to be the upper $\alpha = 1/\log m$ percentile of $V^1_m, V^2_m, \ldots, V^{1000}_m$.

Implement QVS with Lars solution path and compare its performance with the classical methods, such as lasso BIC ('BIC'), lasso cross-validation ('LCV'), Q statistics with Benjamini-Hochberg method ('FDR') and Q statistics with Bonferroni method ('BON'). 'BIC' is that

$$\text{argmin}_{1 \leq k \leq m} \left\{ \frac{1}{\hat{\sigma}^2} \sum_{l=1}^{n} (y_l - \hat{\beta}_0 - \hat{\beta}_1 x_{1l} - \cdots - \hat{\beta}_k x_{kl})^2 + k \log(n) \right\},$$

where $\hat{\sigma}^2$ is estimated error variance. In 'LCV', 10-fold cross-validation is used to select
lasso tuning parameter $\lambda$ so that the prediction error of the model is minimized. 'FDR' is
$$\text{argmax}_k \left\{ k : q_k \leq 0.05k/m \right\}.$$ 'BON' is $\text{argmax}_k \left\{ k : q_k \leq 0.05/m \right\}$.

The simulation results are summarized in the following tables. Denote $\hat{G}$ to be the average
stop point, i.e. the average size of the selected set in QVS. Denote $m_0$ to be the average
index right before entry of first noise variable. Denote $m_1$ to be the average index where
the last true variable enters the solution path. Denote Coverage to be the proportion of
repetitions where the QVS is between $m_0$ and $m_1$.

Denote 'TPP' to be the number of selected true variables divided by the total number
of true variables averaged over all repetitions. Denote "FPP" to be the number of selected

51
noise variables divided by the total number of noise variables averaged over all repetitions. Denote "FDP" to be the number of selected noise variables divided by the size of selected set averaged over all repetitions. The specificity is defined as

\[ \text{specificity} = \frac{TN}{TN + FP} = 1 - FPP. \]

If the specificity is higher, the method includes less noise variables in the selected set. The sensitivity is calculated as

\[ \text{sensitivity} = \frac{TP}{TP + FN} = TPP. \]

If the sensitivity is higher, QVS includes more true variables. When more true variables are included in the selected set, more noise variables are included as well. In the low dimensional setting, it is better when specificity and sensitivity are both greater. However, in the high-dimensional setting, specificity decreases as sensitivity increases. G-measure achieves a balance between the specificity and sensitivity. It is better when G-measure is higher. The g-measure is defined to be

\[ g \text{ measure} = \sqrt{\text{specificity} \times \text{sensitivity}}. \]

A method with greater g-measure achieves the goal of including more true variables without including too many noise variables. In the high-dimensional setting, g-measure can be used to make comparisons between different methods.
### 2.5.2 Upper and lower bound of QVS estimated model size

Table 2.1 Frequency of QVS estimated model size located between lower and upper bound when true variable intensity $\beta^* = 0.3$ and sample size $n = 200$. Dimension $p = 2,000$ or $p = 10,000$. Number of true variables $s$ is 10, 20, 30 or 40. TPP is true positive proportion. Standard errors are in parenthesis. $m_0$ is the entry right before first noise variable. $m_1$ is the entry of last true variable. $F(m_0 < \hat{G} \leq m_1)$ is frequency of QVS estimated model size located between upper and lower bound. $F(m_0 < \hat{G})$ is frequency of QVS estimated model size greater than lower bound. $F(\hat{G} \leq m_1)$ is frequency of QVS estimated model size less than upper bound.

<table>
<thead>
<tr>
<th>p</th>
<th>s</th>
<th>G</th>
<th>TPP</th>
<th>$m_0$</th>
<th>$m_1$</th>
<th>$F(m_0 &lt; G \leq m_1)$</th>
<th>$F(m_0 &lt; G)$</th>
<th>$F(G \leq m_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>10</td>
<td>19.60(5.20)</td>
<td>0.78(0.14)</td>
<td>3.47(1.92)</td>
<td>65.53(46.66)</td>
<td>0.91</td>
<td>1.00</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>39.15(6.14)</td>
<td>0.73(0.11)</td>
<td>3.24(2.24)</td>
<td>131.36(40.95)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>54.44(6.58)</td>
<td>0.70(0.09)</td>
<td>2.71(2.13)</td>
<td>160.47(29.62)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>66.98(6.65)</td>
<td>0.70(0.09)</td>
<td>2.45(2.06)</td>
<td>173.52(22.42)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>10000</td>
<td>10</td>
<td>27.12(6.55)</td>
<td>0.72(0.17)</td>
<td>2.00(1.49)</td>
<td>96.59(50.82)</td>
<td>0.93</td>
<td>1.00</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>50.67(7.20)</td>
<td>0.71(0.13)</td>
<td>1.63(1.43)</td>
<td>141.18(41.50)</td>
<td>0.98</td>
<td>1.00</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>67.48(7.27)</td>
<td>0.71(0.12)</td>
<td>1.24(1.25)</td>
<td>157.21(33.52)</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>78.11(6.48)</td>
<td>0.71(0.12)</td>
<td>0.94(1.09)</td>
<td>163.74(29.26)</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
</tr>
</tbody>
</table>

In 1,000 simulation runs, the $m_0$ and $m_1$ are recorded and the QVS stop points are located between $m_0$ and $m_1$ most of the time.

### 2.5.3 Comparison of QVS performance with other method

In table Table 2.2, we make a comparison between different variable selection methods and QVS. 'LCV' has greater variability in the size of selected set. As true variable intensity increases, true positive proportion increases as well for QVS. As true variables are more dense, QVS has smaller true positive proportion when true variable intensity is strong. These are empirical observations from simulation study. The theoretical mechanism remains to
Table 2.2 A table of comparison between QVS and popular methods when sample size $n = 200$ and dimension $p = 10,000$. Covariance matrix of design matrix follows AR structure. Number of true variables $s = 10, 20$ or 40. True variable intensity $\beta^* = 0.3$ or 0.6. $\hat{G}$ is QVS estimated model size. TPP is true positive proportion. $s^*$ is number of true variables covered in lasso solution path. Standard errors are in parenthesis.

<table>
<thead>
<tr>
<th>Cov(X)</th>
<th>$s$</th>
<th>$\beta^*$</th>
<th>BIC</th>
<th>LCV</th>
<th>BON</th>
<th>FDR</th>
<th>QVS</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>10</td>
<td>0.3</td>
<td>$G$</td>
<td>1.00(0.00)</td>
<td>7.44(12.52)</td>
<td>13.88(3.92)</td>
<td>40.33(8.67)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TPP</td>
<td>0.11(0.05)</td>
<td>0.28(0.32)</td>
<td>0.57(0.18)</td>
<td>0.79(0.15)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$s^*$</td>
<td>7.96(1.35)</td>
<td>7.96(1.35)</td>
<td>7.96(1.35)</td>
<td>7.96(1.35)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td></td>
<td>$G$</td>
<td>10.73(10.69)</td>
<td>36.21(15.06)</td>
<td>26.45(5.25)</td>
<td>52.68(9.68)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TPP</td>
<td>0.51(0.45)</td>
<td>0.99(0.05)</td>
<td>0.99(0.03)</td>
<td>1.00(0.01)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$s^*$</td>
<td>9.99(0.08)</td>
<td>9.99(0.08)</td>
<td>9.99(0.08)</td>
<td>9.99(0.08)</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.3</td>
<td>$G$</td>
<td>1.00(0.00)</td>
<td>7.50(15.02)</td>
<td>31.74(5.97)</td>
<td>70.20(9.32)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TPP</td>
<td>0.06(0.04)</td>
<td>0.17(0.25)</td>
<td>0.57(0.15)</td>
<td>0.77(0.13)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$s^*$</td>
<td>11.70(2.31)</td>
<td>11.70(2.31)</td>
<td>11.70(2.31)</td>
<td>11.70(2.31)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td></td>
<td>$G$</td>
<td>1.02(0.47)</td>
<td>44.88(29.16)</td>
<td>73.94(7.42)</td>
<td>110.55(9.40)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TPP</td>
<td>0.05(0.02)</td>
<td>0.67(0.28)</td>
<td>0.87(0.08)</td>
<td>0.95(0.06)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$s^*$</td>
<td>18.06(1.83)</td>
<td>18.06(1.83)</td>
<td>18.06(1.83)</td>
<td>18.06(1.83)</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.3</td>
<td>$G$</td>
<td>1.00(0.00)</td>
<td>4.63(13.68)</td>
<td>61.66(7.72)</td>
<td>104.09(9.13)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TPP</td>
<td>0.04(0.04)</td>
<td>0.07(0.17)</td>
<td>0.63(0.13)</td>
<td>0.80(0.11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$s^*$</td>
<td>13.49(2.86)</td>
<td>13.49(2.86)</td>
<td>13.49(2.86)</td>
<td>13.49(2.86)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td></td>
<td>$G$</td>
<td>1.00(0.00)</td>
<td>10.78(20.84)</td>
<td>118.39(7.49)</td>
<td>150.97(7.33)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TPP</td>
<td>0.04(0.03)</td>
<td>0.16(0.24)</td>
<td>0.85(0.09)</td>
<td>0.92(0.06)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$s^*$</td>
<td>17.57(3.09)</td>
<td>17.57(3.09)</td>
<td>17.57(3.09)</td>
<td>17.57(3.09)</td>
</tr>
</tbody>
</table>
be established.

For all the simulation settings, in order to provide better visualization of the comparison between all the variable selection methods in terms of various criteria, we make plots across all simulation settings. For the plots, the horizontal axis is the intensity of true variables. The vertical axis is the TPP (true positive proportion), FPP (false positive proportion) and g-measure as defined. The main title states the simulation setting, where $AR$ means that design matrix have covariance matrix $\Sigma = (0.5|i-j|)_{i=1,\ldots,p; j=1,\ldots,p}$.

For the case where $n = 200$ and $p = 100$ (figure Fig. 2.2), it is the low dimensional setting. The true positive proportion of QVS is high when $s = 5$ (more sparse model). But when $s$ increases to 10 and 20 (more dense model), the true positive proportion of the LCV and BIC is slightly better.

The false positive proportion of QVS, FDR and BON is comparable under all three settings. The false positive proportion of LCV in the sparse model ($s=5$) is good, but bad in dense models ($s=10$ and $s=20$). The false positive proportion of BIC is good in the dense model ($s=20$) but bad in the sparse models ($s=5$ and $s=10$).

In terms of the g-measure, QVS has good performance overall except in some cases where $s=10$ and $s=20$ and the intensity is weak 0.2. The g-measure of BON and FDR is generally not as good as that of BIC and LCV. BIC and LCV performs better in low dimensional settings than in high-dimensional settings. In high-dimensional setting, the error variance in BIC is difficult to estimate accurately, adversely impacting its performance. For LCV, when $p$ is much greater than $n$, the cross validation has limited power and the volatility of coefficients estimates is severe.

The FDP of the QVS is not as good as that of BON and FDR in the low dimensional
Figure 2.2 Comparison of QVS with popular methods when sample size $n = 200$ and dimension $p = 100$. Number of true variables $s = 5, 10$ or $20$. Covariance matrix of design matrix follows AR structure. Vertical axis is true positive proportion (TPP), false positive proportion (FPP), g-measure and false discovery proportion (FDP) respectively. Horizontal axis is true variable intensity.
setting. In sparse model ($s=5$), LCV has good FDP. LCV does not have as good FDP in dense models ($s=10$ and $s=20$). BIC has good FDP in dense model ($s=20$) but not as good FDP in sparse models ($s=5$ and $s=10$). The FDP of QVS is relatively stable across all three settings. The FDP of FDR and BON increases slightly as $s$ increases and the model is denser. The FDP of BIC decreases slightly as $s$ increases and the model is denser.

For the case where $n=200$ and $p=400$ (figure Fig. 2.3), it is high-dimensional but the dimension is not ultra-high. The true positive proportion of LCV is the best in all settings in this scenario. The true positive proportion of QVS is the second best in all settings.

Similar to what we observed in the low dimensional case, the false positive proportion of BIC is better in dense model. Whereas the false positive proportion of LCV is better in sparse model. The false positive proportion of QVS, FDR and BON increases as model is denser.

The g-measure of QVS is the best in almost all the settings. The g-measure of LCV is slightly better than QVS when the intensity is 0.3. The g-measure of LCV decreases as intensity increases. The g-measure of FDR and QVS is comparable, and is better than that of BON.

The phenomenon of the FDP is similar to that in the low dimensional case. The false discovery proportion of QVS is worse than that of FDR and BON. The LCV has bad FDP in all three settings. The FDP of BIC is better in dense models, which implies that BIC tends to select sparse models. Across all different intensities, the FDP is relatively stable for all the methods except BIC. The FDP of BIC tends to increase when intensity increases.

For the case where $n=200$ and $p=2000$ (figure Fig. 2.4), the dimension is much higher than the sample size. The phenomenon we observe in this case is very different from what
Figure 2.3 Comparison of QVS with popular methods when sample size $n = 200$ and dimension $p = 400$. Number of true variables $s = 10, 20$ or $40$. Covariance matrix of design matrix follows AR structure. Vertical axis is true positive proportion (TPP), false positive proportion (FPP), g-measure and false discovery proportion (FDP) respectively. Horizontal axis is true variable intensity.
Figure 2.4 Comparison of QVS with popular methods when sample size $n = 200$ and dimension $p = 2,000$. Number of true variables $s = 10, 20$ or 40. Covariance matrix of design matrix follows AR structure. Vertical axis is true positive proportion (TPP), false positive proportion (FPP), g-measure and false discovery proportion (FDP) respectively. Horizontal axis is true variable intensity.
we observed in the previous two cases.

The true positive proportion of QVS is the best when $s=10$ and $s=20$. The true positive proportion of QVS is not as good as that of FDR and BON when $s=40$. The true positive proportion of LCV is good in sparse model ($s=10$ and $s=20$). In the dense model ($s=40$), the true positive proportion of LCV is significantly worse than QVS, FDR and BON. The true positive proportion of BIC is much worse than the other methods in all three settings. The true positive proportion of BIC is close to 0 when $s=40$. The bad estimation of the error variance for the ultra-high-dimensional case is one reason why BIC fails.

Though the true positive proportion of FDR and BON reaches close to 1 as intensity increases, the false positive proportion also grows in an unconstrained fashion to 1. On the contrary, in the dense setting ($s=40$), QVS stops selecting more variables as intensity increases and stabilizes at true positive proportion of 0.8 and false positive proportion of 0.4. In the dense setting ($s=40$), LCV exhibits similar pattern and stabilizes at true positive proportion of 0.7 and false positive proportion of 0.4 as intensity increases. The false positive proportion of QVS, FDR and BON increases as the model is denser.

The g-measure of QVS is the best in almost all the settings. In some settings where $s=10$ and $s=20$ and the intensity is strong, the g-measure of BON is better than that of QVS and FDR. In the dense setting where the intensity is strong, the g-measure of QVS is better than LCV and the other methods. In the dense case, the g-measure of BIC is stable across all intensities.

The FDP of QVS, LCV and FDR are comparable in all three settings. The FDP of BON is much better than QVS, LCV and FDR in the sparse setting. The FDP of QVS increases slightly as the model is denser. The FDP of all the methods is stable across all intensities.
except BIC in the sparse setting.

For the case where \( n=200 \) and \( p=10000 \) (figure Fig. 2.5), the dimension is even higher. The phenomenon is similar to that in the case of \( n=200 \) and \( p=2000 \). The true positive proportion of BIC is the worst in all the settings. The true positive proportion of FDR is the best across all the settings. In the sparse settings (s=10 and s=20), the QVS has the second best true positive proportion. In the dense setting (s=40), the true positive proportion of QVS, FDR and BON is much better than that of LCV and BIC. The QVS does not have as good true positive proportion as FDR and BON in the dense setting. This is completely different from our observations in the low dimensional case. Now the LCV and BIC have much worse performance compared to the other methods. Generally in the low dimensional setting and the high-dimensional setting where \( p \) is not that much greater than \( n \), the LCV has good true positive proportion. But in the ultra-high-dimensional setting, the FDR has good true positive proportion and the LCV has very bad true positive proportion. The performance of QVS is good across all the low dimensional and high-dimensional and ultra-high-dimensional settings. In this sense, the QVS shows adaptivity across all scenarios for model selection.

The false positive proportion of FDR is the highest and the false positive proportion of BIC is the lowest. The BIC and LCV tend to select sparse models when the true model is dense (s=40). The true positive proportion and the false positive proportion are both lowest for BIC and LCV in the dense setting (s=40). In the dense model (s=40), as intensity increases, the QVS stabilizes at true positive proportion of 0.8 and false positive proportion of 0.5, which is very similar to the corresponding setting of \( p=2000 \). The LCV stabilizes at true positive proportion of 0.3 and false positive proportion close to 0. Recall that in
Figure 2.5 Comparison of QVS with popular methods when sample size $n = 200$ and dimension $p = 10,000$. Number of true variables $s = 10, 20$ or $40$. Covariance matrix of design matrix follows AR structure. Vertical axis is true positive proportion (TPP), false positive proportion (FPP), g-measure and false discovery proportion (FDP) respectively. Horizontal axis is true variable intensity.
the case of \( p=2000 \), the LCV stabilizes at true positive proportion of 0.7 and false positive proportion of 0.4. Now the dimension increases from \( p=2000 \) to \( p=10000 \), the LCV stabilizes at a much worse true positive proportion than 0.7. This is another evidence of the lack of adaptivity of LCV. On the other hand, QVS shows superior adaptivity than LCV and BIC.

The g-measure of QVS is the best in almost all the settings. The BON has slightly better g-measure than QVS when the intensity is strong in the sparse model \( (s=10) \). In the dense model \( (s=40) \), the g-measure of QVS is much better than the other methods.

A short-coming is that now the FDP of QVS is among the highest. The FDR method has the highest FDP. The FDR method is proved to control false discovery rate asymptotically under orthogonal design and perfect separation. The false discovery rate control is not feasible in the ultra-high-dimensional setting with significant amount of mixture of noise and true variables along the solution path. The BIC has the lowest FDP, but the true positive proportion of BIC is close to 0 when \( s=20 \) and \( s=40 \).

As we demonstrated using the simulations, the QVS procedure has the best adaptivity among the methods we are comparing in our simulation.

### 2.6 Application to eQTL dataset concerning Down Syndrome

Dataset eQTL is in the area of genetics, including individual gene expression and variants. The number of variants is much greater than sample size. Down Syndrome is a genetic disease that is caused solely by genetic irregularities. We have eQTL data about the critical region of Down Syndrome in the gene CCT8. Our dataset is the eQTL data of the Japanese and Chinese population and the population in Yoruba for gene CCT8.

We implement QVS to the eQTL dataset to identify significant genetic variants associated
with Down Syndrome. In table Table 2.3, we have the covariance test statistics of the first 10 knots in the solution path. The p value is computed from stepwise covariance test using Eq. 2.5. The Q statistic is defined in Eq. 2.6. We have three different populations separated by different countries and continents. We conduct QVS for different populations separately.

**Table 2.3** Table of covariance test statistics, p values and Q statistics along lasso solution path for eQTL data concerning Down Syndrome. There are three populations: European population, Chinese and Japanese population and population in Yoruba. For each population, covariance test statistics, p values and Q statistics are displayed. P values are from covariance test.

<table>
<thead>
<tr>
<th>step</th>
<th>European</th>
<th>Chinese and Japanese</th>
<th>Yoruba</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Covtest</td>
<td>P value</td>
<td>Q statistic</td>
</tr>
<tr>
<td>1</td>
<td>9.81e-03</td>
<td>0.99</td>
<td>0.85</td>
</tr>
<tr>
<td>2</td>
<td>2.94e-02</td>
<td>0.94</td>
<td>0.85</td>
</tr>
<tr>
<td>3</td>
<td>1.64e-02</td>
<td>0.95</td>
<td>0.88</td>
</tr>
<tr>
<td>4</td>
<td>1.78e-03</td>
<td>0.99</td>
<td>0.89</td>
</tr>
<tr>
<td>5</td>
<td>1.05e-02</td>
<td>0.94</td>
<td>0.90</td>
</tr>
<tr>
<td>6</td>
<td>6.99e-03</td>
<td>0.95</td>
<td>0.91</td>
</tr>
<tr>
<td>7</td>
<td>1.06e-03</td>
<td>0.99</td>
<td>0.91</td>
</tr>
<tr>
<td>8</td>
<td>6.49e-03</td>
<td>0.94</td>
<td>0.91</td>
</tr>
<tr>
<td>9</td>
<td>1.03e-04</td>
<td>0.99</td>
<td>0.92</td>
</tr>
<tr>
<td>10</td>
<td>7.83e-03</td>
<td>0.92</td>
<td>0.92</td>
</tr>
</tbody>
</table>

The Covtest is the value of the covariance test statistic defined in Lockhart et al. [57].

The true variables have strong intensity and are sparse in the Japanese and Chinese population. The QVS selects the first two variables which are both true. The LCV procedure has a large variation in its selected sets, as shown in the large standard error of the average selected model size in table Table 2.2. The performance of QVS is close to the setting where \( n = 200, p = 2000 \) and \( s = 10 \) in the AR setting, as shown in figure Fig. 2.4.

From the table (Table 2.3), the Q statistics are indeed a more powerful tool for detecting true variables than the ordered p values from the covariance test statistic. For the European
population, the true variables are too weak and the p values from the covariance test are all very large and are not significant. The Chinese and Japanese population and the population in Yoruba exhibit similar patterns both in p values and Q statistics. The most significant p value is at step 2 for both populations. The Q statistic becomes larger from step 2 to step 3.
CHAPTER 3

HIGH-DIMENSIONAL INFERENCE FOR PERSONALIZED TREATMENT DECISION

3.1 Introduction

Traditionally during the clinical visit, the treatment scheme is designed for the patient based upon the doctor’s expert knowledge in the area. The heterogeneity in the population of patients is often overlooked by practitioners when making the treatment decision. The
individualized optimal treatment regime can be prescribed through the analysis of data, in
the hope of lowering medical cost and improving efficacy of clinical outcome.

Precision medicine has been developed within the framework of causal inference. Ad-

dvantage learning, first proposed in Murphy [65], posits the regression models for the contrast
function of the clinical outcome of the different treatment options. The baseline function
$\mu(\tilde{X})$ represents the main effects of the patient’s variants and does not affect the treatment
decision. The variable selection in the penalized method in precision medicine is focused
upon the selection of the important interactions between the treatment option and the pa-
tient’s variants. When we posit the linear working model for the baseline, we must consider
the effect of possible baseline misspecification upon the interaction.

Previous work has been done upon the variable selection for finding the optimal treat-
ment regimes. Some are based upon the statistical tests, such as the Wald test and the
likelihood ratio test as in [37]. This class of variable selection method is for the relatively
low dimensions and does not account for the randomness of the selected model in the
statistical tests. Other methods focus upon the inference for the interaction. In Lv et al. [58],
the adaptive penalty is applied to conduct the variable selection. It only requires that the
interaction between the biomarker and the treatment is posited correctly but is robust to
the misspecification of the baseline model.

In the high-dimensional setting, lasso may be exercised to construct variable selection
for penalized method in precision medicine. Inference for parameters taking account of
the randomness in the selected model, in other words the post-selection inference for the
parameters has been studied exclusively in the literature, for instance Van De Geer et al.
[86]. We can research essentially how reliable the estimated optimal decision is with the
post-selection inference.

In penalized method for precision medicine, misspecification of baseline model has an effect upon the post-selection inference of interaction between treatment and patients’ variants. In Buhlmann and Van De Geer [12], it studies the high-dimensional post-selection inference of desparsified lasso estimator with misspecified linear model. Desparsified lasso procedure in Zhang and Zhang [100] is used in Buhlmann and Van De Geer [12] that relies upon the nodewise lasso regression in the design matrix. The derivation in Buhlmann and Van De Geer [12] is based upon the assumptions of the error in the misspecified working model. The whole linear model is assumed to be misspecified in Buhlmann and Van De Geer [12]. Our research applies the assumptions of the error in the true statistical model. We utilize the theory in Bunea et al. [14] that analyzes the convergence rate of lasso estimator under model misspecification. We use desparsified lasso procedure in Van De Geer et al. [86] to derive the high-dimensional inference of the interaction parameter in penalized method for precision medicine since the baseline parameters are not of interest. The contrast model is correctly specified while the baseline model is not. We specifically research the effect of the misspecification of the baseline model upon the high-dimensional inference of the contrast model. We study the misspecification of one component in the regression model upon the high-dimensional inference of the other component. In Buhlmann and Van De Geer [12], it researches the effect of the whole misspecified linear model upon the high-dimensional inference of the whole working model. Our research is more specific in terms of the component in the model that is misspecified. We are only interested in the high-dimensional inference of contrast model. The baseline model can be nonlinear which introduces robustness into the parameter estimate in the contrast model. In Bunea et al.
[14], it requires that the distance between the misspecified linear model and the true model should be upper bounded. In Buhlmann and Van De Geer [12], it assumes that the error of the misspecified regression model has bounded moments, which implies that the degree of model misspecification is controlled.

Using desparsifying lasso approach, we are able to pursue the post-selection inference of parameters in penalized method for precision medicine accounting for the misspecification of baseline model. It helps the researchers gain an insight into how reliable the decision from penalized method for precision medicine is even when we posit wrong baseline model. In our research, we provide theoretical work, simulation and related application for the desparsified lasso estimator in penalized method for precision medicine.

3.2 Method and Theory

3.2.1 Desparsified Lasso estimator with misspecified baseline function

Denote $\mathbf{X}$ as the $n \times p$ design matrix, where $n$ is the number of patients and $p$ the number of patients’ features. We consider the setting with $p > n$. Let $\tilde{\mathbf{X}}$ be the design matrix with an additional first column of 1’s. Further, denote $\mathbf{Y}$ as the observed outcome for the patients and $\mathbf{A}$ as the observed treatment assignments for the patients. The conditional mean of the clinical outcome is posited as

$$Y_i = \mu(\tilde{\mathbf{X}}_i) + A_i(\beta^T_0 \tilde{\mathbf{X}}_i) + \epsilon_i,$$  \hspace{1cm} (3.1)
where $\mu(\tilde{X}_i)$ is the unspecified baseline function and $\beta_0$ the unknown coefficients for the interactions of treatment and patients’ variants. We are interested in the inference of $\beta_0$. In real practice, the baseline function $\mu(\tilde{X})$ is unknown and can be misspecified. Therefore, it is crucial to investigate the robustness of the estimate of $\beta_0$ with misspecified baseline function.

In this paper, we involve the propensity score in the interaction term of our regression model [85]. The propensity score is defined as $\pi(\tilde{X}) = P(A=1|\tilde{X}) = E(A|\tilde{X})$. We consider the completely randomized study with pre-specified propensity score $\pi(\tilde{X}) = \pi$. Utilizing the information of the propensity score, we transform the interaction term from $A_i(\beta^T_0 \tilde{X}_i)$ to $(A_i - \pi)(\beta^T_0 \tilde{X}_i)$. Since $E(A_i - \pi|\tilde{X}) = 0$, the transformed interaction term is orthogonal to the baseline function $\mu(\tilde{X})$ given $X$. By doing so, we could protect the estimation of $\beta_0$ from the effect of baseline model misspecification [85]. For notation simplicity, let $\mu_\pi(\tilde{X}) = \mu(\tilde{X}_i) + \pi(\beta^T_0 \tilde{X}_i)$. Therefore, we have

$$Y_i = \mu_\pi(\tilde{X}_i) + (A_i - \pi)(\beta^T_0 \tilde{X}_i) + \epsilon_i. \quad (3.2)$$

Let $\hat{\mu}_\pi(\cdot)$ be an estimator of $\mu_\pi(\cdot)$. The estimator $\hat{\mu}_\pi(\cdot)$ does not have to converge to the true baseline function. Assume $\hat{\mu}_\pi(\cdot) \rightarrow \mu^*_\pi(\cdot)$ uniformly for some function $\mu^*_\pi(\cdot)$. Since $E(A_i - \pi|\tilde{X}) = 0$, then $\mu^*_\pi(\tilde{X}_i)$ and $(A_i - \pi)\beta^T_0 \tilde{X}_i$ are orthogonal, which implies that

$$\beta_0 = \arg\min_{\hat{\beta}} E \left[ Y - \mu^*_\pi(\tilde{X}) - d(A, \pi)\tilde{X} \hat{\beta} \right]^2, \quad (3.3)$$

where $d(A, \pi) = diag \{A_1 - \pi, \ldots, A_n - \pi\}$. Eq. 3.3 tells us that inference of $\beta_0$ based on minimizing mean square error will not be affected much by misspecification of $\mu_\pi(\cdot)$. 

70
We apply lasso to $Y - \hat{\mu}_\pi(\tilde{X})$ versus $d(A, \pi)\tilde{X}$ and get the lasso solution for $\beta_0$ as

$$
\hat{\beta} = \arg\min_{\beta} \left\{ \| Y - \hat{\mu}_\pi(\tilde{X}) - d(A, \pi)\tilde{X}\beta \|^2_2 / n + 2\lambda_{n,p} \| \beta \|_1 \right\},
$$

(3.4)

where $\lambda_{n,p} > 0$ is the tuning parameters. It is well-known that lasso estimator is biased. Furthermore, the true baseline function $\mu_\pi$ is misspecified by $\hat{\mu}_\pi$, which may not even converge to $\mu_\pi$. Therefore, it is difficult to perform inference on $\beta_0$ directly using the limiting distribution of $\hat{\beta}$.

We develop an estimator by desparsifying the lasso estimator, which is similar to the procedures in Van De Geer et al. [86] and [100]. We show that this desparsified lasso estimator is asymptotically unbiased and derive its limiting distribution when baseline function is misspecified. Our desparsified lasso estimator is constructed as

$$
\hat{b}_\beta = \hat{\beta} + \hat{\Theta} \left\{ d(A, \pi)\tilde{X} \right\}^T \left\{ Y - \hat{\mu}_\pi(\tilde{X}) - d(A, \pi)\tilde{X}\hat{\beta} \right\} / n,
$$

(3.5)

where $\hat{\Theta}$ is an estimator of the precision matrix of $d(A, \pi)\tilde{X}$.

### 3.2.2 Limiting distribution of the desparsified Lasso estimator

Our main goal in this section is to derive limiting distribution of $\sqrt{n}H(\hat{b}_\beta - \beta_0)$, where $H$ is a pre-specified $q \times (p + 1)$ matrix with $q = O(1)$, i.e. $q$ is not increasing with $n$. For example, if $H = e_j^T$, where $e_j$ is the $j$th unit column vector, $\sqrt{n}H(\hat{b}_\beta - \beta_0) = \sqrt{n}(\hat{b}_{\beta,j} - \beta_{0,j})$.

We allow arbitrary $H$ with dimension $q \times (p + 1)$ so that we can study the joint distribution of $q$ arbitrary linear contrasts of $\{\sqrt{n}(\hat{b}_{\beta,1} - \beta_{01}), \ldots, \sqrt{n}(\hat{b}_{\beta,p} - \beta_{0p})\}$.

We first decompose $\hat{b}_\beta - \beta_0$ into three terms. Define $\Sigma$ as the covariance matrix of
\[ d(A, \pi) \tilde{X} \] and \( \hat{\Sigma} \) as the sample version of \( \Sigma \). Then

\[ \hat{b}_\beta - \beta_0 = \eta - \Delta_1 - \Delta_2, \quad (3.6) \]

where

\[ \eta = \tilde{\Theta} \{ d(A, \pi) \tilde{X} \}^T \{ Y - \mu^*_{\pi}(\tilde{X}) - d(A, \pi) \tilde{X} \beta_0 \} / n, \quad (3.7) \]

\[ \Delta_1 = \tilde{\Theta} \{ d(A, \pi) \tilde{X} \}^T \{ (\hat{\mu}_{\pi}(\tilde{X}) - \mu^*_{\pi}(\tilde{X})) / n \}, \quad (3.8) \]

\[ \Delta_2 = (\hat{\Theta} \hat{\Sigma} - I)(\hat{\beta} - \beta_0). \quad (3.9) \]

In this section, we derive the limiting distribution of \( \sqrt{n} H (\hat{b}_\beta - \beta_0) \) by showing that \( \| \sqrt{n} H \Delta_1 \|_\infty = o_p(1) \), \( \| \sqrt{n} H \Delta_2 \|_\infty = o_p(1) \), and \( \sqrt{n} H \eta \to_d N(0, G) \).

We first show a preliminary result on the lasso solution \( \hat{\beta} \) with misspecified baseline function. As \( \hat{\beta} - \beta_0 \) is an important components in \( \Delta_2 \), this preliminary result helps to derive the asymptotic properties of \( \sqrt{n} H \Delta_2 \). Consider the following assumptions, where \( C \) denotes a generic constant, possibly varying from place to place:

(A1): The random error \( \varepsilon_i \) of the true model (3.1) are independent and distributed with \( E(\varepsilon_i | \tilde{X}, A) = 0 \) and \( E(\varepsilon_i^2 | \tilde{X}, A) \leq C \).

(A2): There exists a \( \mu^*_{\pi}(\cdot) \) such that

(a) \( \sup_{\tilde{X}_i} | \hat{\mu}_{\pi}(\tilde{X}_i) - \mu^*_{\pi}(\tilde{X}_i) | = o_p(1) \) and

(b) \( E[(\mu^*_{\pi}(\tilde{X}_i) - \mu_{\pi}(\tilde{X}_i))^2] \leq C \).

(A3): Denote \( \tilde{X}^j \) as the \( j \)th column of \( \tilde{X} \). \( \| \tilde{X}^j \|_\infty \leq C \) for all \( 1 \leq j \leq p + 1 \).

(A4): \( \Sigma \) has smallest eigenvalue \( \Lambda_{min}(\Sigma) \geq C > 0 \).

(A5): The number of important interaction terms satisfies that \( s_0 = \| \beta_0 \|_0 = o(\sqrt{n} / \log(p)) \).
Lemma 3.2.1. Consider model (3.1). Assume conditions (A1) - (A5). The lasso solution from (3.4) with \( \lambda_n \approx \frac{p \log p}{\sqrt{n}} \) satisfies

\[
\| \hat{\beta} - \beta_0 \|_1 = o_p(1/\sqrt{\log p}). \tag{3.10}
\]

The convergence rate of lasso estimator \( \hat{\beta} \) is similar to that shown in Buhlmann and Van De Geer [12]. The conditions are imposed upon the error terms of true model in our assumptions rather than misspecified model. In Buhlmann and Van De Geer [12], assumptions of error terms are imposed upon misspecified linear model. The conditions upon the design matrix and covariance matrix are similar to that in Buhlmann and Van De Geer [12]. The assumptions upon the baseline model are needed for our proof.

Proof of lemma 3.2.1:

Define

\[
\epsilon_n = \epsilon - [\mu_{\pi}(\tilde{X}) - \mu_{\pi}(\tilde{X})]. \tag{3.11}
\]

We have

\[
\epsilon_n^T d(A, \pi) \tilde{X} = (\epsilon - [\mu^*(\tilde{X}) - \mu_{\pi}(\tilde{X})] - [\hat{\mu}(\tilde{X}) - \mu^*(\tilde{X})])^T d(A, \pi) \tilde{X} = \epsilon^T d(A, \pi) \tilde{X} - [\mu^*(\tilde{X}) - \mu_{\pi}(\tilde{X})]^T d(A, \pi) \tilde{X} \tag{3.12}
\]

For the first term of Eq. 3.12, it is clear that by (A1),

\[
E(\epsilon_n^T d(A, \pi) \tilde{X}) = E E(\epsilon^T d(A, \pi) \tilde{X} | A, \tilde{X}) = 0.
\]
Further, by moment inequality (from Chapter 6.2.2 of Buhlmann and Van De Geer [11]),

$$E(\|e^T d(A, \pi) \bar{X}\|_\infty^2) \leq 8 \log(2p) \sum_{i=1}^{n} (A_i - \pi)^2 (\max_{1 \leq j \leq p} |X_i^j|)^2 E(\varepsilon_i^2) \leq C n \log(p),$$

where the second inequality is by (A1) and (A3).

Then, by Markov inequality,

$$P(\|e^T d(A, \pi) \bar{X}\|_\infty^2 > n \log(p)) \leq \frac{1}{n \log(p)} E(\|e^T d(A, \pi) \bar{X}\|_\infty^2) \leq C,$$

where $C$ is a generic constant.

Then

$$\|e^T d(A, \pi) \bar{X}\|_\infty^2 = O_p\{n \log(p)\}.\)

Then

$$\|2e^T d(A, \pi) \bar{X} / n\|_\infty = O_p(\sqrt{\log(p)}/\sqrt{n}). \quad (3.13)$$

Next, consider the second term of Eq. 3.12. It is easy to see that since $E[d(A, \pi) \bar{X}] = 0$, we have

$$E([\mu^*_{\pi}(\bar{X}) - \mu_{\pi}(\bar{X})]^T d(A, \pi) \bar{X}) = E E([\mu^*_\pi(\bar{X}) - \mu_{\pi}(\bar{X})]^T d(A, \pi) \bar{X} | \bar{X}) = 0.$$

Then, similar arguments as those leading to Eq. 3.13 combined with (A2b) and (A3) gives

$$\|2[\mu^*_\pi(\bar{X}) - \mu_{\pi}(\bar{X})]^T d(A, \pi) \bar{X} / n\|_\infty = O_p(\sqrt{\log(p)}/\sqrt{n}). \quad (3.14)$$
Finally, by (A2a), the third term of Eq. 3.12

\[ \|2[\hat{\mu}_\pi(\tilde{X}) - \mu_\pi^\star(\tilde{X})]^T d(A, \pi)\tilde{X}/n\|_\infty = o_p(\sqrt{\log(p)}/\sqrt{n}). \] (3.15)

Combining Eq. 3.13 - Eq. 3.15 gives that

\[ \|2\epsilon_\mu^T d(A, \pi)\tilde{X}/n\|_\infty = O_p(\sqrt{\log p}/\sqrt{n}). \] (3.16)

By the construction of \( \hat{\beta} \), we have

\[ \|Y - \hat{\mu}_\pi(\tilde{X}) - d(A, \pi)\tilde{X}\hat{\beta}\|_2^2/n + 2\lambda_{n,p}\|\hat{\beta}\|_1 \leq \|Y - \hat{\mu}_\pi(\tilde{X}) - d(A, \pi)\tilde{X}\beta_0\|_2^2/n + 2\lambda_{n,p}\|\beta_0\|_1. \] (3.17)

By true model in (3.2), we have

\[ \|\mu_\pi(\tilde{X}) + d(A, \pi)\tilde{X}\beta_0 + \epsilon - \hat{\mu}_\pi(\tilde{X}) - d(A, \pi)\tilde{X}\hat{\beta}\|_2^2/n + 2\lambda_{n,p}\|\hat{\beta}\|_1 \leq \|\mu_\pi(\tilde{X}) + d(A, \pi)\tilde{X}\beta_0 + \epsilon - \hat{\mu}_\pi(\tilde{X}) - d(A, \pi)\tilde{X}\beta_0\|_2^2/n + 2\lambda_{n,p}\|\beta_0\|_1. \] (3.18)

Then

\[ \|\mu_\pi(\tilde{X}) + d(A, \pi)\tilde{X}\beta_0 + \epsilon - \hat{\mu}_\pi(\tilde{X}) - d(A, \pi)\tilde{X}\hat{\beta}\|_2^2/n + 2\lambda_{n,p}\|\hat{\beta}\|_1 \leq \|\mu_\pi(\tilde{X}) + \epsilon - \hat{\mu}_\pi(\tilde{X})\|_2^2/n + 2\lambda_{n,p}\|\beta_0\|_1. \] (3.19)

\[ \leq \|\mu_\pi(\tilde{X}) + \epsilon - \hat{\mu}_\pi(\tilde{X})\|_2^2/n + 2\lambda_{n,p}\|\beta_0\|_1. \] (3.20)
Recall the definition of $\epsilon_\hat{\mu}$ in Eq. 3.11, then

$$
\| d(A, \pi) \tilde{X} \beta_0 + \epsilon_\hat{\mu} - d(A, \pi) \tilde{X} \hat{\beta} \|^2 / n + 2 \lambda_{n,p} \| \hat{\beta} \|_1
$$

(3.21)

$$
\leq \| \epsilon_\hat{\mu} \|^2 / n + 2 \lambda_{n,p} \| \beta_0 \|_1.
$$

(3.22)

Then

$$
\| \epsilon_\hat{\mu} - d(A, \pi) \tilde{X} (\hat{\beta} - \beta_0) \|^2 / n + 2 \lambda_{n,p} \| \hat{\beta} \|_1
$$

(3.23)

$$
\leq \| \epsilon_\hat{\mu} \|^2 / n + 2 \lambda_{n,p} \| \beta_0 \|_1.
$$

(3.24)

Then

$$
\| \epsilon_\hat{\mu} \|^2 / n + \| d(A, \pi) \tilde{X} (\hat{\beta} - \beta_0) \|^2 / n - 2 \epsilon_\hat{\mu} \epsilon_\hat{\mu}^T d(A, \pi) \tilde{X} (\hat{\beta} - \beta_0) / n + 2 \lambda_{n,p} \| \hat{\beta} \|_1
$$

(3.25)

$$
\leq \| \epsilon_\hat{\mu} \|^2 / n + 2 \lambda_{n,p} \| \beta_0 \|_1.
$$

(3.26)

Then

$$
\| d(A, \pi) \tilde{X} (\hat{\beta} - \beta_0) \|^2 / n - 2 \epsilon_\hat{\mu} \epsilon_\hat{\mu}^T d(A, \pi) \tilde{X} (\hat{\beta} - \beta_0) / n + 2 \lambda_{n,p} \| \hat{\beta} \|_1
$$

(3.27)

$$
\leq 2 \lambda_{n,p} \| \beta_0 \|_1.
$$

(3.28)

Then

$$
\| d(A, \pi) \tilde{X} (\hat{\beta} - \beta_0) \|^2 / n + 2 \lambda_{n,p} \| \hat{\beta} \|_1
\leq 2 \epsilon_\hat{\mu} \epsilon_\hat{\mu}^T d(A, \pi) \tilde{X} (\hat{\beta} - \beta_0) / n + 2 \lambda_{n,p} \| \beta_0 \|_1.
$$

(3.29)
The first term of the right-hand side

\[ 2\epsilon_T d(A, \pi) \tilde{X}(\hat{\beta} - \beta_0)/n \leq \|2\epsilon_T d(A, \pi) \tilde{X}/n\|_\infty \|\hat{\beta} - \beta_0\|_1, \]

where the order of \( \|2\epsilon_T d(A, \pi) \tilde{X}/n\|_\infty \) is derived in Eq. 3.16.

The rest of the proof is similar to the proof of the second claim of Lemma 2 in Buhlmann and Van De Geer [12], where it is shown that given (A3), (A4), and (A5), the compatibility condition holds with probability tending to 1. Then using Eq. 3.16 and similar arguments in Section 6.2.2 of Buhlmann and Van De Geer [11], the oracle inequality in Eq. 3.10 holds. The details are as follows.

Define \( \beta_{j,S} = \beta_j 1(j \in S) \) and \( \beta_{j,S^c} = \beta_j 1(j \in S^c) \). Then \( \beta = \beta_{j,S} + \beta_{j,S^c} \).

Define \( S_0 \) to be index of true interaction. Then \( \beta_{0,S_0^c} = 0 \).

For the left-hand side,

\[ \|\hat{\beta}\|_1 = \|\hat{\beta}_{S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1 \geq \|\beta_{0,S_0}\|_1 - \|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1. \]

For the right-hand side,

\[ \|\hat{\beta} - \beta_0\|_1 = \|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1. \]

Then considering left-hand and right-hand side, we have

\[ \|d(A, \pi) \tilde{X}(\hat{\beta} - \beta_0)\|_2^2/n + 2\lambda_{n,p}(\|\beta_{0,S_0}\|_1 - \|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1) \leq \]  
\[ O_p(\sqrt{\log p/\sqrt{n}})(\|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1) + 2\lambda_{n,p}\|\beta_0\|_1. \]  

(3.29)
Recall that \( \| \beta_0 \|_1 = \| \beta_{0,S_0} \|_1 \), we have

\[
\| d(A, \pi) \hat{X}(\hat{\beta} - \beta_0) \|_2^2 / n - 2\lambda_{n,p}(\| \hat{\beta}_{S_0} - \beta_{0,S_0} \|_1 - \| \hat{\beta}_{S_0^\perp} \|_1) \leq O_p(\sqrt{\log p / \sqrt{n}})(\| \hat{\beta}_{S_0} - \beta_{0,S_0} \|_1 + \| \hat{\beta}_{S_0} \|_1).
\]

(3.31)

Recall that \( \lambda_{n,p} = O(\sqrt{\log p / \sqrt{n}}) \), we have

\[
\| d(A, \pi) \hat{X}(\hat{\beta} - \beta_0) \|_2^2 / n + \lambda_{n,p} \| \hat{\beta}_{S_0} - \beta_{0,S_0} \|_1 \leq O_p(\sqrt{\log p / \sqrt{n}})\| \hat{\beta}_{S_0} \|_1.
\]

(3.32)

Then

\[
\| d(A, \pi) \hat{X}(\hat{\beta} - \beta_0) \|_2^2 / n + \lambda_{n,p} \| \hat{\beta} - \beta_0 \|_1 \leq O_p(\sqrt{\log p / \sqrt{n}})\| \hat{\beta}_{S_0} \|_1.
\]

(3.33)

\[
= \| d(A, \pi) \hat{X}(\hat{\beta} - \beta_0) \|_2^2 / n + \lambda_{n,p} \| \hat{\beta}_{S_0} - \beta_{0,S_0} \|_1 + \lambda_{n,p} \| \hat{\beta}_{S_0^\perp} \|_1 \leq O_p(\sqrt{\log p / \sqrt{n}})\| \hat{\beta}_{S_0} \|_1.
\]

(3.36)

In (A4), we assume that \( \Sigma \) is positive definite. We use estimated covariance matrix \( \hat{\Sigma} \) such that \( \| \hat{\Sigma} - \Sigma \|_\infty = O_p(\sqrt{\log p / \sqrt{n}}) \). From chapter 6.12 of Buhlmann and Van De Geer [11], the compatibility condition holds with probability tending to 1. For some \( \phi_0 > 0 \), we have

\[
\| \hat{\beta}_{S_0^\perp} \|_1^2 \leq \frac{1}{n} \| d(A, \pi) \hat{X}(\hat{\beta} - \beta_0) \|_2^2 \frac{S_0}{\phi_0^2}.
\]
Then

\begin{align}
\|d(A, \pi)\hat{\beta} - \beta_0\|_1^2 \leq & O_p(\sqrt{\log p} / \sqrt{n}) ||\hat{\beta} - \beta_0||_1 \\
\leq O_p(\sqrt{\log p} / \sqrt{n}) & \frac{1}{\sqrt{n}} \|d(A, \pi)\hat{\beta} - \beta_0\|_2 \frac{\sqrt{s_0}}{\phi_0} \\
\leq & \|d(A, \pi)\hat{\beta} - \beta_0\|_2^2 / n + O_p(\sqrt{\log p} / \sqrt{n})^2 s_0 / \phi_0^2. \quad (3.38)
\end{align}

Then by (A5), we have

\[ \|\hat{\beta} - \beta_0\|_1 \leq O_p(\sqrt{\log p} / \sqrt{n}) o \left( \frac{\sqrt{n}}{\log p} \right) = o_p(1 / \sqrt{\log p}). \]  

Next, we show that \( \sqrt{n}H\Delta_1 \) and \( \sqrt{n}H\Delta_2 \) are at the order of \( o_p(1) \). Similar to Van De Geer et al. [86], we apply lasso for nodewise regression to construct \( \hat{\Theta} \). Consider additional assumptions:

(A6) The pre-specified matrix \( H_{q \times (p+1)} \) satisfies

(a) \( ||H||_\infty \leq C \) and

(b) \( h_t = |\{ j \neq t : H_{t,j} \neq 0 \}| \leq C \) for any \( t \in \{1, \ldots, q\} \).

(A7): The precision matrix \( \Theta \) of \( d(A, \pi)\hat{X} \) satisfies

(a) \( ||\Theta||_\infty \leq C \) and

(b) \( s_j = |\{ k \neq j : \Theta_{j,k} \neq 0 \}| \leq C \) for any \( j \in \{1, \ldots, p\} \).

**Lemma 3.2.2.** Consider model (3.1). Assume conditions (A1) - (A7). Obtain \( \hat{\beta} \) by (3.4) with \( \lambda_{n,p} \geq \sqrt{\log(p) / n} \) and \( \hat{\Theta} \) by nodewise regression with \( \lambda_j \geq \sqrt{\log(p) / n} \) for any \( j \in \{1, \ldots, p\} \). Then

\[ \|\sqrt{n}H\Delta_1\|_\infty = o_p(1) \quad (3.41) \]
\[
\| \sqrt{n} H \Delta_1 \|_\infty = o_p(1) \quad (3.42)
\]

Proof of lemma 3.2.2:

Consider (3.41) first. Recall

\[
\sqrt{n} H \Delta_1 = \frac{1}{\sqrt{n}} H \hat{\Theta} \{ d(A_1, \pi) \hat{X} \}^T \{ \hat{\mu}_\pi(\hat{X}) - \mu^*_\pi(\hat{X}) \}.
\]

Rewrite \( \sqrt{n} H \Delta_1 \) as

\[
\| \sqrt{n} H \Delta_1 \|_\infty = \left\| \frac{1}{\sqrt{n}} H \hat{\Theta} \frac{1}{\sqrt{n}} \sum_{i=1}^n (A_i - \pi) \hat{X}_i \{ \hat{\mu}_\pi(\hat{X}_i) - \mu^*_\pi(\hat{X}_i) \} \right\|_\infty.
\]

\[
\leq \sup_{\hat{X}_i} |\hat{\mu}_\pi(\hat{X}_i) - \mu^*_\pi(\hat{X}_i)| \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n (A_i - \pi) \hat{X}_i \right\|_\infty.
\]

Since E[(A_i - \pi)\hat{X}_i] = 0 and \( \| E[(A_i - \pi)^2 \hat{X}_i \hat{X}_i^T] \|_\infty < \infty \) by (A3), by Multivariate Central Limit Theorem, the \((p + 1)\)-vector

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^n (A_i - \pi) \hat{X}_i = o_p(1).
\]

The \(j\)th element of \( H \hat{\Theta} \frac{1}{\sqrt{n}} \sum_{i=1}^n (A_i - \pi) \hat{X}_i \) is

\[
\begin{align*}
&\sum_{k=1}^{p+1} \sum_{l=1}^{p+1} H_{jk} \hat{\Theta}_{kl} \frac{1}{\sqrt{n}} \sum_{i=1}^n (A_i - \pi) \hat{X}_{li} \\
&= \sum_{k=1}^{p+1} \sum_{l=1}^{p+1} H_{jk} \Theta_{kl} \frac{1}{\sqrt{n}} \sum_{i=1}^n (A_i - \pi) \hat{X}_{li} \\
&\quad + \sum_{k=1}^{p+1} \sum_{l=1}^{p+1} H_{jk} (\hat{\Theta}_{kl} - \Theta_{kl}) \frac{1}{\sqrt{n}} \sum_{i=1}^n (A_i - \pi) \hat{X}_{li}.
\end{align*}
\]
By (A6) and (A7), we have
\[
\sum_{k=1}^{p+1} \sum_{l=1}^{p+1} H_{jk} \Theta_{kl} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (A_i - \pi) \tilde{X}_{li} \leq C \sum_{k=1}^{p+1} H_{jk} s_k O(1) \leq C h_j O(1) = O(1). \tag{3.46}
\]

On the other hand, (A7) combined with similar arguments as in the proof of the first claim of Lemma 2 in Buhlmann and Van De Geer [12] gives
\[
\|\hat{\Theta}^j - \Theta^j\|_1 = o_p(1/\sqrt{\log p}).
\]

The second term
\[
\sum_{k=1}^{p+1} \sum_{l=1}^{p+1} H_{jk} (\hat{\Theta}_{kl} - \Theta_{kl}) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (A_i - \pi) \tilde{X}_{li} \tag{3.47}
\]

\[
\leq \sum_{k=1}^{p+1} H_{jk} o_p(1/\sqrt{\log p}) O(1) \leq C h_j o_p(1/\sqrt{\log p}) O(1) = o_p(1/\sqrt{\log p}). \tag{3.48}
\]

The $j$th element of $H \hat{\Theta} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (A_i - \pi) \tilde{X}_i$ is $O_p(1)$. We know that $H \hat{\Theta} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (A_i - \pi) \tilde{X}_i$ is $q \times 1$, where $q = O(1)$. Summarizing the above with (A6) and Slutsky’s Theorem gives
\[
\left\| H \hat{\Theta} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (A_i - \pi) \tilde{X}_i \right\|_\infty = O_p(1). \tag{3.49}
\]

(3.41) follows from (3.49) and (A2a).

Next, consider (3.42).
\[
\|\sqrt{n}H \Delta\|_\infty = \|\sqrt{n}H(\hat{\Theta} - \Theta - I)(\hat{\beta} - \beta_0)\|_\infty \leq \sqrt{n}||H(\hat{\Theta} - \Theta - I)||_\infty \|\hat{\beta} - \beta_0\|_1.
\]
By (A6),

$$
\| H(\hat{\Theta}\hat{\Sigma} - I) \|_\infty = \max_{l,k} \sum_{j=1}^{p+1} H_{lj}(\hat{\Theta}\hat{\Sigma} - I)_{jk}
\leq \max_l \sum_{j=1}^{p+1} H_{lj} \| \hat{\Theta}\hat{\Sigma} - I \|_\infty
\leq \max_l h_l \| \hat{\Theta}\hat{\Sigma} - I \|_\infty \leq C \| \hat{\Theta}\hat{\Sigma} - I \|_\infty. \tag{3.50}
$$

Similar arguments as those leading to (10) in Van De Geer et al. [86], combined with (A3) and (A7), result in $\| \hat{\Theta}\hat{\Sigma} - I \|_\infty = O_p(\sqrt{\log p}/\sqrt{n})$. Combing the above with Lemma 3.2.1 gives (3.42).

We have

$$
\sqrt{n}H\Delta_2 \|_\infty \leq \sqrt{n} \| H(\hat{\Theta}\hat{\Sigma} - I) \|_\infty \| \hat{\beta} - \beta_0 \|_1 \leq C \sqrt{n}O_p(\sqrt{\log p}/\sqrt{n})o_p(1/\sqrt{\log p}) = o_p(1).
$$

The remaining component of $\sqrt{n}H(\hat{\beta} - \beta_0)$ is $\sqrt{n}H\eta$. We show that $\sqrt{n}H\eta$ converges to a multivariate normal distribution. Since $\eta$ does not involve $\hat{\beta} - \beta_0$, less conditions are needed in the following lemma.

**Lemma 3.2.3.** Consider model (3.1). Assume conditions (A1)-(A3) and (A6)-(A7). Obtain $\hat{\Theta}$ by nodewise regression with $\lambda_j \asymp \log(p)/n$ for any $j \in \{1, \ldots, p\}$. Then

$$
\sqrt{n}H\eta \rightarrow_d N(0, G), \tag{3.53}
$$

where $G = H\Theta D\Theta^T H^T$ is a $q \times q$ nonnegative matrix, $D = E[(A_i - \pi)^2\sigma_\epsilon^2(A_i, \tilde{X}_i)\tilde{X}_i\tilde{X}_i^T] + \pi(1 - \pi)E[(\mu_\pi(\tilde{X}_i) - \mu_\pi^*(\tilde{X}_i))^2\tilde{X}_i\tilde{X}_i^T]$, $\sigma_\epsilon^2(A_i, \tilde{X}_i) = \text{Var}(\epsilon_i|A_i, \tilde{X}_i)$, and $\|G\|_\infty < \infty$. 

82
Proof of lemma 3.2.3:

We first show $E(\sqrt{n}H\eta) = 0$. By definition of $\eta$, we have

$$E(\sqrt{n}H\eta) = E\left[\frac{1}{\sqrt{n}}H\hat{\Theta}\{d(A, \pi)\bar{X}\}^T(Y - \mu_0^\pi\bar{X} - d(A, \pi)\bar{X}^T\bar{X})\right]$$

$$= \frac{1}{\sqrt{n}}HE[\hat{\Theta}\{d(A, \pi)\bar{X}\}^T\epsilon] + \frac{1}{\sqrt{n}}HE[\hat{\Theta}(d(A, \pi)\bar{X})^T(\mu_0^\pi\bar{X} - \mu_0^\pi\bar{X})]$$

The first term of the above is equal to 0 because $E(\epsilon|\bar{X}, A) = 0$ by (A1). The second term is equal to 0 since $E(d(A, \pi)|\bar{X}) = 0$.

Next, rewrite $\sqrt{n}H\eta$ as

$$\sqrt{n}H\eta = H\hat{\Theta}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_i,$$

where

$$W_i = (A_i - \pi)(y_i - \mu_0^\pi(\bar{X}_i) - (A_i - \pi)\beta_0^T\bar{X}_i)\bar{X}_i$$

The second moment of $W_i$ is

$$E\left[(A_i - \pi)^2(\epsilon_i + \mu_0^\pi(\bar{X}_i) - \mu_0^\pi(\bar{X}_i))^2\bar{X}_i\bar{X}_i^T\right] = E\left[(A_i - \pi)^2\epsilon_i^2\bar{X}_i\bar{X}_i^T\right] + E\left[(A_i - \pi)^2(\mu_0^\pi(\bar{X}_i) - \mu_0^\pi(\bar{X}_i))^2\bar{X}_i\bar{X}_i^T\right].$$

The first term

$$E\left[(A_i - \pi)^2\epsilon_i^2\bar{X}_i\bar{X}_i^T\right] = E\left[(A_i - \pi)^2\epsilon_i^2|\bar{X}, A\bar{X}_i\bar{X}_i^T\right] = E\left[(A_i - \pi)^2\epsilon_i^2|\bar{X}_i\bar{X}_i^T\right]$$
The second term

\[
E \left[ (A_i - \pi)^2 (\mu_{\pi}(\tilde{X}_i) - \mu^*_{\pi}(\tilde{X}_i))^2 \tilde{X}_i \tilde{X}_i^T \right] = E \left[ E((A_i - \pi)^2 | \tilde{X}_i) (\mu_{\pi}(\tilde{X}_i) - \mu^*_{\pi}(\tilde{X}_i))^2 \tilde{X}_i \tilde{X}_i^T \right]
\]

\[
= \pi (1-\pi) E \left[ (\mu_{\pi}(\tilde{X}_i) - \mu^*_{\pi}(\tilde{X}_i))^2 \tilde{X}_i \tilde{X}_i^T \right].
\]

Then the second moment of \( W_i \) is \( D \). Note that \( \sigma_i^2(A_i, \tilde{X}_i) < \infty \) by (A1). This combined (A2b) and (A3) gives \( \| D \|_\infty < \infty \). Consequently, \( \| H \Theta D \Theta^T H^T \|_\infty < \infty \) is implied by (A6)-(A7).

On the other hand, (A7) combined with similar arguments as in the proof of the first claim of Lemma 2 in Buhlmann and Van De Geer [12] gives \( \| \hat{\Theta}_j - \Theta_j \|_1 = o_p(1/\sqrt{\log p}) \). By Multivariate Central Limit Theorem and Slutsky’s Theorem,

\[
H \hat{\Theta} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \rightarrow_d N(0, H \Theta D \Theta^T H^T).
\]

Result in (3.53) follows.

We summarize the above results in the following theorem.

**Theorem 3.2.1.** Consider model (3.1). Assume (A1) - (A7). Obtain \( \hat{\beta} \) by (3.4) with \( \lambda_{n,p} \approx \sqrt{\log(p)/n} \) and \( \hat{\Theta} \) by nodewise regression with \( \lambda_j \approx \sqrt{\log(p)/n} \) for any \( j \in \{1, \ldots, p\} \). Then,

\[
\sqrt{n} (H \hat{\beta} - \beta_0) \rightarrow_d N(0, G),
\]

where \( G \) is defined in (3.53).

**Remark:** Compared to Buhlmann and Van De Geer [12] where conditions on the error term of the misspecified linear model are required, our study requires conditions only on the error term of the true model. The limiting distribution in theorem 3.2.1 is robust to the
specification of baseline model.

### 3.2.3 Interval estimation and hypothesis testing

The theoretical result on the asymptotic normality of the desparsified estimator $\hat{\beta}$ can be utilized to construct confidence intervals for the coefficients of interest and to perform hypothesis testing. The variance-covariance matrix $G$ can be approximated by $H\hat{\Theta}\hat{\Theta}^T H^T$ where

$$\hat{D} = \frac{1}{n} \sum_{i=1}^{n} (A_i - \pi)^2 \left\{ Y_i - \hat{\mu}_\pi(\tilde{X}_i) - (A_i - \pi)\hat{\beta}^T \tilde{X}_{ij} \right\}^2 \tilde{X}_i\tilde{X}_i^T. \quad (3.54)$$

Therefore, a pointwise $(1 - \alpha)$ confidence interval of $\beta_{0,j}$ can be constructed as

$$\{ \hat{\beta}_{\beta,j} - c(\alpha, n), \hat{\beta}_{\beta,j} + c(\alpha, n) \}, \quad (3.55)$$

where $c(\alpha, n) = \Phi^{-1}(1 - \alpha/2)\sqrt{(H\hat{\Theta}\hat{\Theta}^T H^T)_{jj}/n}$, and $\Phi(\cdot)$ is the c.d.f of $N(0, 1)$.

One can also calculate the asymptotic $p$-value for testing

$$H_0 : \beta_{0,j} = 0 \quad \text{vs.} \quad H_A : \beta_{0,j} \neq 0$$

for a given $j \in \{1, \ldots, p\}$. A non-zero $\beta_{0,j}$ means that variant $X_j$ is relevant in making treatment decision for a patient.
3.2.4 Some special case of \( \hat{\mu}_\pi \)

In real applications, we need to choose an estimator \( \hat{\mu}_\pi \) for the baseline function. The only requirement for \( \hat{\mu}_\pi \) is condition (A2), which is very general. Here we discuss two special cases of \( \hat{\mu}_\pi \).

Case 1: \( \hat{\mu}_\pi(X_i) = \) constant.

Case 2: \( \hat{\mu}_\pi(X_i) = \gamma^T X_i \).

In case 1, (A2a) holds trivially. (A2b) is implied by the follows.

\[(B1): E(Y_i^2) \leq C.\]

The following corollary presents the limiting distribution of \( \sqrt{n}H\hat{\beta} \) in this case. Proof of this corollary 3.2.1 is straightforward and, thus, omitted.

**Corollary 3.2.1.** Consider model (3.1). Implement \( \hat{\mu}_\pi(X_i) = \) constant. Assume (A1), (B1), (A3) - (A7). Obtain \( \hat{\beta} \) by (3.4) with \( \lambda_{n,p} \simeq \sqrt{\log(p)/n} \) and \( \hat{\Theta} \) by nodewise regression with \( \lambda_j \simeq \sqrt{\log(p)/n} \) for any \( j \in \{1, \ldots, p\} \). Then,

\[\sqrt{n}H(\hat{\beta} - \beta_0) \rightarrow_d N(0, G),\]

where \( G \) is defined in (3.53).

When \( \hat{\mu}_\pi(X_i) = \hat{\gamma}^T X_i \), where \( \hat{\gamma} \) is obtained by

\[(\hat{\gamma}, \hat{\beta}) = \arg\min_{\gamma, \beta} \left\{ ||Y - X\gamma - d(A, \pi)X\beta||^2_2/n + 2\lambda_{n,p}(||\gamma||_1 + ||\beta||_1) \right\}. \quad (3.56)\]

We will show that the solution \( \hat{\beta} \) from (3.56) is equivalent to the solution from (3.4). In this case, we replace (A2) with the following conditions.

\[\lambda_j \simeq \sqrt{\log(p)/n} \] for any \( j \in \{1, \ldots, p\} \).
(B2): Define \( \gamma^* = \arg\min \gamma E(\mu_{\pi}(\tilde{X}_i) - \gamma^T \tilde{X}_i)^2; \gamma^* \) satisfies

(a) \( s_{\gamma} = \|\gamma^*\|_0 = o(\sqrt{n}/\log(p)) \) and

(b) \( E(\gamma^{*T} \tilde{X}_i - \mu_{\pi}(\tilde{X}_i))^2 \leq C. \)

The proof of the following corollary is not trivial. Details are shown in its proof.

**Corollary 3.2.2.** Consider model (3.1). Obtain \( \hat{\gamma} \) and \( \hat{\beta} \) from (3.56) with \( \lambda_{n,p} \approx \sqrt{\log(p)/n} \). Assume (A1), (B2), (A3) - (A7). Obtain \( \hat{\Theta} \) by nodewise regression with \( \lambda_{j} \approx \sqrt{\log(p)/n} \) for any \( j \in \{1, \ldots, p\} \). Then,

\[
\sqrt{n}H(\hat{\beta} - \beta_0) \overset{d}{\to} N(0, G),
\]

where \( G \) is defined in (3.53).

Remark: In Buhlmann and Van De Geer [12], the whole linear model is misspecified and the limiting distribution of desparsified lasso estimator is derived. In corollary 3.2.2, the baseline model is misspecified and the contrast model is correctly specified. The limiting distribution of the desparsified lasso estimator of contrast model coefficient is robust to the misspecification of baseline model.

**Proof of Corollary 3.2.2:**

After obtaining \( \hat{\gamma} \) and \( \hat{\beta} \) from (3.56), implement \( \hat{\mu}_{\pi}(\tilde{X}_i) = \hat{\gamma}^T \tilde{X}_i \). Define

\[
\beta^* = \arg\min_{\beta} \left\{ \|Y - \tilde{X}\hat{\gamma} - d(A, \pi)\tilde{X}\beta\|_2^2/n + 2\lambda_{n,p}\|\beta\|_1 \right\}.
\]

First, we show that \( \hat{\beta} = \beta^* \). By the construction of \( \beta^* \),

\[
\|Y - \tilde{X}\hat{\gamma} - d(A, \pi)\tilde{X}\beta^*\|_2^2/n + 2\lambda_{n,p}\|\beta^*\|_1 \leq \|Y - \tilde{X}\hat{\gamma} - d(A, \pi)\tilde{X}\hat{\beta}\|_2^2/n + 2\lambda_{n,p}\|\hat{\beta}\|_1. \tag{3.57}
\]
On the other hand, the construction of $\hat{\beta}$ implies

$$\|Y - \hat{X} \hat{\gamma} - d(A, \pi) \hat{X} \hat{\beta}\|^2/n + 2\lambda_{n,p}(\|\hat{\gamma}\|_1 + \|\hat{\beta}\|_1)$$

$$\leq \|Y - \hat{X} \hat{\gamma} - d(A, \pi) \hat{X} \beta^*\|^2/n + 2\lambda_{n,p}(\|\hat{\gamma}\|_1 + \|\beta^*\|_1)$$

$$\leq \|Y - \hat{X} \hat{\gamma} - d(A, \pi) \hat{X} \hat{\beta}\|^2/n + 2\lambda_{n,p}(\|\hat{\gamma}\|_1 + \|\hat{\beta}\|_1),$$

where the second inequality is by (3.57). Therefore, By the uniqueness of the solution of convex optimization, $\hat{\beta} = \beta^*$.

Secondly, we show that

$$\|\hat{\gamma} - \gamma^*\|_1 + \|\hat{\beta} - \beta_0\|_1 = o_p(1/\sqrt{\log p}). \quad (3.58)$$

Given the fact that $\mu_{\pi}(\tilde{X}_i)$ and $(A_i - \pi)\beta_0^T \tilde{X}_i$ are orthogonal and $E(\epsilon_i | \tilde{X}) = 0$ from (A1), definition of $\gamma^*$ implies

$$(\gamma^*, \beta_0) = \arg\min_{\gamma, \beta} E(Y_i - \gamma^T \tilde{X}_i - (A_i - \pi)\beta^T \tilde{X}_i)^2.$$ 

Define

$$\xi_i := Y_i - \gamma^*^T \tilde{X}_i - (A_i - \pi)\beta^* \tilde{X}_i = \epsilon_i + \mu_{\pi}(\tilde{X}_i) - \gamma^*^T \tilde{X}_i.$$ 

Note that (A1) and (B2b) implies

$$E(\xi_i)^2 \leq C. \quad (3.59)$$ 

By the second claim of Lemma 2 of Buhlmann and Van De Geer [12], (3.59) combined with (B2a), (A3), (A4) and (A5), gives (3.58). The eigenvalue condition on $\Sigma$ implies the same
condition on augmented $\tilde{\Sigma}$, where $\tilde{\Sigma}$ is block diagonal with 1 and $\Sigma$ on the diagonal. The eigenvalue of $\tilde{\Sigma}$ is 1 and the same eigenvalues of $\Sigma$.

Next, it is easy to see that (A2a) is implied by (3.58) and (A3), and (A2b) holds trivially given (B2b). Therefore, (A2) is satisfied and the rest of the proof is the same as the proof of Theorem 3.2.1.

### 3.3 Simulation

#### 3.3.1 Simulation setting

We study the robustness of our desparsified lasso estimator when the baseline function is linear or nonlinear. We use simulated data from true baseline model in three different cases proposed in Lv et al. [58]. In the first case, data is simulated from the true model with linear baseline function. In the second case, the true baseline model is quadratic. In the third case, the true baseline model is trigonometric.

- **Case 1:** $Y = 1 + X\gamma + A(\tilde{X}\beta) + \epsilon$,

- **Case 2:** $Y = 1 + 0.5(1 + X\gamma)^2 + A(\tilde{X}\beta) + \epsilon$,

- **Case 3:** $Y = 1 + 1.5\sin(\pi X\gamma) + X_1^2 + A(\tilde{X}\beta) + \epsilon$, where $\pi$ is propensity score and $X_1$ is the first covariate in design matrix.

In the three cases above, we set the error term $\epsilon$, covariance matrix of $X$, sample size $n$, dimension $p$, and the parameters $\gamma_j$ and $\beta_j$ as follows:

1. The random error $\epsilon$ is simulated from $N(0,1)$. 
2. The covariance matrix of design matrix $X$ is $\Sigma_X = AR(0.5)$. Design matrix $X$ is simulated from multivariate normal distribution $MN(0, \Sigma_X)$.

3. The sample size $n$ is 200. The number of predictors is $p$, where $p = 100$, 200 or 300.

4. Baseline model parameter $\gamma$ has 2 nonzero coefficients: $\gamma = (1, 1, 0, \cdots, 0)^T$. The intensity is 1.

5. Contrast model parameter $\beta$ has $s$ nonzero coefficients with intensity 1 or 1.5 where $s = 5$ or 10. Contrast model includes treatment main effect and interaction effects between treatment and covariates. The intensity of treatment main effect is zero. Among interaction effects, $s$ of them are nonzero.

6. We have 100 simulations.

We derive the confidence interval of our desparsified lasso estimator $\hat{b}_\beta$ defined in (Eq. 3.5) using (Eq. 3.55). We denote the confidence interval of $\beta_0^j$ as $CI_j$ and let $S_{true} = \{j \in \{2, \ldots, p+1\}: \beta_0^j \neq 0\}$. The following measures are reported in Table 3.1-Table 3.2 to show the coverage property of confidence interval in (Eq. 3.55) and the estimation property of $\hat{b}_\beta$ in (Eq. 3.5).

- Coverage for noise variables: the empirical version of $(p - s)^{-1} \sum_{j \in S_{true}} \mathbb{P}[0 \in CI_j]$.
- Coverage for relevant variables: the empirical version of $s^{-1} \sum_{j \in S_{true}} \mathbb{P}[\beta_0^j \in CI_j]$.
- Length of CI for noise variables: the empirical version of $(p - s)^{-1} \sum_{j \in S_{true}} \text{length}(CI_j)$.
- Length of CI for relevant variable: the empirical version of $s^{-1} \sum_{j \in S_{true}} \text{length}(CI_j)$.
• MAB(\(\hat{b}\)): the mean absolute bias of \(\hat{b}\) for relevant variables, calculated as
\[
s^{-1} \sum_{j \in \mathbb{S}_{true}} |\hat{b}_j - \beta_j^0|.
\]

• MAB(\(\hat{\beta}\)): the mean absolute bias of lasso for relevant variables, calculated as
\[
s^{-1} \sum_{j \in \mathbb{S}_{true}} |\hat{\beta}_j - \beta_j^0|.
\]

### 3.3.2 Coverage of confidence interval and mean absolute bias of desparsified lasso estimator

Table 3.1-Table 3.2 summarize the performance measures for case 1-3 with \(p = 300\) and \(s = 5\) or 10. The results show that the coverage properties of confidence intervals are quite consistent for case 1-3, where case 1 has the linear baseline function and case 2-3 have the nonlinear baseline functions. The coverage for noise variable is better than the coverage for true variables. The length of confidence intervals increases from case 1 to case 3. As the baseline model deviates farther from linear function, the uncertainty of interaction coefficient estimator \(\hat{b}_p\) increases as well. In (Eq. 3.54) where \(\hat{D}\) is defined, the difference \(Y_i - \hat{\gamma}^T \hat{X}_i\) increases. As a result, the estimated variance of \(\hat{b}_j\) increases as \(\hat{D}\) has greater elementwise values. The mean absolute bias of the desparsified estimate \(\hat{b}\) is much smaller than that of lasso estimator, which shows the bias-correction effect of \(\hat{b}\). More simulation results for \(p = 100\) and 200 are presented in Appendix.
**Table 3.1** Coverage of confidence interval in (Eq. 3.55) based upon limiting distribution of desparsified estimator and mean absolute bias of desparsified estimator in (Eq. 3.5) when $p = 300$ and $s = 5$. True variable intensity is $1$ or $1.5$. Case 1 is simulated from linear baseline model. Case 2 is simulated from quadratic baseline model. Case 3 is simulated from trigonometric baseline model. Standard errors are in parenthesis.

<table>
<thead>
<tr>
<th></th>
<th>Intensity=$1$</th>
<th></th>
<th>Intensity=$1.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 3</td>
</tr>
<tr>
<td>Coverage for noise variables</td>
<td>0.95(0.01)</td>
<td>0.95(0.01)</td>
<td>0.95(0.03)</td>
</tr>
<tr>
<td>Coverage for relevant variables</td>
<td>0.88(0.13)</td>
<td>0.89(0.15)</td>
<td>0.85(0.18)</td>
</tr>
<tr>
<td>Length of CI for noise variables</td>
<td>0.65(0.04)</td>
<td>1.50(0.17)</td>
<td>1.21(0.91)</td>
</tr>
<tr>
<td>Length of CI for relevant variables</td>
<td>0.65(0.04)</td>
<td>1.49(0.16)</td>
<td>1.20(0.92)</td>
</tr>
<tr>
<td>MAB($\hat{b}$)</td>
<td>0.17(0.06)</td>
<td>0.40(0.14)</td>
<td>0.30(0.10)</td>
</tr>
<tr>
<td>MAB($\hat{\beta}$)</td>
<td>0.23(0.06)</td>
<td>0.50(0.14)</td>
<td>0.39(0.11)</td>
</tr>
</tbody>
</table>

**Table 3.2** Coverage of confidence interval in (Eq. 3.55) based upon limiting distribution of desparsified estimator and mean absolute bias of desparsified estimator in (Eq. 3.5) when $p = 300$ and $s = 10$. True variable intensity is $1$ or $1.5$. Case 1 is simulated from linear baseline model. Case 2 is simulated from quadratic baseline model. Case 3 is simulated from trigonometric baseline model. Standard errors are in parenthesis.

<table>
<thead>
<tr>
<th></th>
<th>Intensity=$1$</th>
<th></th>
<th>Intensity=$1.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 3</td>
</tr>
<tr>
<td>Coverage for noise variables</td>
<td>0.95(0.03)</td>
<td>0.94(0.04)</td>
<td>0.95(0.03)</td>
</tr>
<tr>
<td>Coverage for relevant variables</td>
<td>0.86(0.14)</td>
<td>0.86(0.16)</td>
<td>0.84(0.14)</td>
</tr>
<tr>
<td>Length of CI for noise variables</td>
<td>0.69(0.05)</td>
<td>1.58(0.18)</td>
<td>1.19(0.13)</td>
</tr>
<tr>
<td>Length of CI for relevant variables</td>
<td>0.69(0.05)</td>
<td>1.58(0.18)</td>
<td>1.18(0.13)</td>
</tr>
<tr>
<td>MAB($\hat{b}$)</td>
<td>0.19(0.04)</td>
<td>0.42(0.10)</td>
<td>0.33(0.08)</td>
</tr>
<tr>
<td>MAB($\hat{\beta}$)</td>
<td>0.24(0.06)</td>
<td>0.51(0.10)</td>
<td>0.40(0.09)</td>
</tr>
</tbody>
</table>
Table 3.3 Coverage of confidence interval in (Eq. 3.55) based upon limiting distribution of desparsified estimator and mean absolute bias of desparsified estimator in (Eq. 3.5) when dimension $p = 100$ and number of true variables $s = 5$. True variable intensity is 1 or 1.5. Case 1 is simulated from linear baseline model. Case 2 is simulated from quadratic baseline model. Case 3 is simulated from trigonometric baseline model. Standard errors are in parenthesis.

<table>
<thead>
<tr>
<th></th>
<th>Intensity=1</th>
<th></th>
<th>Intensity=1.5</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 3</td>
<td>Case 1</td>
</tr>
<tr>
<td>Coverage for noise</td>
<td>0.92(0.16)</td>
<td>0.86(0.28)</td>
<td>0.88(0.23)</td>
<td>0.87(0.26)</td>
</tr>
<tr>
<td>variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage for relevant</td>
<td>0.86(0.22)</td>
<td>0.79(0.30)</td>
<td>0.79(0.25)</td>
<td>0.80(0.29)</td>
</tr>
<tr>
<td>variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length of CI for</td>
<td>0.67(0.19)</td>
<td>1.88(3.78)</td>
<td>1.36(2.16)</td>
<td>0.95(1.69)</td>
</tr>
<tr>
<td>noise variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length of CI for</td>
<td>0.67(0.19)</td>
<td>1.87(3.77)</td>
<td>1.35(2.15)</td>
<td>0.94(1.68)</td>
</tr>
<tr>
<td>relevant variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAB($\hat{b}$)</td>
<td>0.17(0.07)</td>
<td>0.37(0.12)</td>
<td>0.30(0.11)</td>
<td>0.35(1.71)</td>
</tr>
<tr>
<td>MAB($\hat{\beta}$)</td>
<td>0.23(0.10)</td>
<td>0.50(0.16)</td>
<td>0.40(0.14)</td>
<td>0.43(1.71)</td>
</tr>
</tbody>
</table>

Table 3.4 Coverage of confidence interval in (Eq. 3.55) based upon limiting distribution of desparsified estimator and mean absolute bias of desparsified estimator in (Eq. 3.5) when dimension $p = 100$ and number of true variables $s = 10$. True variable intensity is 1 or 1.5. Case 1 is simulated from linear baseline model. Case 2 is simulated from quadratic baseline model. Case 3 is simulated from trigonometric baseline model. Standard errors are in parenthesis.

<table>
<thead>
<tr>
<th></th>
<th>Intensity=1</th>
<th></th>
<th>Intensity=1.5</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 3</td>
<td>Case 1</td>
</tr>
<tr>
<td>Coverage for noise</td>
<td>0.85(0.29)</td>
<td>0.86(0.27)</td>
<td>0.90(0.22)</td>
<td>0.87(0.26)</td>
</tr>
<tr>
<td>variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage for relevant</td>
<td>0.77(0.29)</td>
<td>0.79(0.27)</td>
<td>0.79(0.23)</td>
<td>0.78(0.27)</td>
</tr>
<tr>
<td>variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length of CI for</td>
<td>2.33(10.71)</td>
<td>2.54(6.97)</td>
<td>2.16(9.65)</td>
<td>0.93(2.44)</td>
</tr>
<tr>
<td>noise variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length of CI for</td>
<td>2.32(10.67)</td>
<td>2.51(6.74)</td>
<td>2.18(9.88)</td>
<td>0.93(2.49)</td>
</tr>
<tr>
<td>relevant variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAB($\hat{b}$)</td>
<td>0.19(0.08)</td>
<td>1.45(10.02)</td>
<td>0.32(0.09)</td>
<td>0.18(0.05)</td>
</tr>
<tr>
<td>MAB($\hat{\beta}$)</td>
<td>0.24(0.11)</td>
<td>1.52(10.01)</td>
<td>0.39(0.11)</td>
<td>0.22(0.06)</td>
</tr>
</tbody>
</table>
Table 3.5 Coverage of confidence interval in (Eq. 3.55) based upon limiting distribution of desparsified estimator and mean absolute bias of desparsified estimator in (Eq. 3.5) when dimension $p = 200$ and number of true variables $s = 5$. True variable intensity is 1 or 1.5. Case 1 is simulated from linear baseline model. Case 2 is simulated from quadratic baseline model. Case 3 is simulated from trigonometric baseline model. Standard errors are in parenthesis.

<table>
<thead>
<tr>
<th></th>
<th>Intensity=1</th>
<th>Intensity=1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case 1</td>
<td>Case 2</td>
</tr>
<tr>
<td>Coverage for noise variables</td>
<td>0.94(0.08)</td>
<td>0.94(0.06)</td>
</tr>
<tr>
<td>Coverage for relevant variables</td>
<td>0.87(0.19)</td>
<td>0.84(0.19)</td>
</tr>
<tr>
<td>Length of CI for noise variables</td>
<td>0.66(0.04)</td>
<td>1.56(0.84)</td>
</tr>
<tr>
<td>Length of CI for relevant variables</td>
<td>0.65(0.04)</td>
<td>1.55(0.83)</td>
</tr>
<tr>
<td>MAB($\hat{b}$)</td>
<td>0.16(0.06)</td>
<td>0.62(2.17)</td>
</tr>
<tr>
<td>MAB($\hat{\beta}$)</td>
<td>0.23(0.07)</td>
<td>0.73(2.13)</td>
</tr>
</tbody>
</table>

Table 3.6 Coverage of confidence interval in (Eq. 3.55) based upon limiting distribution of desparsified estimator and mean absolute bias of desparsified estimator in (Eq. 3.5) when dimension $p = 200$ and number of true variables $s = 10$. True variable intensity is 1 or 1.5. Case 1 is simulated from linear baseline model. Case 2 is simulated from quadratic baseline model. Case 3 is simulated from trigonometric baseline model. Standard errors are in parenthesis.

<table>
<thead>
<tr>
<th></th>
<th>Intensity=1</th>
<th>Intensity=1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case 1</td>
<td>Case 2</td>
</tr>
<tr>
<td>Coverage for noise variables</td>
<td>0.95(0.05)</td>
<td>0.94(0.06)</td>
</tr>
<tr>
<td>Coverage for relevant variables</td>
<td>0.86(0.15)</td>
<td>0.85(0.16)</td>
</tr>
<tr>
<td>Length of CI for noise variables</td>
<td>0.69(0.05)</td>
<td>1.54(0.19)</td>
</tr>
<tr>
<td>Length of CI for relevant variables</td>
<td>0.69(0.05)</td>
<td>1.54(0.18)</td>
</tr>
<tr>
<td>MAB($\hat{b}$)</td>
<td>0.18(0.06)</td>
<td>0.42(0.12)</td>
</tr>
<tr>
<td>MAB($\hat{\beta}$)</td>
<td>0.24(0.06)</td>
<td>0.51(0.10)</td>
</tr>
</tbody>
</table>
3.3.3 Additional simulation table

3.4 Application in estimating treatment decision rule in STAR*D study

The major depressive disorder (MDD) is a common recurrent and chronic episodic disorder. We consider the dataset from multi-site, randomized, multi-step STAR*D (Sequenced Treatment Alternatives to Relieve Depression) study, which includes the patients with MDD [76].

In level 1, the patients receive citalopram (CIT) which is a selective serotonin reuptake inhibitor antidepressant. Patients who have had unsatisfactory outcome in level 1 are included in level 2. We are interested in comparing the clinical outcome for the two treatments switch options in level 2: sertraline (SER) and bupropion (BUP). Patients which receive treatment at level 2 but have not showed sufficient improvement will be randomized at level 2A.

Our data contains patients who receive SER or BUP in level 2 of STAR*D study. In our data, 319 patients and 308 covariate variables are present. Treatment SER or BUP is assigned to each patient. In our data, 48% receives treatment BUP and 52% receives treatment SER. The outcome is measured as the Quick Inventory of Depressive Symptomatology-Self-report QIDS-SR16. We transform the outcome to be the negative of the outcome in our analysis. The clinical outcome is recorded for each patient. In order to compare the effectiveness of SER and BUP, we examine the covariates interaction with the binary treatment option and determine whether SER or BUP is better for each individual patient.
3.4.1 P value from desparsifying lasso approach

We assume the model in (3.2) for our data. We have developed the desparsifying lasso procedure for arbitrary baseline model. We have also derived the confidence intervals and hypothesis tests of the interaction coefficients. We apply desparsifying lasso procedure in penalized regression for precision medicine to the STAR*D data. We obtain the p values of the interaction coefficients based upon desparsifying lasso approach.

Table 3.7 is a table of p values for the interactions which have the smallest p values. In Table 3.7, variables are the treatment interactions with the corresponding covariates. The p values help us find the most important interactions that are significant for deciding whether SER or BUP is better for each individual patient.

From the p value in Table 3.7, we select the variables with the p value less than 0.05: "qccur_r_rate", "URNONE", "NVTRM", "IMPWR", "hWL", "GLT2W", "DSMTD", "IMSPY", "hMNIN", "EARNG", "URPN", "PETLK", "NVCRD" and "EMSTU". We use the selected model to estimate the optimal decision rule to assign SER or BUP to each individual patient.

3.4.2 Estimated optimal decision rule

In our data, BUP is coded as 0 and SER is coded as 1 in the treatment A. In the statistical model 3.2, the contrast model $\tilde{X} b_\beta$ is the individual expected clinical outcome for SER minus that for BUP. If the estimated contrast function $\hat{\tilde{X}} b_\beta$ is positive, it implies that the SER is significantly better for the individual patient. Otherwise, BUP is significantly better for the individual patient.

First the coefficients in the optimal decision rule are based upon the desparsified lasso
Table 3.7 Table of p values for significant interaction between treatment and patients’ covariates. P values are from the limiting distribution of desparsified lasso estimator. Variables are treatment interactions with the variables. The meaning of these variables are as follows. Qccur_r_rate: QIDS-C score changing rates. URNONE: no symptoms in patients’ urination category. NVTRM: Tremors. IMPWR: indicating whether patients thought they have special powers. hWL: hRS Weight loss. DSMTD: Recurrent thoughts of death, recurrent suicidal ideation, or suicide attempt. hMNIN: hRS Middle insomnia. EMSTU: Did you worry a lot that you might do something to make people think that you were stupid or foolish? [18] P values are computed using limiting distribution of desparsified lasso estimator.

<table>
<thead>
<tr>
<th>Index</th>
<th>Variables</th>
<th>p values</th>
<th>b</th>
<th>se of b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>qccur_r_rate</td>
<td>0.0056</td>
<td>-1.84</td>
<td>0.66</td>
</tr>
<tr>
<td>2</td>
<td>URNONE</td>
<td>0.0079</td>
<td>2.35</td>
<td>0.88</td>
</tr>
<tr>
<td>3</td>
<td>NVTRM</td>
<td>0.0097</td>
<td>-1.45</td>
<td>0.56</td>
</tr>
<tr>
<td>4</td>
<td>IMPWR</td>
<td>0.010</td>
<td>-1.47</td>
<td>0.57</td>
</tr>
<tr>
<td>5</td>
<td>hWL</td>
<td>0.011</td>
<td>-1.58</td>
<td>0.62</td>
</tr>
<tr>
<td>6</td>
<td>GLT2W</td>
<td>0.013</td>
<td>-1.52</td>
<td>0.61</td>
</tr>
<tr>
<td>7</td>
<td>DSMTD</td>
<td>0.017</td>
<td>-1.50</td>
<td>0.63</td>
</tr>
<tr>
<td>8</td>
<td>IMSPY</td>
<td>0.019</td>
<td>-1.36</td>
<td>0.58</td>
</tr>
<tr>
<td>9</td>
<td>hMNIN</td>
<td>0.020</td>
<td>1.40</td>
<td>0.60</td>
</tr>
<tr>
<td>10</td>
<td>EARNG</td>
<td>0.028</td>
<td>-1.43</td>
<td>0.65</td>
</tr>
<tr>
<td>11</td>
<td>URPN</td>
<td>0.029</td>
<td>-1.23</td>
<td>0.56</td>
</tr>
<tr>
<td>12</td>
<td>PETLK</td>
<td>0.039</td>
<td>-1.18</td>
<td>0.57</td>
</tr>
<tr>
<td>13</td>
<td>NVCRD</td>
<td>0.041</td>
<td>1.17</td>
<td>0.57</td>
</tr>
<tr>
<td>14</td>
<td>EMSTU</td>
<td>0.046</td>
<td>-1.39</td>
<td>0.69</td>
</tr>
</tbody>
</table>
Figure 3.1 The 95% confidence interval of selected interaction in Table 3.7. The horizontal axis is the index of variables in Table 3.7. The vertical axis is the desparsified lasso estimate.
estimate from the regression of clinical outcome versus the design matrix. Denote

\[ \hat{b}_\beta = (-1.84, 2.35, -1.45, -1.47, -1.58, -1.52, -1.50, -1.36, 1.40, -1.43, -1.23, -1.18, 1.17, -1.39). \]

Denote the design matrix containing only columns of selected variables in Table 3.7 to be \( \tilde{X} \). The estimated optimal decision rule is that SER is optimal for the individual if

\[ \tilde{X} \hat{b}_\beta > 0. \]  \hspace{1cm} (3.60)

BUP is optimal for the individual if

\[ \tilde{X} \hat{b}_\beta < 0. \]  \hspace{1cm} (3.61)

We use the estimated optimal decision rule to make treatment assignment for individual patients in level 2 of STAR*D study.

3.4.3 Illustration of advantage in estimated optimal decision rule using bootstrap or cross validation

In STAR*D study, treatment SER or BUP is randomly assigned to each individual patient. We compare estimated optimal treatment assignment scheme to the strategy of assigning only SER or the strategy of assigning only BUP. To evaluate the performance of estimated optimal personalized treatment decision rule in (Eq. 3.60), we take bootstrap samples of size \( n = 319 \) from our data. Bootstrap samples are random samples taken with replacement. We intend to obtain the sampling distribution of the clinical outcomes from the treatment
strategies we compare by using bootstrap. For each bootstrap sample, we compute the average outcomes of estimated optimal personalized treatment strategy, assigning only SER and assigning only BUP. We take 1000 bootstrap samples from our data.

We use the inverse probability weighted estimator (IPWE) described in [97] to estimate the value function of the treatment regime. The IPWE is defined as

\[ \hat{V}(d(\tilde{X})) = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i 1(A_i = d(\tilde{X}))}{\hat{\pi}(d(\tilde{X}))}, \] (3.62)

where \( \hat{V}(\cdot) \) is the estimated value function, and \( \hat{\pi} \) is the estimated propensity score of the strategy to be evaluated. The boxplot of the average IPWE from the 1000 bootstrap samples is in figure (Fig. 3.2). We can see that the optimal decision is better than assigning only SER or assigning only BUP.

By using bootstrap method, we evaluate the performance of the estimated optimal treatment regime from the complete dataset. By using cross validation method, we evaluate the performance of the selected variables from our procedure applied to the complete dataset.

We also apply cross validation to evaluate the variable selection from our procedure. We randomly divide the data in half into the training dataset and the testing dataset. On the training dataset, we use the 14 selected variables to compute the least squares estimator of interaction coefficients. Then we formulate the estimated optimal treatment regime using the selected variables and least squares estimates of interaction coefficients. On the corresponding testing dataset, we compute the average IPWE of our estimated optimal treatment regime, the strategy of assigning only SER and the strategy of assigning only BUP. The steps described above are repeated 1000 times to make the boxplot in (Fig. 3.3). We can
see that our estimated optimal treatment regime is better than the strategies of assigning only SER or only BUP.

**Figure 3.2** Boxplot of average IPWE from bootstrap samples. We compare the estimated optimal personalized treatment strategy with the strategy of assigning only SER and the strategy of assigning only BUP. Horizontal axis is average clinical outcome. In vertical axis, ‘optimal’ is estimated optimal personalized treatment regime. ‘all.BUP’ is the strategy of assigning only BUP. ‘all.SER’ is the strategy of assigning only SER.
Figure 3.3 Boxplot of average IPWE from cross validation testing samples. We compare the estimated optimal personalized treatment strategy with the strategy of assigning only SER and the strategy of assigning only BUP. Horizontal axis is average clinical outcome. In vertical axis, 'optimal' is estimated optimal personalized treatment regime. 'all.BUP' is the strategy of assigning only BUP. 'all.SER' is the strategy of assigning only SER.
BIBLIOGRAPHY


