Estimation of Total Effect of Selected Variables on an Outcome in Presence of Numerous Weak Effects

BY

YARU SHI
B.S., Beijing Institute of Technology, Beijing, China, 2011
M.S., Georgetown University, Washington, D.C., 2012

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Public Health Sciences
in the Graduate College of the
University of Illinois at Chicago, 2018

Chicago, Illinois

Defense Committee:
Dr. Hua Yun Chen, Chair and Advisor
Dr. Maria Argos
Dr. Sanjib Basu
Dr. Jie Yang, Department of Mathematics, Statistics, and Computer Science
Dr. Chunyu Liu, SUNY Upstate Medical University

ACKNOWLEDGMENT

I would like to take this opportunity to thank all the people who have been helping and supporting me through this most important journey in my life.

Firstly, I would like to thank my thesis advisor, Dr Hua Yun Chen, for the continuous encouragement and help of my Ph.D study and research. His immense knowledge, patience, and insightful advice have helped me tremendously in all the time of research and writing of this thesis. I am especially grateful for his rigorous attitudes in research, which helps me to understand the nature of scientific research. I could not have thought of a better choice of the mentor for my Ph.D.

I would also like to express my sincere gratitude to my dissertation committee, Dr. Maria Argos, Dr. Sanjib Basu, Dr. Chunyu Liu, and Dr. Jie Yang, for their kindness and flexibility. Their critical comments and valuable suggestions have helped me widen my thesis from various perspective.

Last but not least, I am very grateful to my parents and boyfriend, for their unconditional love and patience. I would also like to thank my friends, Zoey and Liyuan, who have provide me with moral and emotional support through this process.

TABLE OF CONTENTS

CHAPTER		PAGE
1	**INTRODUCTION**	1
1.1	The Problem of Identifying Causal Genetic Variants of a Trait	1
1.2	The Challenge of Detecting Numerous Weak Signals	6
1.3	The Proposed Work	11
2	**THE WEAK SIGNAL PROBLEM AND POSSIBLE SOLUTIONS**	14
2.1	Problem Formulation	14
2.2	Variable Selection	17
2.3	Estimation of Total Variation Explained by Selected Covariates	18
2.3.1	Methods by assembly of individual estimators	19
2.3.2	Direct estimating the explained variation as a parameter	24
2.4	Simulation Designs and Results	27
2.4.1	Experiment 1	28
2.4.2	Experiment 2	33
2.4.3	Experiment 3	36
2.4.4	Experiment 4	39
2.4.5	Experiment 5	42
2.4.6	Experiment 6	45
2.4.7	Experiment 7	48
2.4.8	Experiment 8	51
2.4.9	Experiment 9	54
2.4.10	Experiment 10	57
2.5	Summary and Comments	60
3	**ESTIMATING TOTAL EFFECT OF SELECTED VARIABLES: THE CASE OF INDEPENDENT COVARIATES**	63
3.1	The LMM with Sample Split	63
3.2	The Proposed Subsampling Approach	64
3.3	Simulation Study and Method Comparison	72
3.3.1	Experiment 1	75
3.3.2	Experiment 2	79
3.3.3	Experiment 3	83
3.3.4	Experiment 4	87
3.4	Application to Real Data	91
4	**ESTIMATING TOTAL EFFECT OF SELECTED VARIABLES: THE CASE OF CORRELATED COVARIATES**	93

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

4.1 Modified Subsampling Approach for Correlated Covariates . . 93
4.2 Simulation Designs and Data Generation 97
4.3 Simulation Results 100
4.3.1 Experiment 1 101
4.3.2 Experiment 2 105
4.3.3 Experiment 3 109
4.3.4 Experiment 4 113
4.4 Application to Real Data 118

5 CONCLUSION AND DISCUSSION 121

APPENDICES 124
Appendix A 125
Appendix B 127

CITED LITERATURE 128

VITA 133

LIST OF TABLES

TABLE		PAGE
I	RESULTS IN EXPERIMENT 1 FOR CURRENT METHODS . .	31
II	RESULTS IN EXPERIMENT 2 FOR CURRENT METHODS . .	34
III	RESULTS IN EXPERIMENT 3 FOR CURRENT METHODS . .	37
IV	RESULTS IN EXPERIMENT 4 FOR CURRENT METHODS . .	40
V	RESULTS IN EXPERIMENT 5 FOR CURRENT METHODS . .	43
VI	RESULTS IN EXPERIMENT 6 FOR CURRENT METHODS . .	46
VII	RESULTS IN EXPERIMENT 7 FOR CURRENT METHODS . .	49
VIII	RESULTS IN EXPERIMENT 8 FOR CURRENT METHODS . .	52
IX	RESULTS IN EXPERIMENT 9 FOR CURRENT METHODS . .	55
X	RESULTS IN EXPERIMENT 10 FOR CURRENT METHODS .	58
XI	MSE IN EXPERIMENT 1 FOR SUBSAMPLING APPROACH	77
XII	90% CI BY SUBSAMPLING APPROACH IN EXPERIMENT 1	78
XIII	MSE IN EXPERIMENT 2 FOR SUBSAMPLING APPROACH	81
XIV	90% CI BY SUBSAMPLING APPROACH IN EXPERIMENT 2	82
XV	MSE IN EXPERIMENT 3 FOR SUBSAMPLING APPROACH	85
XVI	90% CI BY SUBSAMPLING APPROACH IN EXPERIMENT 3	86
XVII	MSE IN EXPERIMENT 4 FOR SUBSAMPLING APPROACH	89
XVIII	90% CI BY SUBSAMPLING APPROACH IN EXPERIMENT 4	90

LIST OF TABLES (Continued)

TABLE **PAGE**

XIX	RESULTS OF THE ANALYSIS OF BRAIN TISSUE eQTL DATA	92
XX	MSE IN EXPERIMENT 1 BY LASSO	102
XXI	MSE IN EXPERIMENT 1 BY INDIVIDUAL TESTS	104
XXII	EFFICIENCY COMPARISON IN EXPERIMENT 1	105
XXIII	MSE IN EXPERIMENT 2 BY LASSO	106
XXIV	MSE IN EXPERIMENT 2 BY INDIVIDUAL TESTS	108
XXV	EFFICIENCY COMPARISON IN EXPERIMENT 2	109
XXVI	MSE IN EXPERIMENT 3 BY LASSO	110
XXVII	MSE IN EXPERIMENT 3 BY INDIVIDUAL TESTS	112
XXVIII	EFFICIENCY COMPARISON IN EXPERIMENT 3	113
XXIX	MSE IN EXPERIMENT 4 BY LASSO	114
XXX	MSE IN EXPERIMENT 4 BY INDIVIDUAL TESTS	116
XXXI	EFFICIENCY COMPARISON IN EXPERIMENT 4	117
XXXII	90% CI FOR CORRELATED CASES	119
XXXIII	SUMMARY OF ANALYSES RESULTS IN eQTL DATA	120
XXXIV	SUMMARY STATISTICS FOR LDPE	125
XXXV	MSE OF FDE WITH VARIOUS SETTINGS	127

LIST OF FIGURES

FIGURE		PAGE
1	Box plot with varied sparsity	16
2	Mean estimation with varied α for part approaches in Experiment 1	30
3	Box plot with varied α for part approaches in Experiment 1	32
4	Mean estimation with varied α for part approaches in Experiment 2	33
5	Box plot with varied α for part approaches in Experiment 2	35
6	Mean estimation with varied α for part approaches in Experiment 3	36
7	Box plot with varied α for part approaches in Experiment 3	38
8	Mean estimation with varied α for part approaches in Experiment 4	39
9	Box plot with varied α for part approaches in Experiment 4	41
10	Mean estimation with varied α for part approaches in Experiment 5	42
11	Box plot with varied α for part approaches in Experiment 5	44
12	Mean estimation with varied α for part approaches in Experiment 6	45
13	Box plot with varied α for part approaches in Experiment 6	47
14	Mean estimation with varied α for part approaches in Experiment 7	48
15	Box plot with varied α for part approaches in Experiment 7	50
16	Mean estimation with varied α for part approaches in Experiment 8	51
17	Box plot with varied α for part approaches in Experiment 8	53
18	Mean estimation with varied α for part approaches in Experiment 9	54

LIST OF FIGURES (Continued)

FIGURE		PAGE
19	Box plot with varied α for part approaches in Experiment 9	56
20	Mean estimation with varied α for part approaches in Experiment 10	57
21	Box plot with varied α for part approaches in Experiment 10	59
22	Demonstration of Yang et al.'s approach under variable selection . .	63
23	Box plot of subsampling approach with varied α in Experiment 1 . .	76
24	Box plot of subsampling approach with varied α in Experiment 2 . .	80
25	Box plot of subsampling approach with varied α in Experiment 3 . .	84
26	Box plot of subsampling approach with varied α in Experiment 4 . .	88
27	Box plot with varied λ in Experiment 1	101
28	Box plot with varied α in Experiment 1	103
29	Box plot with varied λ in Experiment 2	107
30	Box plot with varied α in Experiment 2	107
31	Box plot with varied λ in Experiment 3	111
32	Box plot with varied α in Experiment 3	111
33	Box plot with varied λ in Experiment 4	115
34	Box plot with varied α in Experiment 4	115

LIST OF ABBREVIATIONS

GWAS	Genome Wide Association Study
SNP	Single-nucleotide polymorphism
LMM	Linear mixed model
LD	Linkage disequilibrium
EM	Expectation maximization
REML	Restricted maximum likelihood
LDPE	Low Dimensional Projection Estimator
FDE	Functional De-biased Estimator
FDR	False Discovery Rate
FWER	Family-wise error rate
Lasso	Least Absolute Shrinkage and Selection Operator
SCAD	Smoothly Clipped Absolute Deviations
ANOVA	Analysis of Variance
CV	Cross Validation
MSE	Mean Squared Error

SUMMARY

Estimation of the total effect of a set of selected variables in high-dimensional linear model under sparsity assumption is complex due to the selection bias. This task is even more challenging when the sparsity assumption is violated and individual variable effects are weak, which is common in genomic studies. Without variable selection, Yang et al have proposed an effective approach to estimating the total effects of single-nucleotide polymorphisms (SNPs) on a quantitative trait when effects of SNPs are numerous and weak. In the thesis, we extended Yang et al's approach to estimating the total effect of a set of selected variables in a linear model. The extension allows us to effectively reduce the scope of search for the causal SNPs for a quantitative trait in presence of numerous weak effects. We also modify our proposed approach to make it suitable for correlated SNPs. We perform extensive simulation studies to demonstrate the effectiveness of the proposed approach in comparison to alternative approaches to this problem. The method is applied to detecting the expression quantitative trait locus (eQTL) in gene-expression study of human brain tissues.

CHAPTER 1

INTRODUCTION

1.1 The Problem of Identifying Causal Genetic Variants of a Trait

The rapid technology advancement has enabled researchers in biological sciences and other fields to collect massive amount of data both in volume and in dimensionality in a relatively short time period. For example, in a genome-wide association study (GWAS), genetic variants in single-nucleotide polymorphism (SNP) in hundreds of thousands of loci are collected in each subject. One major goal of the data collection is to identify causal genetic variants such as SNPs that are responsible for complex disorders. Traditional statistical methods may not be appropriate for handling such data because the data have more variables than observations. To make use of the massive high-dimensional data to answer important questions, it is critical to develop statistical methods for the analyses of such data. In the past two decades, a number of novel statistical approaches have been proposed for variable selection, outcome prediction, and statistical inferences for high-dimensional data (Tibshirani (1996), Sun and Zhang (2012), Yang et al. (2010), Guo et al. (2016), Zhang and Zhang (2014)).

One simple and frequently used method for identifying important SNPs that are associated with a complex disorder is to perform statistical tests of the association of the complex disorder with the SNPs one-at-a-time. A cutoff value that adjusts for the multiple testing by controlling the family-wise error rate (FWER) (Dunn, 1959) or false discovery rate (FDR) (Benjamini and

Hochberg, 1995) may be used for screening the SNPs. FWER is the probability of incorrectly rejecting at least one true null hypothesis among all the tests. A classical approach to control FWER is Bonferroni adjustment (Dunn, 1959) which controls the FWER strictly by increasing the significant level in individual tests according to the total number of hypotheses tested. No additional assumptions are required for this approach to work. In other words, the overall error is always under control as long as the significance of individual tests is adequately adjusted irrespective to the number of tests conducted or whether the tests are independent. Using FWER adjustment for multiple testing is too conservative. When the data include a large number of SNPs, Bonferroni procedure sets a very stringent cutoff value which makes it difficult for a causal SNP to be detected. A less conservative approach is to control the false discovery rate (FDR), which measures the expected proportion of the falsely rejected null hypotheses among all the rejections. The original FDR method, called BH-FDR, is introduced by Simes (1986) and further developed by Benjamini and Hochberg (1995). The basic procedure is to rank the tests according to the p-value and compare each individual p-value with certain critical values. Compared with Bonferroni procedure, the FDR gains more power when more non-nulls are present. These methods can detect a SNP with a big effect relative to the noise level. For SNPs with weak effects, a relaxed criterion needs to be applied to detecting them. However, the false negative rate is increased with a relaxed criterion.

Many improvements on the BH-FDR method have been proposed. Benjamini and Hochberg (2000) improved the BH-FDR by incorporating the estimated number of true null hypotheses. The new approach is called adaptive BH-FDR. Besides the adjustment procedures in indepen-

dent or weak positive dependent tests, a procedure called BY-FDR is also introduced to account for the FDR in correlated tests (Benjamini and Yekutieli, 2001). Storey (2002) extended the FDR method to control for pFDR, the expected false discover rate given the positive discovery event occurs. Generally, pFDR is more powerful compared with BH-FDR. Benjamini et al. (2006) introduced another method by considering the distribution of p-value, and the testing power is further improved through the process. Efron (2008) introduced "fdr" to estimate the local FDR for each case. Compared with traditional FDR, fdr method is more readily interpretable as it is the empirical Bayesian posterior probability of the false positive event conditional on the observed test events. In general, the FDR methods are more powerful than the Bonferroni procedure, they have the best performance when the tests are independent. When the number of tests are large and most signals are weak, FDR method can be computationally costly and it may increase the false positive rate unless the sample size is sufficiently large. In addition, SNPs detected by the individual tests are less likely to be causal SNPs than those detected by model-based approaches because fewer confounding effect adjustments are included in individual tests.

A more sophisticated approach to the causal SNP detection is the model-based methods that model the effects of SNPs on the disorder and apply one of the variable selection approaches to the SNPs. One popular approach is to use the l^1 penalty of the regression parameters to automatically select the variables. With l^1 penalty, the small coefficients are shrunken to zero while large coefficients, though also shrunken, remain nonzero. Therefore, the Lasso approach (Tibshirani, 1996) can select variables by adjusting the penalty levels. Many extension and im-

provement of the Lasso approach have been proposed for dealing with high-dimensional data. Zou and Hastie (2005) added an l^2 penalty to the Lasso estimation function, termed "elastic net", to ensure unique solutions when covariates are highly correlated. Yuan and Lin (2006) proposed "group Lasso" to allow for a pre-assigned groups of covariates to be selected into the model together. In addition to assigning penalty on the sum of the squared coefficients, Tibshirani et al. (2005) introduced the Fused Lasso to extend the penalty to the differences between the coefficients. As a result, in the Fused Lasso, not only the coefficients are sparse, their differences are also sparse. Usually, the cross-validation approach is used to find the optimal penalty level, which is associated with the true noise level. To avoid such search, Belloni et al. (2011) proposed a method called square-root Lasso, using the square-root of the error variance estimate as the estimation function. It has been shown that the square-root Lasso can give an equivalent result as Lasso without using information of noise level. Furthermore, the computation is more efficient by formulating the solution as a conic programming. Stadler et al. (2010) maximized a joint log-likelihood of coefficients and noise level with l^1 penalty. Although the estimation function is non-convex, which means no global optimum is guaranteed, this approach still gives a better solution compared with the non-penalized maximization likelihood. Sun and Zhang (2012) improved Stadler et al.'s approach by using an iterative strategy. Their approach, termed the scaled Lasso, guarantees the convergence through transforming the function into a joint convex loss (Antoniadis, 2010). In addition, the noise level estimator in the scaled Lasso is shown to be consistent and asymptotic normal under certain regularity conditions. Sun and Zhang (2012) also suggested to re-estimate the coefficients and noise level

using the ordinary least square after the Lasso selection, which can further reduce the bias of estimators. Since the optimal model selected by the scaled Lasso usually have a smaller number of coefficients than the sample size, some statistical inference can be conducted. The statistical inference based on the scaled Lasso limits to the non-zero coefficients, an improved approach is proposed to expand the inferences to all coefficients (Zhang and Zhang (2014), Geer et al. (2014)). This method is called "Low Dimensional Projection Estimator (LDPE)" in Zhang and Zhang (2014). By approximating the inverse of covariance-variance matrix for covariates using Lasso approach, LDPE can estimate the single coefficient with much smaller bias. This method also gives an estimation of standard errors of the estimated coefficient, which makes statistical inference possible for high-dimensional data. Simulation studies show that the confidence intervals constructed by the proposed methods have a good overall coverage for the true coefficients.

In addition to the Lasso approach and its extensions, l^p penalty has also been used in high-dimensional data. When $p = 2$, it is ridge regression and it shrinks the estimators proportionally without reducing the model size in a high dimensional model. Another extension is the bridge regression which does not take the pth root of the penalty in comparison to the l^p penalty. When $p = 1$ or $p = 2$, bridge regression reduces to the Lasso regression or Ridge regression, respectively. When $p > 2$, bridge regression tends to shrink small estimators with smaller rate and large estimators with large rate (Fu, 1998). Therefore, for $p > 2$ with large signals, bridge regression cannot capture the large signals as well as Lasso. Fan and Li (2001) proposed a quadratic spline function with knots at penalty level to further reduce the bias of the estimators,

and they call it "smoothly clipped absolute deviation (SCAD)" penalty. By imposing different penalties on coefficients in different scales, this approach sets small coefficients to zero, shrinks not-so-small coefficients towards zero, and does not shrink large coefficients. As a result, SCAD produces a sparse and approximately unbiased coefficients for large coefficients. One of the problems with the SCAD approach is that the penalty function is concave, which makes the computation of the estimator very challenging because fast computational approaches such as the convex programming, cannot be directly applied to computing the estimator. In contrast, Lasso type of approaches can usually be solved by the convex programming approach or even linear programming approach, which is very important for handling high-dimensional data. Adaptive Lasso assigns different weights to different coefficients. It helps correct the bias in Lasso, but requires a set of pre-estimated weights which may not be obtainable in the high-dimensional problem (Efron et al. (2004), Friedman et al. (2007), Friedman et al. (2010), Zuo and Hastie (2005)). The Lasso regression has also been extended beyond the simple linear regression model to the generalized linear models and other regression models. Many variations have been proposed for different situations (Breiman (1995), Geer (2008), Tibshirani (1997), Goeman (2010)).

1.2 The Challenge of Detecting Numerous Weak Signals

The Lasso approaches have good properties in prediction and variable selection. Under some regularity conditions, the Lasso approach can achieve the optimal prediction error rivaling knowing the true model *a priori* (Bickel et al., 2009). The Lasso approach can also achieve selection consistency (Zhao and Yu, 2006) under reasonable conditions. Parameter estimation

and inferences have been well studied by Zhang and Zhang (2014) and Geer et al. (2014). One important requirement for those results to hold is the sparsity assumption on variables in the model. For the GWAS data, this means that, conditional on a small set of SNPs, the disorder is uncorrelated with or independent of the rest of SNPs, and the effects of the set of associated SNPs are reasonably strong. In genomic study of a complex disorder or trait, this requirement often fails to hold. In other words, the set of SNPs that are jointly correlated with the complex disorder or trait is not small and the effects of the SNPs are weak. As a result, Lasso-type of approaches can detect only a small number of SNPs associated with the disorder and, collectively, they can only explain a small proportion of the variation of the complex disorder or trait attributable to the SNPs.

To tackle the problem of having many causal SNPs with weak effects in the genomic studies, Yang et al. (2010) proposed a working linear mixed effects model (LMM) to estimating the total effects of the causal SNPs tagged by the typed SNPs in a GWAS. This method has been successfully applied to many GWASs in estimating the total variation of a complex disorder or trait explained by typed SNPs, a.k.a., the narrow-sense heritability. This quantity ties to the traditional heritability also termed "broad-sense" heritability. As a fundamental concept in genetics, heritability summarizes how much variation of a quantitative trait can be explained by genetic, or genetic-environmental effects (Naomi and Visscher (2008) Griffiths et al. (2000)). A quantitative trait is a manifested numeric characteristic having genetic determinants. In statistical model, gene and environment factors are considered two major factors that impact the quantitative trait. Families studies have been the primary method of estimating heritability.

Since the variance of environmental and genetic factors cannot be measured directly, heritability needs to be estimated through carefully varying the levels of genetic variability among studying subjects. Therefore, members within families are often used for such analyses as they share some common genetic factors and differ in others. Twin studies, relying on the fact that identical twins share the 100% genes while fraternal twins share 50% genes on average and both set share 100% environment in theory, can disentangle the genetic and environmental effects to certain degree so that the heritability can be estimated (Render et al. (1990), Poderman et al. (2015)). More generally, family studies with sample of siblings, parents, and offsprings are more available (Jang (2005), Alexander et al. (2014), Chen et al. (2007)). Analysis of studies usually use regression models or analysis of variances (ANOVA). Although family studies can reveal the risk of passing down a certain diseases within the whole family, the genetic and environmental factors are not easy to be separated within the family. Therefore, the inflated heritability is often reported when effect of shared environment is ignored (Alexander et al., 2014). One solution to remove the effect of shared environment effect is to conduct a corresponding twin study. Studies conducted within family have been able to reveal variances explained by genes in many phenotype (e.g. heights, hair color, etc). Nevertheless, for some common, complex diseases, family studies have been proved to be hard to reproduce the results (Strachan and Read, 2010). On the other hand, the rapid development of GWAS has provided an alternative solution to estimating heritability estimation in complex disorders. Compared with twin and family studies, GWASs are conducted among distant-related subjects, therefore less confounding of environmental and genetic factors. As the model only considers additive genetic effects, the

heritability computed by GWAS often refers as "narrow-sense heritability". Jiang et al. (2016) showed that the working linear mixed model approach yields a consistent estimator of the narrow-sense heritability even if the linear mixed model holds only on a subset of a large pool of SNPs. Their numeric comparison also clearly demonstrated the better performance of Yang et al's approach in comparison with the Lasso-type of approaches in cases with numerous weak signals. Additional simulation comparisons can be found in (Vattikuti et al. (2012), Lee et al. (2012), Listgarten et al. (2012)).

Although the LMM approach can be applied to estimating the total variations explained by the typed SNPs in a GWAS, no information is gained on identifying SNPs that are responsible for the effects in using this approach. It is important in practice to identify the specific SNPs responsible for the effects on the complex disorder or trait, and when this is not possible, to narrow down the scope of search for the set of SNPs responsible for the trait among tens or hundreds of thousands of SNPs included in a GWAS. To accomplish the latter task, we may apply the individual testing strategy with relaxed cutoff values or a Lasso-type variable selection approach with relative small penalty parameter values to select a subset of SNPs. Conditional on selection of covariates, the estimated coefficient is no longer unbiased. The sampling distribution of the estimated coefficient is a truncated normal distribution (Garner (2007), Zhong and Prentice (2008)). The bias depends on the true coefficient value, the corresponding standard error, and the significance level. For variables with weak signals, the chance of their being selected is low and the bias can be large if selected. Such phenomenon is called "winner's curse" (Capen et al., 1971). If the linear mixed model is applied to the selected variables, the estimated

variation explained by the selected covariates from the linear mixed model is in general subject to upward bias. A frequently used approach in practice is to split the data into two halves. One half is used for model selection and the other half is used for estimation. The resulting estimator for the variation explained is subject to large variation. In addition, it yields a biased estimator for the variation explained by the selected SNPs from the whole sample.

Lasso and its extensions can also be used for signal estimation. The scaled Lasso (Sun and Zhang, 2012) can in principle be used to obtain the total variation explained by selected covariates through subtracting the noise from the total variation of the outcome. Since LDPE (Zhang and Zhang, 2014) corrects the bias in the coefficient estimation of Lasso, the estimators may be used to estimate the total variation. However, small individual bias can accumulate if many nonzero effects are involved. Guo et al. (2016) proposed an approach to adjust for the total bias in the estimation of the overall variation explained by a set of covariates. The method is termed "Functional De-biased Estimator (FDE)". By assuming the negligible error term, FDE adjusts the overall bias in the sum of squared scaled Lasso coefficients by estimating the interaction term between estimated coefficient and corresponding error term. The estimation is shown to achieve the optimal rate of convergence. In addition, the simulation studies show it consistently outperforms other Lasso approaches in estimating the total variation.

Certain assumptions are needed for both Yang et al.'s approach and Lasso-related approaches. For Yang et al.'s approach, the coefficients are assumed to be random, while Lasso approaches usually make assumptions on the sparsity of the regression coefficients and the knowledge of noise level. Janson et al. (2017) proposed an approach called EigenPrism to esti-

mate the variation in high-dimensional data without assuming anything on the coefficients, instead, they assume the covariates are normally distributed (See also Dicker (2014)). EigenPrism provides estimated standard errors of the signal estimators under the normality assumption on the covariates. Thereby, the EigenPrism approach allows direct inference on the total variation.

To estimate the variation explained by a set of selected variables, one may first correct the bias in the estimators for the selected variables, and then sum over the squared estimators. Zhong and Prentice (2008) derived the conditional expectation of estimated coefficient, and adjusted the estimator through subtracting the estimated bias. A similar approach proposed by Ghosh et al. (2008) used a conditional maximum likelihood to correct the bias in the estimated coefficients. Such approach, though can reduce the bias of coefficient estimators, it is not unbiased and bias may also be introduced when the estimator is squared to obtain a total variation estimator. This is because bias in the total variation estimator can be large due to the cumulative effect even though bias in individual estimator of the coefficients can be negligible after correction. Chen (2016) proposed an approach to directly adjust the bias in the squared coefficient estimators. Under weak conditions, his method is shown to be consistent. When the selection criterion is less stringent, more covariates can be included in the subset, and correcting individual estimator may not be an efficient approach and more bias can be included when we sum over the corrected estimators.

1.3 The Proposed Work

In this thesis, we conduct a series of studies focusing on the estimation of total variation explained by a subset of selected covariates based on the whole sample. We first examine

performance of current approaches through an extensive simulation study. Next, we propose a subsampling approach to estimating the variation explained by a set of SNPs selected from the individual testing approach based on the whole sample. The subsampling approach was studied as an extension to jackknife approach for variance estimation in Shao (1989) and Shao and Wu (1989), and as an alternative to the bootstrapping approach to constructing confidence interval by Politis and Romano (1994). For variable selection, the subsampling approach was proposed in Meinshausen and Buhlmann (2010). Recently, Bin et al. (2006) performed an empirical study of the subsampling approach to variable selection and found it more appealing than the bootstrapping approach. The problem we consider in this paper is estimation after selection with non-sparse weak signals whereby the subsampling approach is particularly attractive. The method allows us to reduce the scope in hunting for the causal SNPs in a GWAS. We also extend the proposed method to selections by Lasso-type approaches for correlated covariates. As a byproduct, the subsampling approach also provides variance estimates for the proposed estimators, which makes inference possible. We also compare the variance estimates by the proposed approach and EigenPrism.

The remainder of this thesis is organized as follows. In Chapter 2, we formulate the problem in a linear mixed model and demonstrate why the conventional sample splitting approach does not work well. A simulation study is conducted to compare the major approaches to solving this problem. In Chapter 3, we propose a subsampling approach with adjustment to cutoff values in selection for estimating and for making inference on the variation explained by a set of selected SNPs. Comprehensive simulation studies are performed to evaluate the proposed

approach in comparison to alternative approaches and applies the proposed approach to analyzing the expression quantitative trait loci (eQTL) in human brain tissues. Chapter 4 extends the proposed approach to correlated covariates. Chapter 5 discusses potential improvement and possible additional theoretical justification for the proposed approach.

CHAPTER 2

THE WEAK SIGNAL PROBLEM AND POSSIBLE SOLUTIONS

2.1 Problem Formulation

Let y be the outcome, usually the measurement of a complex disorder or trait. Let $x_1, \cdots, x_m$ be covariates denoting the typed SNP values. The SNP effects on the complex disorder or on a quantitative trait is modeled by a linear model as,

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \epsilon, \tag{2.1}$$

where $\beta_1, \beta_2, \ldots, \beta_p$ are independently distributed as $N(0, \sigma_g^2/p)$ and $\beta_{p+1}, \ldots, \beta_m$ are zeros, and ϵ is the random error distributed as $N(0, \sigma_\epsilon^2)$. Note that β_0 could be a function of other non-SNP confounders to be adjusted. For notation simplicity, we suppress them in the discussion. Let $Y = (y_1, \cdots, y_n)^t$ be the observed outcomes and $X_{n\times m}$ be the observed covariate matrix. The vector form of the linear model appears as

$$Y = \beta_0 \mathbf{1} + X\beta + \varepsilon,$$

where $\mathbf{1} = (1, \cdots, 1)^t$, $\beta = (\beta_1, \cdots, \beta_m)^t$, and $\varepsilon = (\epsilon_1, \cdots, \epsilon_n)^t$. The variance matrix of Y is

$$\mathrm{var}(Y) = \frac{1}{p} E(X_p X_p^t)\sigma_g^2 + I_n \sigma_\epsilon^2, \tag{2.2}$$

where X_p is the submatrix of X retaining the columns with indices corresponding to the nonzero indices of β. Note that $\sigma_g^2 = E||\beta||_2^2$. The total variation of a subject's trait explained by the covariate $x = (x_1, \cdots, x_m)^t$ is $\frac{1}{p}E\left(\sum_{j=1}^{p} x_j^2\right)\sigma_g^2$, which is σ_g^2 when $E(x_j^2) = 1$ for $j = 1, \cdots, p$. In this case, the proportion of variation explained by all the typed SNPs is

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\epsilon^2},$$

which is termed the narrow-sense heritability in genetic literature.

Under Equation 2.1, we may intuitively estimate individual β_j by the least-square approach based on the following linear model,

$$y = \beta_{0j} + \beta_j x_j + e_j, \tag{2.3}$$

for $j = 1, \cdots, m$, where $\text{var}(\beta_j) = \sigma_{ej}^2$. Since we do not know p, nor which covariates have nonzero effects, one might use $\sum_{j=1}^{m} \hat{\beta}_j^2$ to estimate σ_g^2. With large m, this estimator performs very poorly. Yang et al. (2010) proposed a restricted maximum likelihood approach to estimating σ_g^2 which is amount to using the approximation

$$\frac{1}{p}X_p X_p^t \approx \frac{1}{m}XX^t.$$

Numeric studies (Goeman, 2010; Yang et al., 2011; Deloukas et al., 2012; Lee et al., 2012) demonstrated that Yang et al's approach to estimating σ_g^2, and thus the narrow-sense heritability

h^2, has very good performance in presence of numerous weak effects. Jiang et al. (2016) showed that, when $p/m \to c$ and $n/m \to d$ where $c \leq 1$ and $d > 0$, Yang et al's approach yields a consistent estimator of σ_g^2.

Yang et al.'s approach can handle problems with varying levels of effect sparsity. We conduct a simple simulation study to demonstrate the robustness of the approach. The data are simulated using steps 1-3 of the algorithm in the simulation section in Chapter 3 with parameters $(n, m, \sigma_g^2, \sigma_e^2) = (200, 4000, 6, 4)$, and p varies from 20 to 4000. Figure 1 shows the box plots of the simulation results. Although LMM method assumes all the βs are normally distributed with mean zero, the estimation of σ_g^2 is unbiased even when the coefficients are sparse. For $p = 20$, Yang et al.'s approach performs similarly to that when $p = 4000$. This means Yang et al's approach is very robust to a wide range of the effect sparsity.

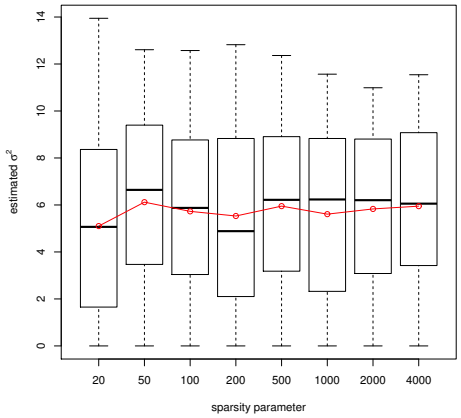


Figure 1: Box plot with varied sparsity using LMM method with n = 200, m =4000, $p = (20, 50, 100, \ldots, 4000)$, $\sigma_g^2 = 6$, $\sigma_e^2 = 4$, the red line is the truth

2.2 Variable Selection

Yang et al's approach does not yield a specific set of SNPs that are responsible for the explained variation. To further investigate this problem, we propose to select a subset of SNPs that are more plausible to account for the explained variation. Let S_λ be the set of indices of X selected with the tuning parameter value λ. Since the individual effect estimators are subject to large variation, we would like to estimate the variation explained by the selected SNPs instead, i.e., to estimate

$$\sigma^2_\lambda = E\left(\sum_{j\in S_\lambda} \beta_j^2\right). \tag{2.4}$$

Many variable selection approaches may be used. One simple selection approach is thresholding by individual tests based on the coefficient estimates from Equation 2.1. Let $\hat{\beta}_j$ and $\hat{\sigma}_{\beta j}$ denote respectively the estimator of β_j and the standard error estimate of $\hat{\beta}_j$. For a given significance level α, X_j is selected if

$$|\frac{\hat{\beta}_j}{\hat{\sigma}_{\beta j}}| \geq z_{1-\alpha/2}, \tag{2.5}$$

where $z_{1-\alpha/2}$ denotes the normal cutoff point for a given α, i.e., $\Phi(z_{1-\alpha/2}) = 1-\alpha$ where Φ is the normal distribution function. A simple selection using thresholding is to run the univariate regression of Y on a single X_j. This method is attractive when the covariates are independent.

Although the simple thresholding method can also be applied to the case with correlated covariates, which is more frequently occurred in practice, variable selection by the Lasso approach is more attractive when covariates are correlated. The Lasso approach minimizes

$$||y - \beta_0 - X\beta||_2^2 + \lambda||\beta||_1,$$

where $||a||_2^2 = \sum_{k=1}^{p} a_k^2$ and $||a||_1 = \sum_{k=1}^{p} |a_k|$ for any $a = (a_1, \cdots, a_p)$. Lasso and its extensions (Tibshirani (1996), Sun and Zhang (2012), Guo et al. (2016), Zhang and Zhang (2014)) have been shown to be very useful in the analyses of high-dimensional data. Given a penalty parameter λ, Lasso estimator of β is defined as

$$\hat{\beta}^L = \arg\min \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - X_i\beta)^2 + \lambda||\beta||_1 \right\}. \tag{2.6}$$

Since the elliptical contour of $\sum_{i=1}^{N}(y_i - \beta_0 - X_i\beta)^2$ is easy to hit corners, at which the corresponding regression coefficient is zero. The Lasso estimators are sparse, setting small βs to zero and shrinking the large βs towards 0. Lasso approach can perform automatic variable selection and estimation.

2.3 Estimation of Total Variation Explained by Selected Covariates

Upon variable selection, the total variation explained by the selected covariates can be estimated using different approaches. The approaches can be divided into two categories: one is to assemble individual estimators for the selected covariates. This approach usually involves correcting the bias in the individual estimators first, and then directly adding adjusted estimators

together with possibly additional corrections. The other approach uses the group of selected covariates to re-estimate $\sigma_\alpha^2 = E\left(\sum_{j\in S_\alpha} \beta_j^2\right)$ as a single target.

2.3.1 Methods by assembly of individual estimators

The simplest approach is the plug-in estimator. Let $\hat{\beta}_j$ be the estimator for β_j. The plug-in estimator estimates σ_S^2 by

$$\hat{\sigma}_{S1}^2 = \sum_{j\in S} \hat{\beta}_j^2, \tag{2.7}$$

where S represents the selected set of covariates. When $\hat{\beta}_j$ is an unbiased estimator of β_j conditional on jth variable being selected, this estimator overestimates the attributable variation to the selected covariates. This is because

$$E\left(\hat{\beta}_j^2 \mid j \in S\right) = \beta_j^2 + Var(\hat{\beta}_j \mid j \in S).$$

If we have a unbiased estimator of the variance of the β_j estimator, that is,

$$E(\hat{\sigma}_{\beta j}^2 \mid j \in S) = Var(\hat{\beta}_j \mid j \in S),$$

a natural unbiased estimator that corrects the upward bias in $\hat{\sigma}_{S1}^2$ is

$$\hat{\sigma}_{S2}^2 = \sum_{j\in S} (\hat{\beta}_j^2 - \hat{\sigma}_{\beta j}^2) \tag{2.8}$$

where $\hat{\sigma}^2_{\hat{\beta}j}$ is the estimated variance of $\hat{\beta}_j$.

In many practical applications, we only have an unbiased estimator of β_j and usually a (nearly) unbiased estimator of the variance of β_j without conditional on the variable being selected. That is,

$$E(\hat{\beta}_j) = \beta_j \text{ and } E(\hat{\sigma}^2_j) = Var(\hat{\beta}_j).$$

Conditional on the variable being selected, $\hat{\beta}_j$ is no longer unbiased for β_j. That is,

$$E(\hat{\beta}_j \mid j \in S) \neq \beta_j.$$

In this case, a correction of the selection bias is needed before the use of $\hat{\sigma}^2_{S2}$.

To correct the selection bias (also known as "winner's curse" in Galton (1886)), let $\hat{\beta}_j$ and $\hat{\sigma}^2_j$ be respectively the unbiased estimators (without conditional on being selected) for β_j and $var(\hat{\beta}_j)$ respectively. Given a cutoff value c, the sampling distribution of $\hat{\beta}_j$ (Zhong and Prentice, 2008) conditional on the jth variable being selected,

$$f_{\hat{\beta}_j || \frac{\hat{\beta}_j}{\hat{\sigma}_j} | \geq c}(x, \beta_j) = \frac{\frac{1}{\sigma_j}\phi(\frac{x-\beta_j}{\sigma_j})}{\Phi(\frac{\beta_j}{\sqrt{\sigma_j}} - c) + \Phi(-\frac{\beta_j}{\sqrt{\sigma_j}} - c)} 1(|\frac{x}{\sigma_j}| \geq c) \tag{2.9}$$

Given that variable j is selected, the expectation of $\hat{\beta}_j$ can be written as

$$E\left(\hat{\beta}_j \middle| |\frac{\hat{\beta}_j}{\hat{\sigma}_j}| \geq c\right) = \int_{|x|\geq c\sigma_j} x f_{\hat{\beta}_j||\frac{\hat{\beta}_j}{\hat{\sigma}_j}|\geq c}(x, \beta_j)dx$$
$$= \beta_j + \frac{\phi(\frac{\beta_j}{\sqrt{\sigma_j}} - c) - \phi(-\frac{\beta_j}{\sqrt{\sigma_j}} - c)}{\Phi(\frac{\beta_j}{\sqrt{\sigma_j}} - c) + \Phi(-\frac{\beta_j}{\sqrt{\sigma_j}} - c)} \tag{2.10}$$

By solving the Equation 2.10, we can obtain the adjusted β_j after variable selection. Denote the winner's curse adjusted estimator by $\hat{\beta}_{wj}$. A plug-in estimator for the variation explained by the selected covariates is

$$\hat{\sigma}^2_{S3} = \sum_{j\in S} \hat{\beta}^2_{wj}, \tag{2.11}$$

If we have also a variance estimate $\hat{\sigma}_{wj}$ for the $\hat{\beta}_{wj}$, an estimator with further correction is

$$\hat{\sigma}^2_{S4} = \sum_{j\in S} (\hat{\beta}^2_{wj} - \hat{\sigma}^2_{wj}), \tag{2.12}$$

which is called a double correction approach.

Another approach proposed by Chen (2016) is based on an approximately unbiased estimator of $\hat{\beta}^2_j$ conditional on being selected. The estimator appears as

$$\hat{\beta}^2_{cj} = \hat{\sigma}_{\beta j}\Big\{(\frac{\hat{\beta}^2_j}{\hat{\sigma}_{\beta j}} - 1)1_{j\in S} - 2\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}}_{\beta j}}\frac{\sqrt{1+\xi^2}}{\sqrt{2\pi}}\big[\exp(-\frac{\xi^2\hat{a}^2_j}{2}) - \exp(-\frac{\xi^2\hat{b}^2_j}{2})\big]$$
$$+ \frac{(\xi^2-1)\sqrt{1+\xi^2}}{2\pi}\big[\hat{a}_j\exp(-\frac{\xi^2\hat{a}^2_j}{2}) - \hat{b}_j\exp(-\frac{\xi^2\hat{b}^2_j}{2})\big]\Big\} \tag{2.13}$$

where $\hat{a}_j = z_{1-\frac{\alpha}{2}} - \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}_{\beta j}}}$, $\hat{b}_j = z_{\frac{\alpha}{2}} - \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}_{\beta j}}}$, and ξ is a constant which may be set between 4 and 10. An estimator of the variation explained by the selected covariates is

$$\hat{\sigma}^2_{S5} = \sum_{j \in S} \hat{\beta}^2_{cj}. \tag{2.14}$$

All of the above approaches aim at adjusting individual β_j estimates and then add all the adjusted estimates together. Although many of the adjusted β estimators have been shown to be asymptotically unbiased for estimating individual β_j, the additive effect of a large number of small errors in individual β_j estimates can accumulate to lead to a large bias.

The individual unbiased estimators for each β_j may be obtained by the marginal regression based on Equation 2.3. This works well when the covariates are independent. When covariates are dependent, Lasso approach may be used to select and estimate the regression coefficients. However, the Lasso estimator of the regression parameters are biased and the distributions of of the Lasso estimators are difficult to characterize. Zhang and Zhang (2014) and Geer et al. (2014) proposed an approach called low-dimensional projection estimator (LDPE) which can reduce the bias in the Lasso estimators and provide a variance estimate for the debiased Lasso estimators.

The LDPE estimator can be expressed as,

$$\hat{\beta}_j = \hat{\beta}^0_j + \frac{\boldsymbol{z_j^T\{y - X\hat{\beta}^0\}}}{\boldsymbol{z_j^T x_j}} \tag{2.15}$$

where $\hat{\beta}^0$ is the estimators from scaled Lasso, and $\boldsymbol{z}_j$ is the relaxed projection.

Given $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\epsilon \sim N(0, \sigma^2\boldsymbol{I})$, we have the estimated error of Equation 2.15 is

$$\hat{\beta}_j - \beta_j = \frac{\boldsymbol{z}_j^T\boldsymbol{\epsilon}}{\boldsymbol{z}_j^T\boldsymbol{x}_j} + \frac{1}{\boldsymbol{z}_j^T\boldsymbol{x}_j}\sum_{k\neq j}\boldsymbol{z}_j^T\boldsymbol{x}_k(\beta_k - \hat{\beta}_k^0) \tag{2.16}$$

The first term in the Equation 2.16 is defined as the noise term and the second term is the approximation error term controlled by the $\max_{k\neq j}|\boldsymbol{z}_j^T\boldsymbol{x}_k|/||\boldsymbol{z}_j||_2$. To control both the noise and approximation error, we need to find a proper λ_j for the estimation of z_j.

The relaxed projection matrix for the sparse $\boldsymbol{X}_j$ to the orthogonal complement of the column space of $\boldsymbol{X}_{-j}$

$$\boldsymbol{z}_j = \boldsymbol{x}_j - \boldsymbol{X}_{-j}\hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\gamma}}_j = \arg\min_{\gamma_j}\left\{\frac{||\boldsymbol{x}_j - \boldsymbol{X}_{-\boldsymbol{j}}\boldsymbol{\gamma}_{\boldsymbol{j}}||_2^2}{2n} + \lambda_j||\boldsymbol{\gamma}_{\boldsymbol{j}}||_1\right\} \tag{2.17}$$

λ_j is chosen based on the error and variance in the final estimation of β. By following the iterative strategy in Zhang and Zhang (2014), we can find a reasonable λ_j with relatively large but acceptable bias and small variance. Finally, the estimation of $\sigma_{ej} = \frac{||z_j||_2}{|z_j^T x_j|}\sigma$ based on Equation 2.16 where $\frac{\boldsymbol{z}_j^T\boldsymbol{\epsilon}}{||\boldsymbol{z}_j||_2} \sim N(0, \sigma^2)$ and error term is bounded with suitable z_j.

A theoretical justification of LDPE is provided in Zhang and Zhang (2014). If the data is satisfied with the regularity condition, $\hat{\beta}_j - \beta_j$ approximates normality with mean 0 and variance $\hat{\sigma}^2\frac{||\boldsymbol{z}_j^T||_2^2}{|\boldsymbol{z}_j^T\boldsymbol{x}_j|^2}$. Since the variance estimator of $\hat{\beta}_j$ is $\hat{\sigma}^2$ times the noises of the estimator, the estimator is quite dominated by the noise estimator in the scaled Lasso regression, which in turn, depends on the sparsity assumption of the data. Furthermore, the iterative approach of searching for proper λ_j in the Z matrix estimation through loosing the noise τ_j may widen

confidence interval of β_j. With $\hat{\beta}_j$ and the corresponding $\hat{\sigma}_{ej}$ from LDPE, we can apply the same individual tests in Equation 2.5 to select X.

Since no code is available for LDPE approach, before applying this approach in our simulation, we wrote our own R codes to implement LDPE approach, and validate our coded by repeating the simulation results published in Zhang and Zhang (2014). See Appendix A for more details.

2.3.2 Direct estimating the explained variation as a parameter

Adjusting bias in individual coefficient estimates and adding the adjusted estimates together are easy. However, the approach may not be efficient. Furthermore, the cumulative effect may result in a large bias in the final estimator even through the biases are small in individual coefficient estimators. To resolve the problems, we consider direct estimation approaches in this section.

The scaled Lasso (Sun and Zhang, 2012) can in principle be used to estimate the total effects as follows. Following the solution path of iteratively estimating noise level σ, β, and λ, the scaled Lasso algorithm converges to a set of $\hat{\beta}^L$ and $\hat{\sigma}$ such that

$$\hat{\beta}^L, \hat{\sigma} = \arg\min_{\beta,\sigma} \Big\{ \frac{\sum_{i=1}^{N}(y_i - \beta_0 - X_i\beta)^2}{2\sigma n} + \frac{\sigma}{2} + \lambda||\beta||_1 \Big\} \tag{2.18}$$

The initial penalty level of the scaled Lasso is set to $\sqrt{(2/n)\log m}$. The estimation of β and σ^2 are bounded as long as $\lambda = A\sqrt{(2/n)\log(m/\epsilon)}$ given some $A > 1$ and $\epsilon \in (0, 1]$. Since $\hat{\sigma}$ in the scaled Lasso is consistent and asymptotically normally distributed when the non-zero

coefficients are sparse and reasonably strong (Sun and Zhang, 2012), we can estimate σ_g^2 by subtracting the $\hat{\sigma}^2$ from the total variance. The following variance equality

$$\sigma_g^2 = Var(Y) - \sigma^2 \tag{2.19}$$

immediately implies a plug-in estimator.

Recently, Guo et al. (2016) proposed a new estimator for $\sum_{j=1}^{m} \beta_j^2$ under the high-dimensional linear model framework. Their approach uses the scaled Lasso estimation at the initial step, adjusts the $\sum_{j=1}^{m} \beta_j^2$ estimate by estimating the $< \hat{\beta}, \hat{\beta} - \beta >$, where $\hat{\beta}$ is the coefficient estimator from the scaled Lasso. From

$$\sum_{j=1}^{m} \left(\hat{\beta}_j^2 - \beta_j^2 \right) = 2 < \hat{\beta}, \hat{\beta} - \beta > - \sum_{j=1}^{m} (\hat{\beta} - \beta)^2, \tag{2.20}$$

since the second term on the right hand side in Equation 2.20 is negligible, to correct the bias in $\sum_{j=1}^{m} \hat{\beta}^2$, we need to estimate the quantity $< \hat{\beta}, \beta - \hat{\beta} >$. To estimate the $< \hat{\beta}, \beta - \hat{\beta} >$, a projection vector u is identified to control the following difference,

$$\frac{1}{n} u^t X^t (Y - X\hat{\beta}) - < \hat{\beta}, \beta - \hat{\beta} > = (u^t \hat{\Sigma} - \hat{\beta})(\beta - \hat{\beta}) + \frac{1}{n} u^t X^t \epsilon \tag{2.21}$$

where $\hat{\Sigma} = \frac{X^t X}{n}$. The u can be solved through the following constraint optimization.

$$\arg\min_{u} \left\{ u^t \hat{\Sigma} u : ||\hat{\Sigma} u - \hat{\beta}||_\infty \leq ||\hat{\beta}||_2 \frac{\lambda_1}{\sqrt{n/2}} \right\} \tag{2.22}$$

Equation 2.22 can be solved through its equivalent Lagrange dual problem. The idea of Equation 2.22 is to control the upper bound of $(u^t\hat{\Sigma} - \hat{\beta})(\beta - \hat{\beta})$ and the variance of $\frac{1}{n}u^t X^t \epsilon$ in Equation 2.21. After obtaining $\hat{u}$, $< \hat{\beta}, \beta - \hat{\beta} >$ is then estimated by $\frac{1}{n}u^t X^t (Y - X\hat{\beta})$ as the right hand side is canceled out. They called the estimator thus obtained the functional de-biased estimator (FDE).

This method appears promising, as it treats $\sum_{j=1}^{m} \beta^2$ as a single parameter, and directly adjust the total bias in the scaled Lasso estimator. On the other hand, the sparsity requirement may limit its performance when the requirement is satisfied. In addition, since the error bound of FDE is positively associated with $||\beta||_2$, the bias of the quadratic form estimation increases as the scale of the signal strength increases.

In our simulation, we implemented FDE using R code we wrote. To verify the correctness of the codes and to demonstrate the possible limitations discussed above, simulations are conducted with similar set-up in the (Guo et al., 2016) and additional cases. See Appendix B for details.

Another approach to estimating the total variation explained by the selected covariates is to apply Yang et al's approach directly to the selected covariates. We call this approach the naive application of Yang et al.'s approach. This is because, the approach is no longer asymptotically unbiased for estimating the explained variation once the covariates are selected non-randomly. To correct this bias, we also consider an alternative approach based on Yang et al., which split the sample into two halves. One half is used to select the covariates. The other half is used to estimate the variation explained by the selected covariates.

2.4 Simulation Designs and Results

In this section, I conduct extensive simulation study to compare the methods discussed earlier. In generating the simulation data, I consider two different scenarios. Simulated cases in the first scenario are consistent with the assumption of Yang et al.'s approach. Non-zero β_js are random and follows a normal distribution, and all the covariates are independent. The set-up is as follows.

1. $\mathbf{X}$: Generate m random numbers following $U(0.05, 0.5)$ distribution, representing as $\mathbf{p}$; generate $n \times m$ matrix with each column following $Binomial(2, \mathbf{p}_i)$;
2. $\boldsymbol{\beta}$: for $\beta_i \in p$, it follows $N(0, \sigma_g^2/p)$, and $\beta_j = 0$ otherwise;
3. ϵ: follows $N(0, \sigma_e^2)$
4. $\boldsymbol{Y}$: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$

The parameter $(m, n, p, \sigma_g^2, \sigma_e^2)$ are varied to account for different situations. m is fixed at 4000 and p is fixed at 40 here. n is varied in 400 and 1000 to see if any approaches might be influenced by the sample size. Furthermore, I vary the size of $\frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$ from 0.4 to 0.6 to see that how these approaches will differ with different signal ratios. The total variation $\sigma_g^2 + \sigma_e^2$ is chosen between 10 and 50 to account for the random noises in penalized regression.

In the second scenario of data generation, SNPs are correlated and and their effects are fixed. The specific set-up is as follows,

1. $\mathbf{X}$: Generate X_j following $N(0, \Sigma_x)$, where the ij entry of Σ_x, $\Sigma_x(i, j) = \rho^{|i-j|}$ for a given $\rho > 0$

2. $\boldsymbol{\beta}$: for $\beta_j = \tau/2 * (1 + j/p)$ for $j = 1, \cdots, p$ and $\beta_j = 0$ otherwise;

3. ϵ: follows $N(0, \sigma_e^2)$

4. $\boldsymbol{Y}$: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$

The analysis of the simulated data includes two steps. The first step selects variables based on certain criteria. The second step applies the methods to estimate variation explained by the selected covariates. For the variable selection, two approaches are used here, one is the simple linear marginal estimation; and the other is the LDPE approach. For the explained variation estimation, the following approaches are used: simple adjustment, error adjustment, winner's curse adjustment, doubly adjustment, Chen's approach, LMM-based approach, scaled Lasso estimation, LMM-based approach with data split (1:1), and FDE approach. The thresholding levels are set to $-\log \alpha = 0, \cdots, 9$, where α is the test significance level. Simulation results are based on 100 replicates.

2.4.1 Experiment 1

For the first scenario, Table I-VII display the mean square error of various estimators at different selection criteria. The mean of estimates of different approaches at various α level are plotted in figures (see Figure 2-16. The boxplots (Figure 3-17) of the estimates from each approaches are plotted alongside of the truth for easy comparison. Similarly, Table IX-X, Figure 18-21 are result displays for the second simulation scenario.

For the first set with n = 400, m =4000, p = 40, $\sigma_g^2 = 3$, $\sigma_e^2 = 7$, Table I shows that scaled Lasso gives the smallest MSE value when $-\log(\alpha)$ is small. As the selection criterion becomes stricter, LMM approach consistently outperforms the other approaches. Since MSE is a

measurement combining the estimator means and variances, we also plot the mean of estimators across simulations to see how the approaches performs respect to the means. Figure 2 shows that both Chen's approach and LMM with split follow the true trend quite well. Specifically, Chen's approach consistently outperforms the others when $-\log(\alpha)$ is greater than 10. On the other side, as the $-\log(\alpha)$ becomes larger, the LMM split estimator differs more. For the LMM, scaled Lasso, and FDE approaches, the estimators do not follow the trend quite well, suggesting that they might have a relatively large bias and small variances. Figure 3 displays the mean variance trade-off with more details. It shows that Chen's approach consistently cover the truth with larger variances when $-\log(\alpha) \leq 5$. For the other approaches (i.e. LMM, LMM split, scaled Lasso and FDE) which give smaller variances, the mean of estimators are too off compared with truth.

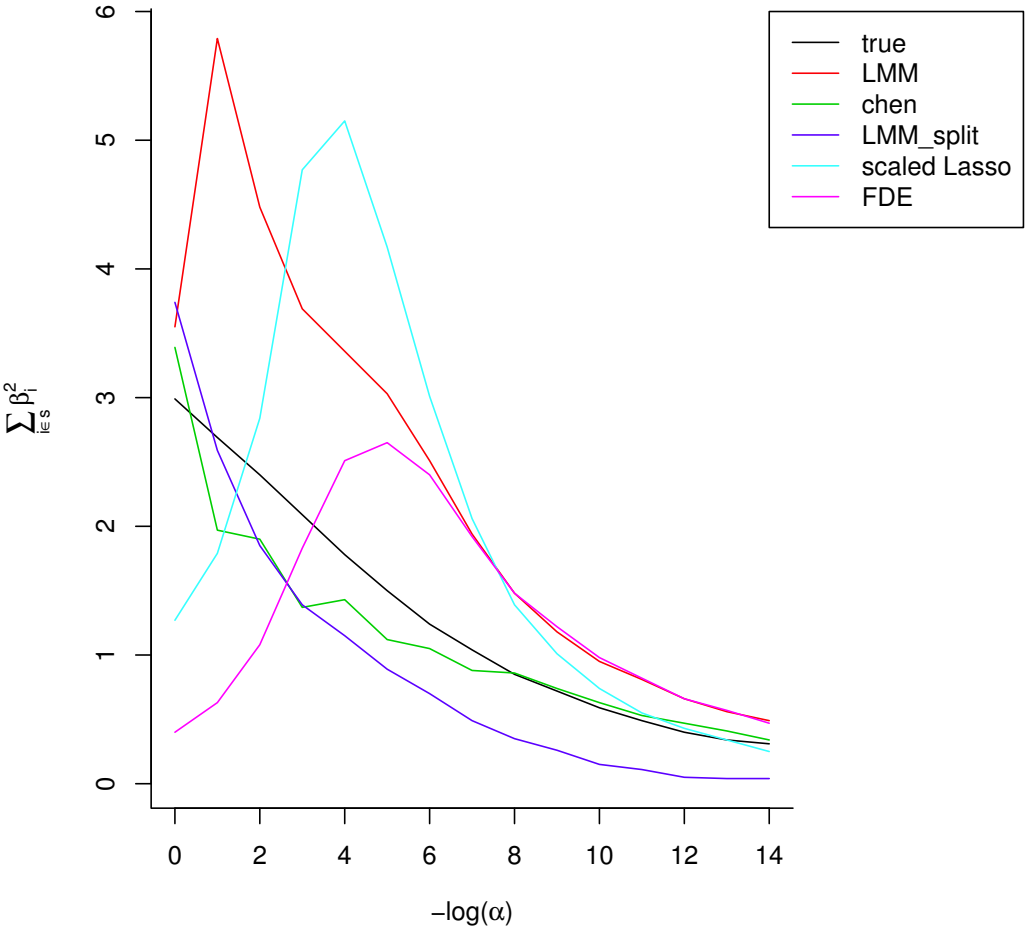
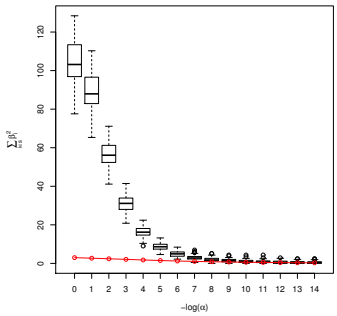


Figure 2: Mean estimation with varied α for part approaches in Experiment 1

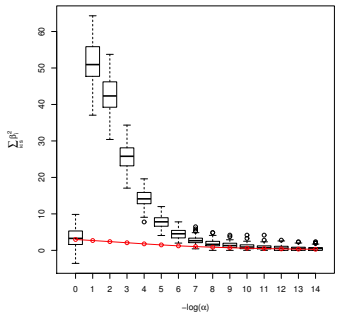
TABLE I: RESULTS IN EXPERIMENT 1 FOR CURRENT METHODS [a]

$-\log(\alpha)$	simple	error	winner's curse	doubly adjustment	Chen	LMM	LMM-split	scaled Lasso	FDE
0	10420.84	6.66	10420.84	6.66	6.66	5.42	9.89	3.41	7.06
1	7552.04	2409.19	182.44	598.9	82.2	9.9	6.02	1.28	4.6
2	2976.88	1633.01	2.87	165.61	45.27	4.61	3.03	0.62	2.06
3	863.86	577.65	0.9	35.98	26.19	2.83	1.85	7.5	0.34
4	220.12	162.27	1.28	9.38	6.87	2.74	1.23	11.69	0.77
5	55.34	43.05	1.15	3.48	4.74	2.58	0.9	7.49	1.55
6	14.64	11.76	0.91	1.71	1.94	1.8	0.69	3.42	1.56
7	4.42	3.61	0.78	1.12	1.01	0.96	0.61	1.36	1.01
8	1.71	1.41	0.67	0.85	0.71	0.54	0.53	0.59	0.68
9	0.78	0.65	0.56	0.67	0.39	0.32	0.45	0.45	0.46
10	0.42	0.35	0.42	0.48	0.24	0.22	0.4	0.38	0.37
11	0.31	0.27	0.39	0.44	0.25	0.19	0.35	0.37	0.31
12	0.23	0.2	0.31	0.35	0.2	0.15	0.29	0.31	0.25
13	0.17	0.15	0.24	0.27	0.16	0.11	0.23	0.27	0.23
14	0.14	0.12	0.21	0.23	0.17	0.09	0.19	0.41	0.21

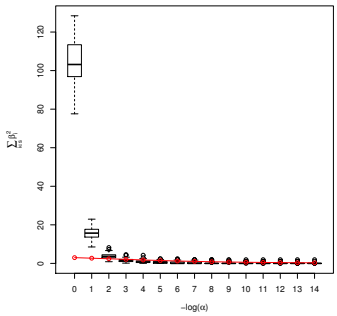
[a] n = 400, m =4000, p = 40, $\sigma_g^2 = 3$, $\sigma_e^2 = 7$, the marginal β_j and σ_j are estimated through univariate linear model.



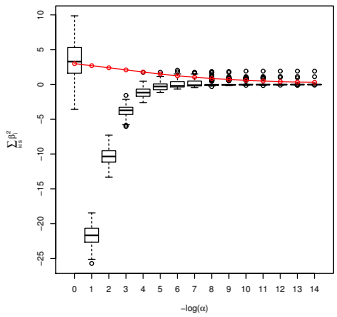
(a) Simple approach



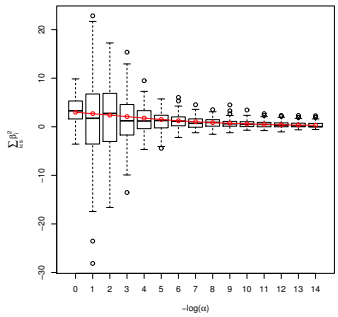
(b) Error approach



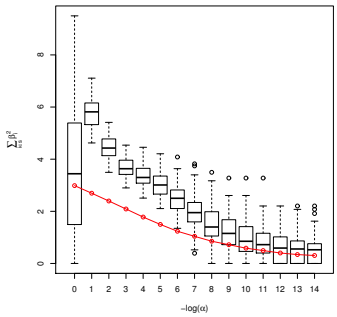
(c) Winner's Curse approach



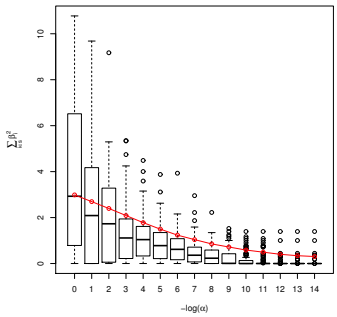
(d) Doubly approach



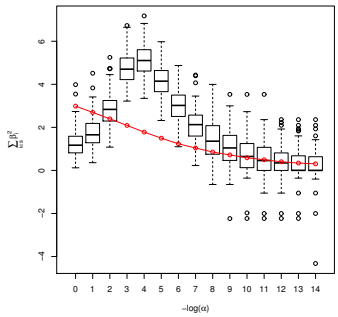
(e) Chen's approach



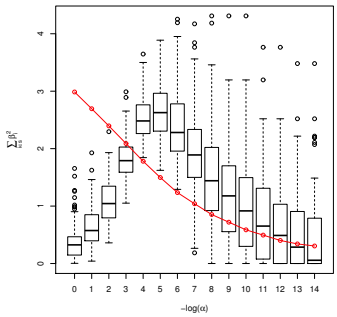
(f) LMM approach



(g) LMM Split approach



(h) scaled Lasso approach



(i) FDE approach

Figure 3: Box plot with varied α for part approaches in Experiment 1

2.4.2 Experiment 2

Table II, Figure 4-5 present the results when $\sigma_g^2 = 4$ and $\sigma_e^2 = 6$. It can be seen that LMM gives a similar MSE with the scaled Lasso when $-\log(\alpha) = 0$. Although Chen's approach give a very large MSE at the beginning, it quickly drops down after $-\log(\alpha) \geq 4$. The mean estimation figure gives the similar results as previous. The boxplot in Figure 5 shows that the LMM with split effectively reduces the bias in the LMM upon selection.

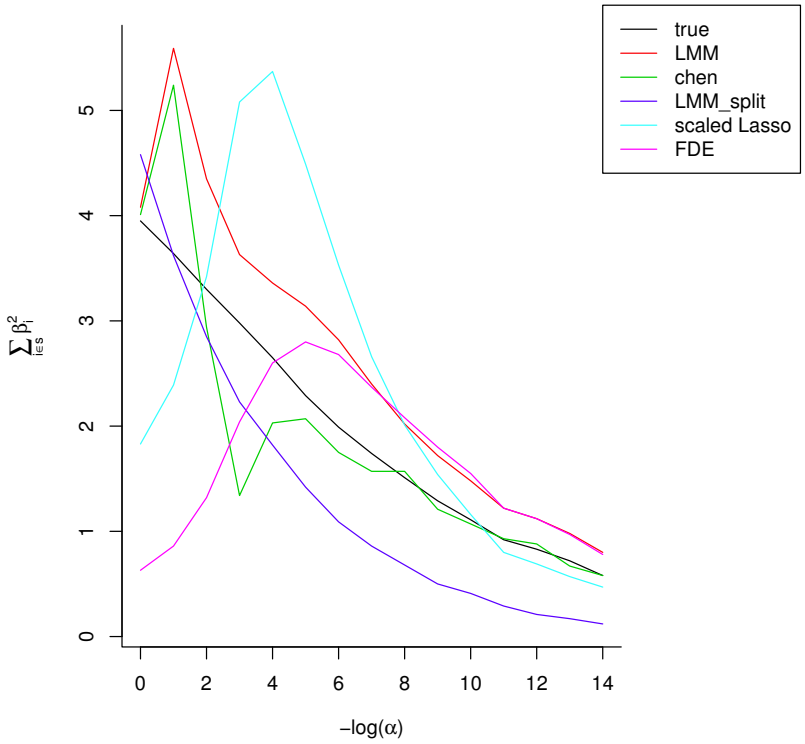
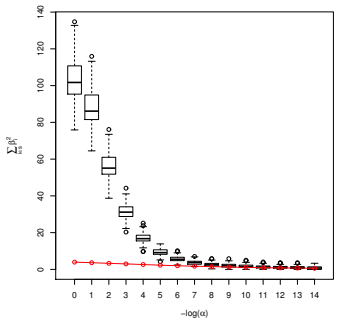


Figure 4: Mean estimation with varied α for part approaches in Experiment 2

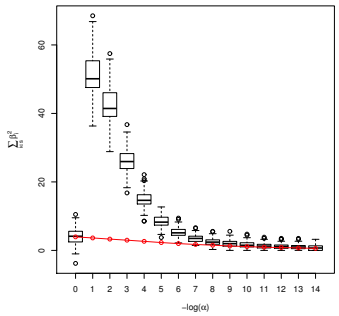
TABLE II: RESULTS IN EXPERIMENT 2 FOR CURRENT METHODS [a]

$-\log(\alpha)$	simple	error	winner's curse	doubly adjustment	Chen	LMM	LMM-split	scaled Lasso	FDE
0	9956.11	5.7	9956.11	5.7	5.7	4.92	11.28	4.91	11.47
1	7199.57	2300.13	167.32	581.34	129.58	4.15	8.28	1.97	8.15
2	2852.18	1561.66	2.21	163.6	63.35	1.45	3.38	0.4	4.24
3	844.05	561.1	1.37	39.33	34.58	0.78	1.88	4.78	1.21
4	212.56	155.44	2.14	11.9	8.65	0.83	1.64	7.73	0.31
5	52.76	40.6	1.87	4.81	4.94	0.95	1.41	5.12	0.46
6	14.95	11.8	1.83	3.04	2.68	0.89	1.34	2.66	0.69
7	4.54	3.63	1.53	2.09	1.27	0.58	1.18	1.16	0.62
8	1.87	1.51	1.26	1.55	0.7	0.37	1	0.6	0.55
9	1.11	0.91	1.13	1.33	0.53	0.3	0.94	0.44	0.53
10	0.67	0.56	1.04	1.19	0.44	0.25	0.8	0.46	0.47
11	0.42	0.36	0.78	0.87	0.37	0.19	0.71	0.58	0.38
12	0.32	0.27	0.73	0.81	0.38	0.18	0.7	0.54	0.35
13	0.25	0.22	0.65	0.71	0.41	0.16	0.55	0.5	0.32
14	0.21	0.18	0.5	0.55	0.26	0.13	0.48	0.44	0.29

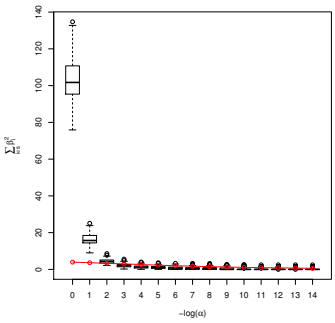
[a] n = 400, m =4000, p = 40, $\sigma_g^2 = 4$, $\sigma_e^2 = 6$, the marginal β_j and σ_j are estimated through univariate linear model.



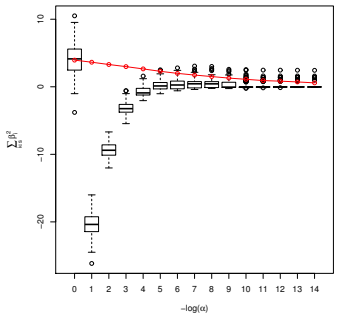
(a) Simple approach



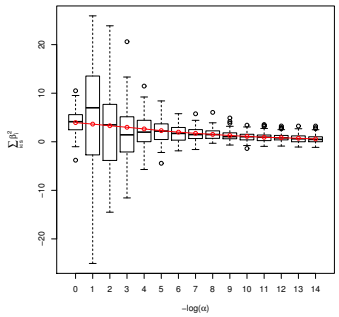
(b) Error approach



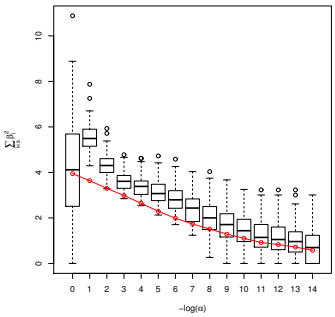
(c) Winner's Curse approach



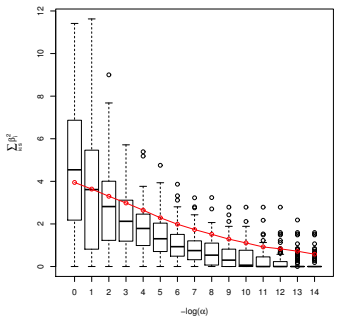
(d) Doubly approach



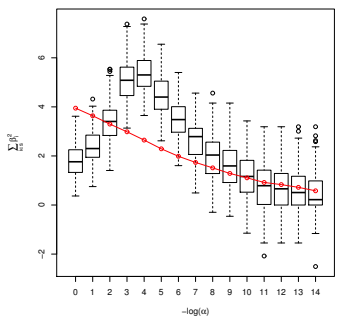
(e) Chen's approach



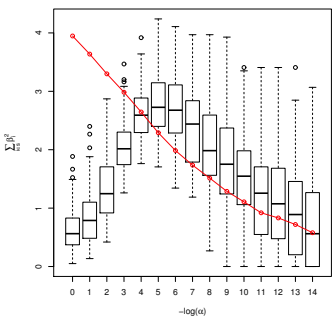
(f) LMM approach



(g) LMM Split approach



(h) scaled Lasso approach



(i) FDE approach

Figure 5: Box plot with varied α for part approaches in Experiment 2

2.4.3 Experiment 3

In addition to the marginal linear estimations, for the set of $\sigma_g^2 = 4$, $\sigma_e^2 = 6$, we also did the LDPE for the variable selection. For those developed to adjust bias in single linear estimation, the MSE increases dramatically especially when $-\log(\alpha)$ is small. On the other hand, the marginal estimator from LDPE seem serve better in the variable selection for those directly-estimation approach.

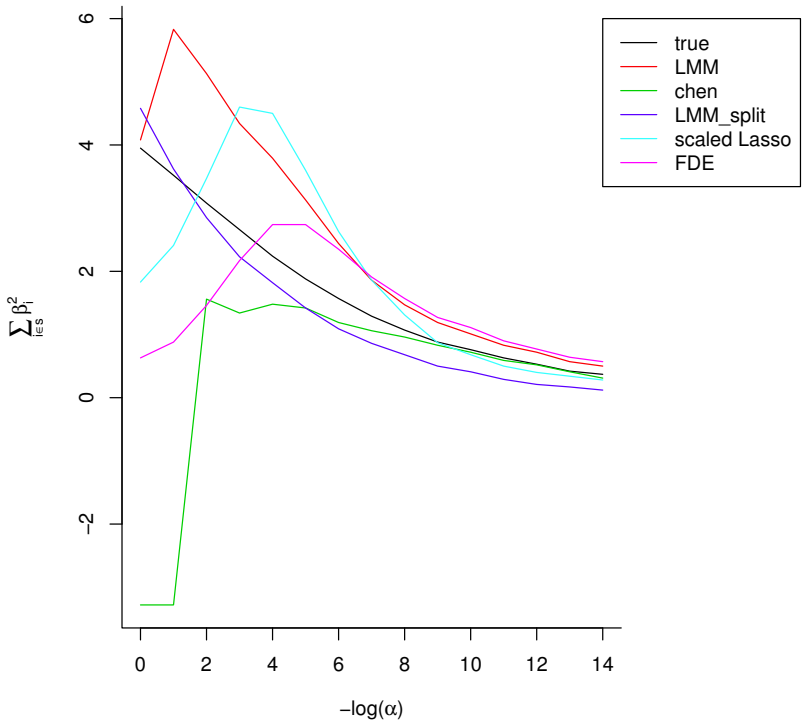
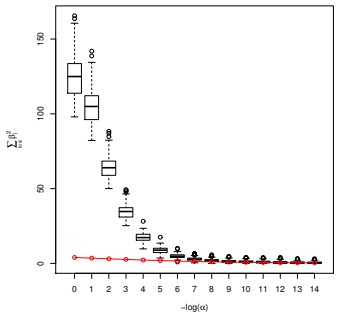


Figure 6: Mean estimation with varied α for part approaches in Experiment 3

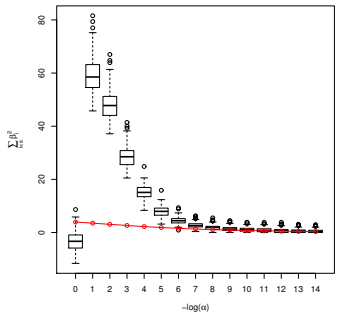
TABLE III: RESULTS IN EXPERIMENT 3 FOR CURRENT METHODS [a]

$-\log(\alpha)$	simple	error	winner's curse	doubly adjustment	Chen	LMM	LMM-split	scaled Lasso	FDE
0	14983.89	66.22	14983.89	66.22	66.22	4.92	11.28	4.91	11.47
1	10573.57	3175.12	265.92	923.43	251.92	5.7	8.26	1.63	7.43
2	3865.94	2082.58	2.22	241.54	103.4	4.59	3.33	0.59	2.95
3	1035.78	684.22	1.61	49.06	43.91	3.13	1.55	4.16	0.5
4	237.06	172.86	2.02	12.38	12.09	2.73	1.11	5.43	0.5
5	52.24	40.13	1.77	4.51	6.14	1.8	0.87	3.24	0.93
6	12.85	10.14	1.52	2.49	3.18	0.94	0.71	1.43	0.83
7	3.63	2.92	1.4	1.85	1.49	0.49	0.56	0.62	0.6
8	1.32	1.07	1.1	1.34	0.91	0.26	0.43	0.36	0.48
9	0.7	0.58	0.92	1.07	0.45	0.19	0.4	0.46	0.38
10	0.41	0.35	0.73	0.83	0.53	0.16	0.36	0.43	0.35
11	0.32	0.27	0.57	0.64	0.38	0.13	0.32	0.45	0.3
12	0.25	0.21	0.46	0.52	0.33	0.11	0.28	0.52	0.27
13	0.16	0.14	0.36	0.4	0.24	0.08	0.21	0.29	0.21
14	0.14	0.12	0.34	0.38	0.2	0.07	0.26	0.34	0.19

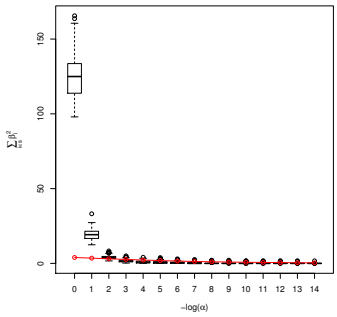
[a] n = 400, m =4000, p = 40, $\sigma_g^2 = 4$, $\sigma_e^2 = 6$, the marginal β_j and σ_j are estimated through LDPE approach.



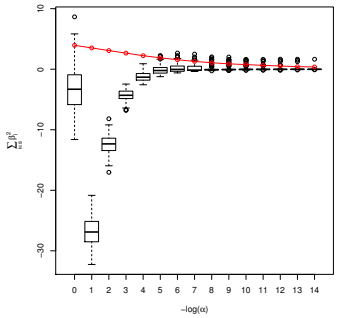
(a) Simple approach



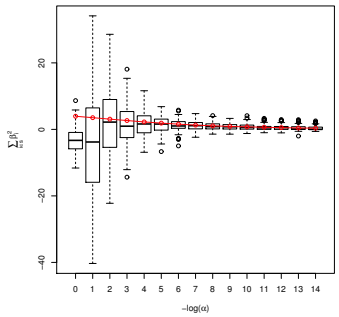
(b) Error approach



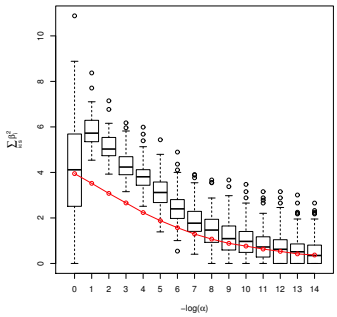
(c) Winner's Curse approach



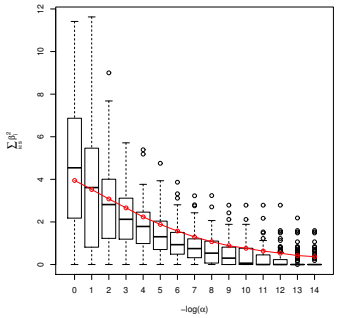
(d) Doubly approach



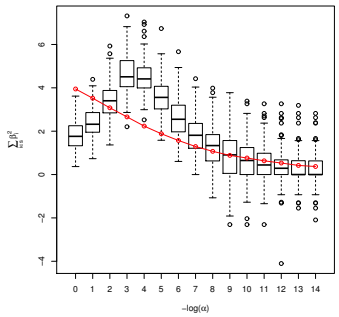
(e) Chen's approach



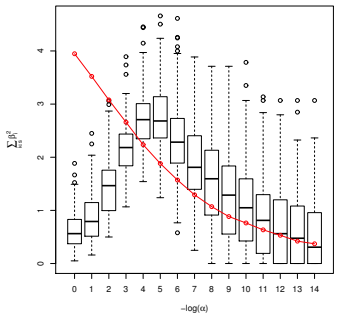
(f) LMM approach



(g) LMM Split approach



(h) scaled Lasso approach



(i) FDE approach

Figure 7: Box plot with varied α for part approaches in Experiment 3

2.4.4 Experiment 4

To check the effects of sample size, we also increase the sample size from 400 to 1000. In Table IV, it can be seen that all the estimators' MSE are reduced when sample size is large. Specifically, scaled Lasso gives a very precise estimation in this case due to the satisfactory of sparsity condition. LMM also outperforms the others with smaller MSE. In Figure 8, Chen's approach is very unstable at the beginning of the selection, and it represents the truth very well after $-\log(\alpha) = 4$. Figure 9 shows the estimators in the simulation can cover the truth better compared with that when $n = 400$.

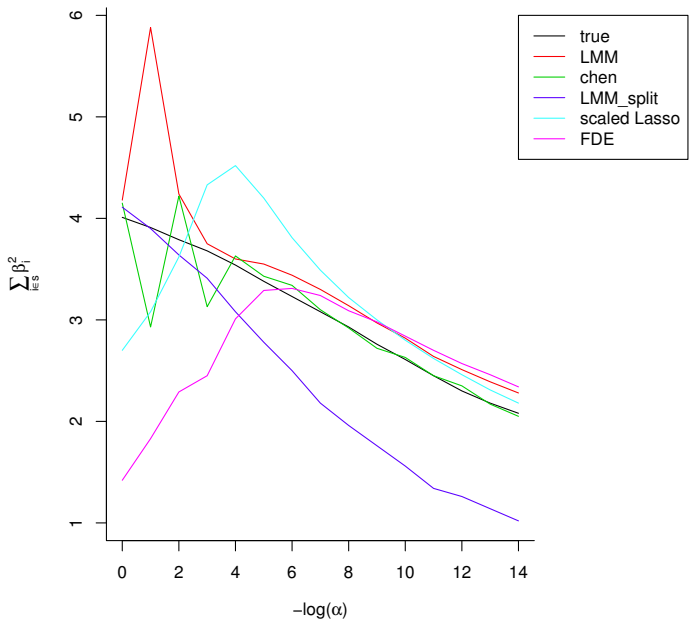
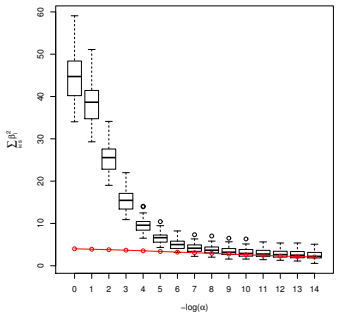


Figure 8: Mean estimation with varied α for part approaches in Experiment 4

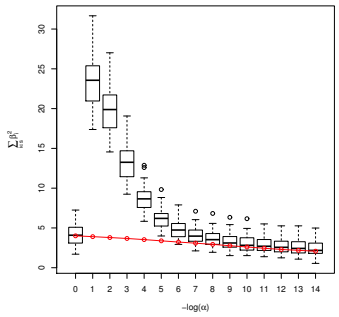
TABLE IV: RESULTS IN EXPERIMENT 4 FOR CURRENT METHODS [a]

$-\log(\alpha)$	simple	error	winner's curse	doubly adjustment	Chen	LMM	LMM-split	scaled Lasso	FDE
0	1665.74	1.19	1665.74	1.19	1.19	1.07	4.5	1.96	7.03
1	1205.23	382.04	12.34	138.65	16.49	4.07	1.96	0.89	4.66
2	472.33	258.11	0.37	30.14	8.9	0.43	0.7	0.19	2.52
3	141.39	93.37	0.54	7.59	4.21	0.24	0.54	0.55	1.67
4	38.23	27.54	0.81	3	1.3	0.17	0.67	1.11	0.42
5	10.72	8	1.03	2.01	0.68	0.14	0.78	0.82	0.11
6	3.65	2.74	1.38	1.99	0.59	0.15	0.96	0.48	0.12
7	1.56	1.18	1.53	1.95	0.37	0.15	1.27	0.32	0.14
8	0.93	0.72	1.75	2.09	0.26	0.16	1.4	0.24	0.17
9	0.6	0.47	1.64	1.9	0.32	0.17	1.5	0.23	0.24
10	0.46	0.36	1.57	1.79	0.21	0.15	1.6	0.2	0.26
11	0.37	0.3	1.57	1.75	0.31	0.14	1.74	0.19	0.26
12	0.34	0.28	1.47	1.62	0.22	0.14	1.61	0.18	0.26
13	0.31	0.26	1.51	1.65	0.2	0.12	1.6	0.17	0.27
14	0.25	0.21	1.56	1.69	0.2	0.11	1.61	0.16	0.26

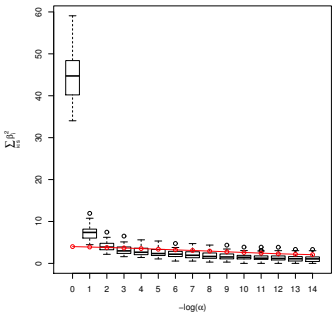
[a] n = 1000, m =4000, p = 40, $\sigma_g^2 = 4$, $\sigma_e^2 = 6$, the marginal β_j and σ_j are estimated through the univariate linear model.



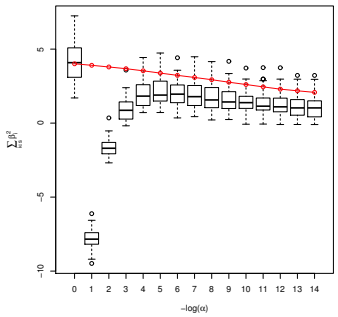
(a) Simple approach



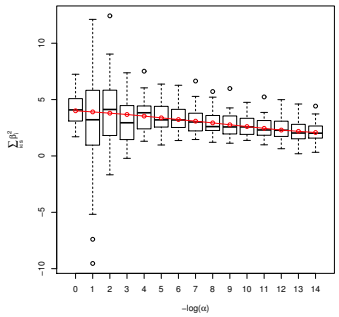
(b) Error approach



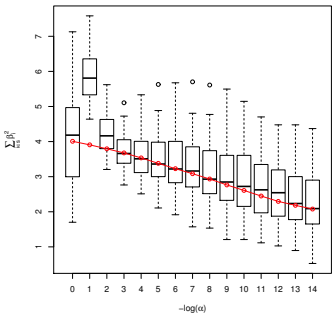
(c) Winner's Curse approach



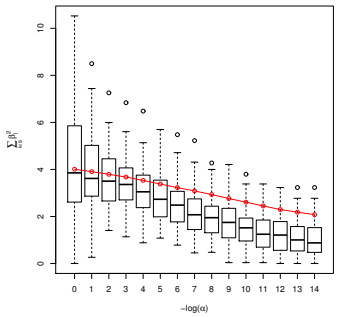
(d) Doubly approach



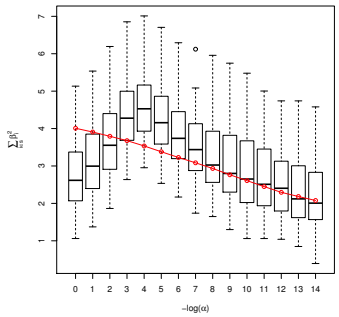
(e) Chen's approach



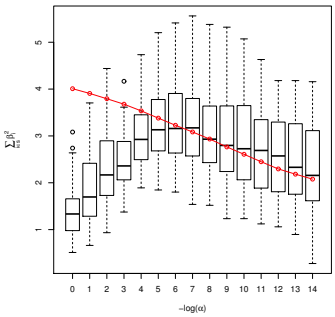
(f) LMM approach



(g) LMM Split approach



(h) scaled Lasso approach



(i) FDE approach

Figure 9: Box plot with varied α for part approaches in Experiment 4

2.4.5 Experiment 5

Since all the previous σ_g^2s are not very large, we would like to see if the good performances of scaled Lasso is due to its preference in reducing small βs to zero. σ_g^2 and σ_e^2 are increased to 20 and 30, respectively. So the heritability does not change. It can be seen in Table V that the scaled Lasso no longer work as well as LMM when $-\log(\alpha) = 0$. When the selection criterion become stricter, LMM is still the most favorable approach. In Figure 10, Chen's approach still gives the most similar trend with truth while its estimators are not quite stable.

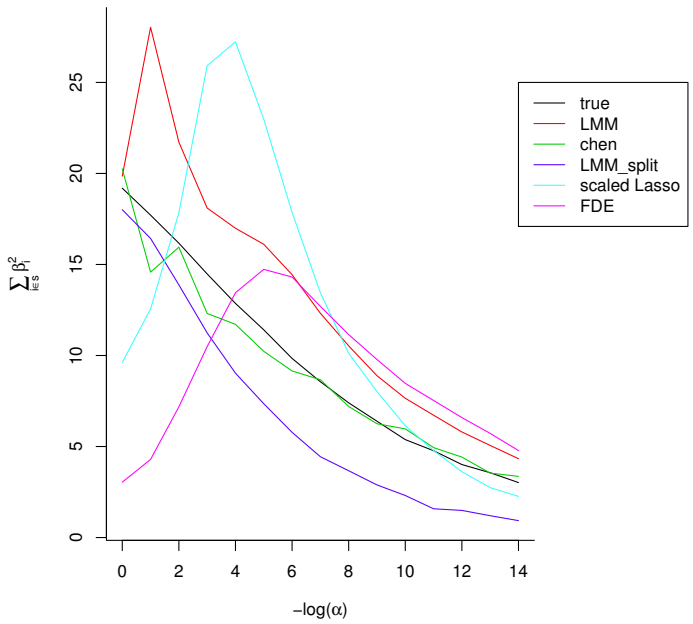
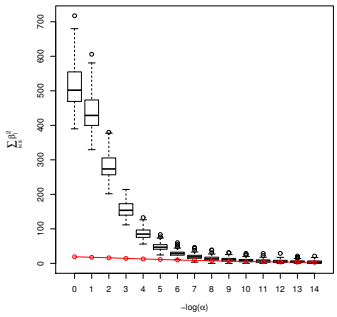


Figure 10: Mean estimation with varied α for part approaches in Experiment 5

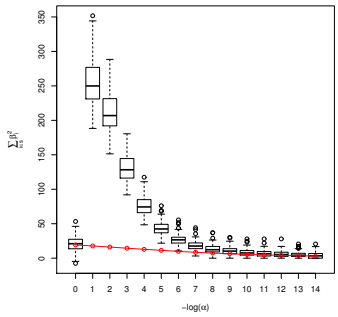
TABLE V: RESULTS IN EXPERIMENT 5 FOR CURRENT METHODS [a]

$-\log(\alpha)$	simple	error	winner's curse	doubly adjustment	Chen	LMM	LMM-split	scaled Lasso	FDE
0	250674.33	128.84	250674.33	128.84	128.84	99.64	288.85	109.1	273.33
1	181449.79	57997.18	16960.32	3175.88	2389.59	112.83	158.11	42.64	186.65
2	71715.98	39373.23	952.2	1602.78	1182.3	38.51	70.9	16.15	89.2
3	21113	14109.94	59.82	463.64	521.67	20.06	42.13	139.94	22.34
4	5575.84	4097.83	25.73	191.07	260.74	24.59	35.76	214.57	8.28
5	1431.67	1106.31	30.2	86.37	145.89	28.78	32.51	142.24	17.97
6	410.23	326.45	29.11	51.4	55.62	27.01	31.06	73.28	28.24
7	141.81	114.76	29.8	41.37	39.47	19.85	33.41	33.44	26.09
8	61.68	50.68	26.78	33.35	21.04	15.81	27.13	18.53	24.92
9	30.45	25.38	23.51	27.52	19.44	11.95	22.52	14.55	21.88
10	20.53	17.43	20.3	23.26	11.99	10.56	17.77	14.55	19.7
11	15.45	13.24	19.29	21.71	11.75	8.8	18.94	17.86	17.89
12	12.3	10.7	13.34	14.84	7.45	7.95	13.39	17.03	17.62
13	8.67	7.62	12.82	14.03	8.53	6.04	13.26	18.04	14.57
14	6.68	5.93	11.57	12.47	5.5	5.25	12.35	15.25	11.46

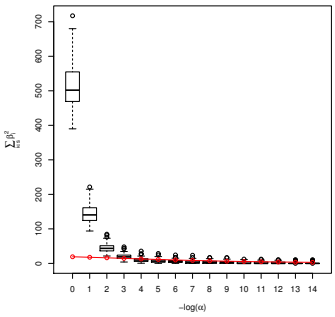
[a] n = 1000, m =4000, p = 40, $\sigma_g^2 = 20$, $\sigma_e^2 = 30$, the marginal β_j and σ_j are estimated through the univariate linear model.



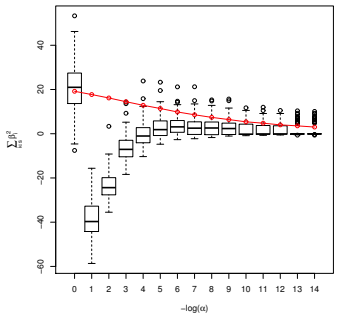
(a) Simple approach



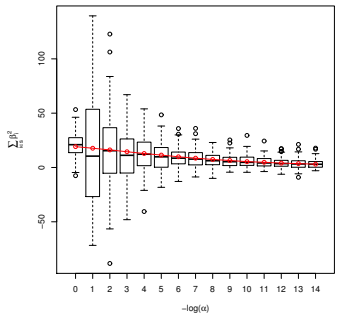
(b) Error approach



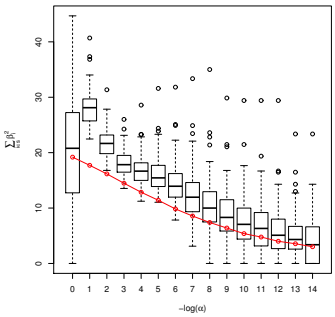
(c) Winner's Curse approach



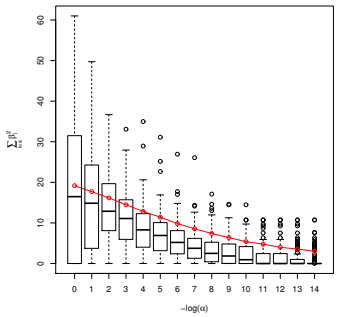
(d) Doubly approach



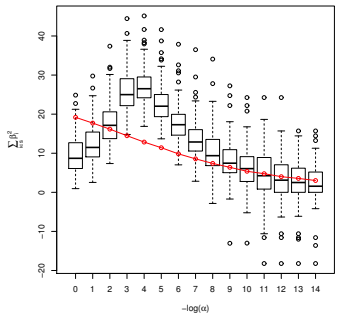
(e) Chen's approach



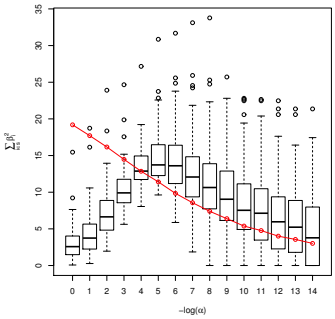
(f) LMM approach



(g) LMM Split approach



(h) scaled Lasso approach



(i) FDE approach

Figure 11: Box plot with varied α for part approaches in Experiment 5

2.4.6 Experiment 6

The following are the results using LDPE as the marginal estimators. As the effects become more varied, LDPE approach does not perform as well as the linear regression when lots of small signals are involved. Furthermore, Chen's approach collapses if LDPE is used in the adjustment. Figure 12 shows that LMM with split work the best under this situation.

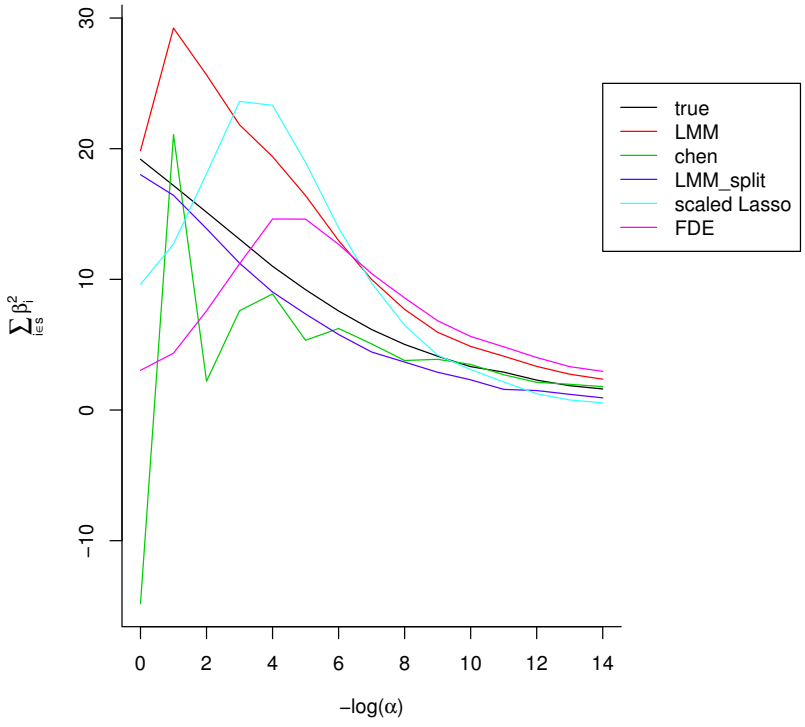
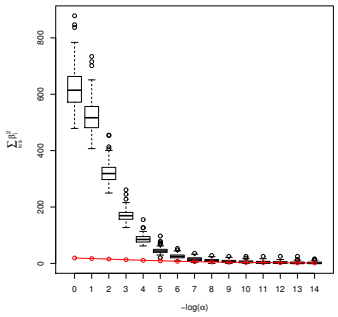


Figure 12: Mean estimation with varied α for part approaches in Experiment 6

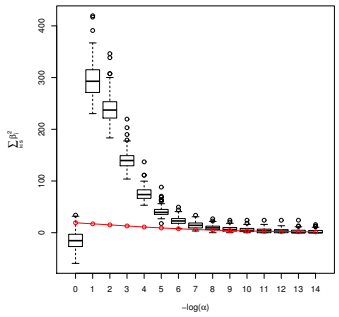
TABLE VI: RESULTS IN EXPERIMENT 6 FOR CURRENT METHODS [a]

$-\log(\alpha)$	simple	error	winner's curse	doubly adjustment	Chen	LMM	LMM-split	scaled Lasso	FDE
0	367055.84	1514.41	367055.84	1514.41	1514.41	99.64	288.85	109.1	273.33
1	258548.27	78378.7	25179.43	5085.17	3872.65	152.05	158.04	35.97	171.81
2	95508.49	51422.54	1328.31	2247.95	2090.46	118.08	67.14	22.69	66.36
3	25424.31	16817.91	58.9	603.21	768.11	85.19	36.32	122.86	10.61
4	5902.37	4314.87	19.83	197.76	315.39	76.85	24.3	160.68	19.86
5	1426.98	1098.72	28.51	85.98	196.98	58.46	19.93	105.23	35.92
6	367.31	291.12	25.1	45.9	69.04	35.84	15.71	50.07	33.55
7	108.33	87.52	23.48	32.89	37.07	21.25	17.53	22.15	29.36
8	38.54	31.59	17.83	22.23	28.14	12.14	12.39	12.36	23.13
9	16.28	13.52	14.51	17.05	15.9	7.95	9.77	12.38	17.88
10	11.34	9.55	12.25	13.97	7.94	6.48	7.55	17.24	14.11
11	7.75	6.6	11.12	12.42	10.47	5.29	9	16.56	12.8
12	5.95	5.11	8.41	9.29	7.32	4.63	7.1	17.46	12.38
13	4.91	4.25	7.12	7.89	5.21	4.14	6.85	16.78	10.15
14	3.55	3.09	5.29	5.82	4	2.84	5.29	16.22	9.32

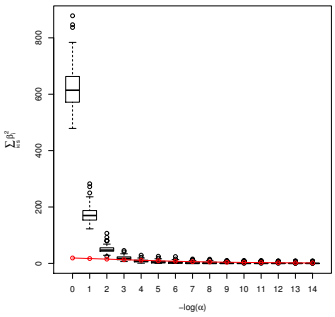
[a] n = 1000, m =4000, p = 40, $\sigma_g^2 = 20$, $\sigma_e^2 = 30$, the marginal β_j and σ_j are estimated through LDPE approaches.



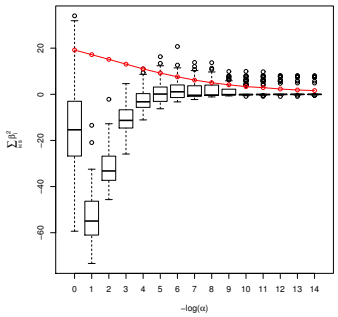
(a) Simple approach



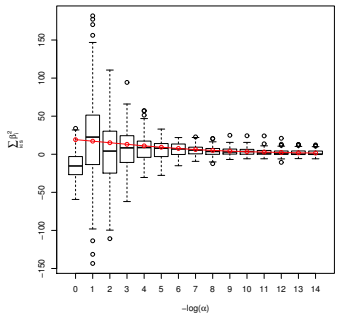
(b) Error approach



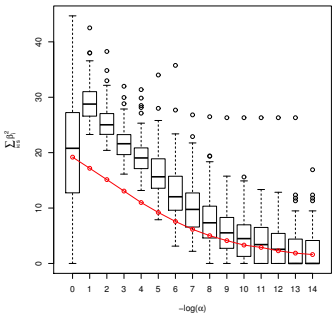
(c) Winner's Curse approach



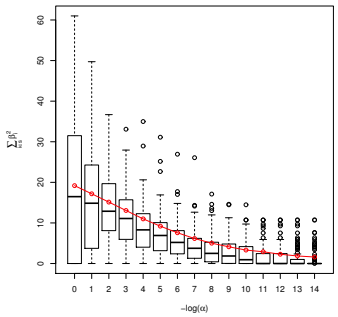
(d) Doubly approach



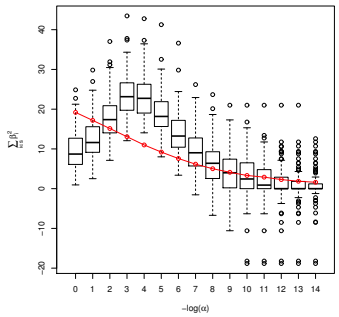
(e) Chen's approach



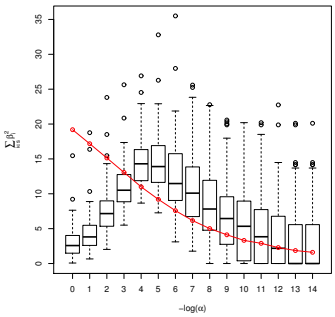
(f) LMM approach



(g) LMM Split approach



(h) scaled Lasso approach



(i) FDE approach

Figure 13: Box plot with varied α for part approaches in Experiment 6

2.4.7 Experiment 7

The following are the results when $\sigma_g^2 = 6$, $\sigma_e^2 = 4$, n = 400, m =4000, and p = 40. Compared with the previous set-up with smaller signals, scaled Lasso and FDE give larger MSE while Chen's approach, LMM, and LMM with split do not change much. In Figure 14, Chen's approach performs well when $-\log(\alpha)$ is large while LMM with split can give a good estimator with small $-\log(\alpha)$s. Both FDE and scaled Lasso approach performs well when $-\log(\alpha)$ is large.

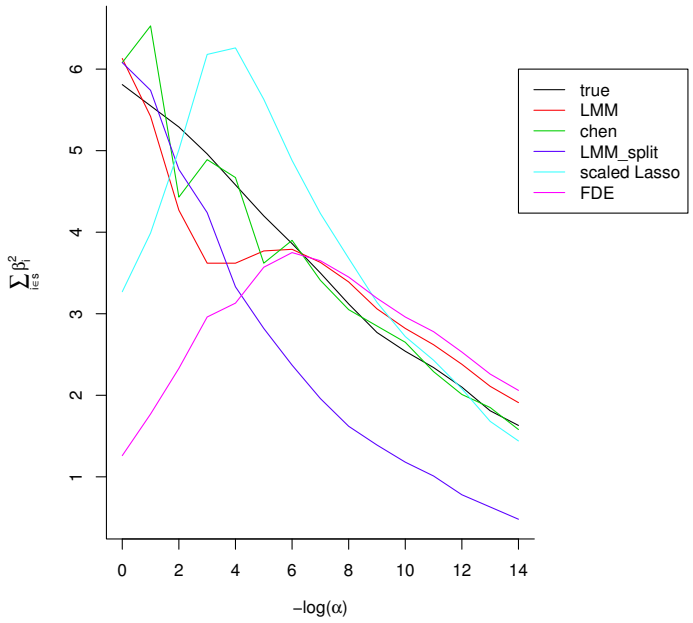
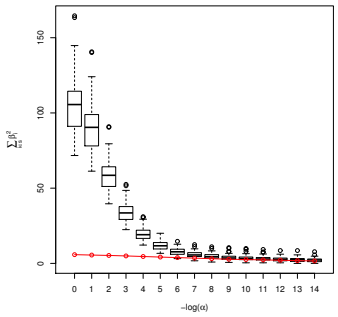


Figure 14: Mean estimation with varied α for part approaches in Experiment 7

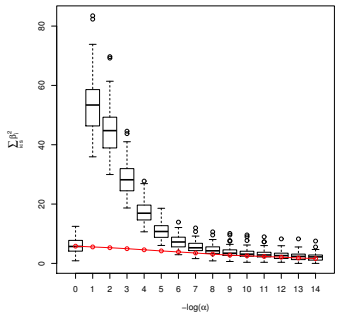
TABLE VII: RESULTS IN EXPERIMENT 7 FOR CURRENT METHODS [a]

$-\log(\alpha)$	simple	error	winner's curse	doubly adjustment	Chen	LMM	LMM-split	scaled Lasso	FDE
0	10164.99	6.72	10164.99	6.72	6.72	5.33	10.19	7.32	21.76
1	7356.33	2352.21	188.96	565.92	113.79	0.62	6.21	3.16	15.32
2	2918.43	1597.8	3.77	163.53	54.78	1.69	4.52	0.66	9.36
3	867.3	576.04	2.71	43.1	24.95	2.5	2.95	1.97	4.41
4	230.57	167.77	3.86	15.33	9.96	1.42	2.92	3.29	2.49
5	62.35	47.43	4.58	9.1	5.94	0.46	3.03	2.54	0.63
6	17.89	13.84	4.5	6.65	3.15	0.21	3.36	1.49	0.25
7	6.9	5.38	4.19	5.46	1.74	0.17	3.53	0.93	0.26
8	3.59	2.84	3.6	4.4	1.48	0.23	3.25	0.71	0.45
9	2.1	1.68	3.09	3.66	1.04	0.24	2.91	0.55	0.61
10	1.52	1.24	3.21	3.66	0.89	0.23	2.78	0.77	0.63
11	1.19	0.98	3.2	3.59	0.8	0.23	2.55	0.83	0.72
12	0.88	0.73	2.96	3.26	0.85	0.23	2.55	0.92	0.79
13	0.75	0.63	2.36	2.59	0.63	0.22	2.16	1.11	0.97
14	0.62	0.52	2	2.19	0.57	0.19	2.04	1.14	0.95

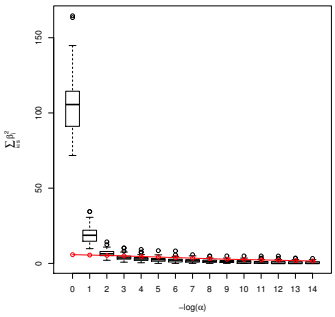
[a] n = 400, m =4000, p = 40, $\sigma_g^2 = 6$, $\sigma_e^2 = 4$, the marginal β_j and σ_j are estimated through the univariate linear model.



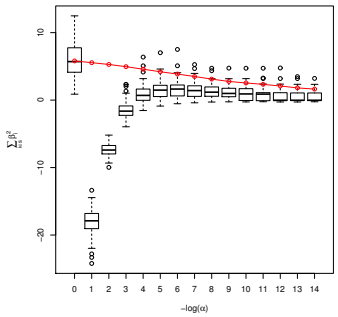
(a) Simple approach



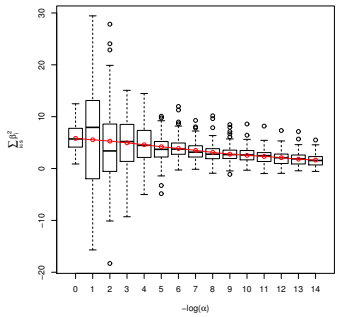
(b) Error approach



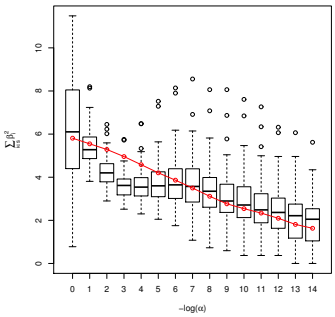
(c) Winner's Curse approach



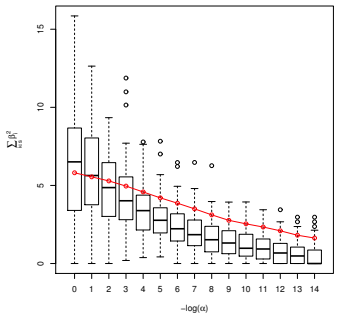
(d) Doubly approach



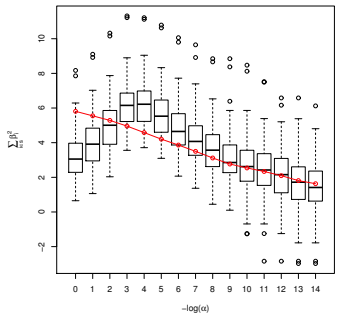
(e) Chen's approach



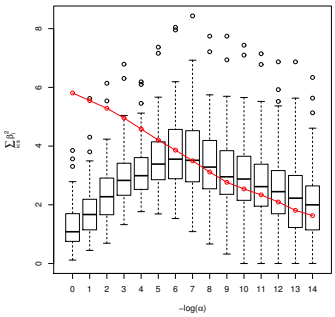
(f) LMM approach



(g) LMM Split approach



(h) scaled Lasso approach



(i) FDE approach

Figure 15: Box plot with varied α for part approaches in Experiment 7

2.4.8 Experiment 8

If we use the LDPE approach to estimate the marginal means when $\sigma_g^2 = 6$ and $\sigma_e^2 = 4$, the influence on the estimation is not as much as that in small signal situations. For those approaches (i.e. LMM, LMM-split, scaled Lasso, and FDE) are independent of the marginal estimators, selection using LDPE gives the similar performance as the linear selection. For those approaches are dependent, LDPE gives larger MSE especially when selection criterion is loose.

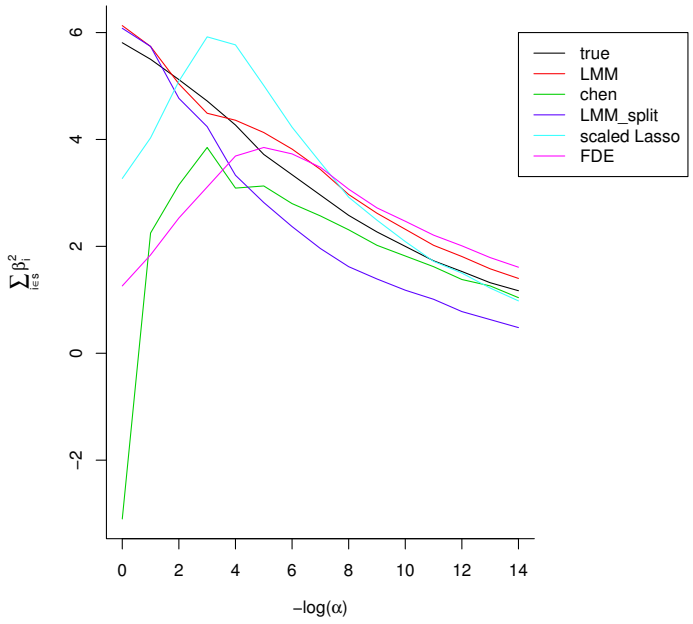
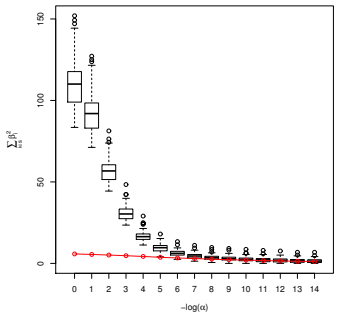


Figure 16: Mean estimation with varied α for part approaches in Experiment 8

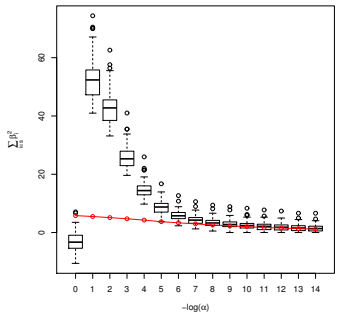
TABLE VIII: RESULTS IN EXPERIMENT 8 FOR CURRENT METHODS [a]

$-\log(\alpha)$	simple	error	winner's curse	doubly adjustment	Chen	LMM	LMM-split	scaled Lasso	FDE
0	11077.43	94.29	11077.43	94.29	94.29	5.33	10.19	7.32	21.76
1	7707.39	2258.18	156.79	781.68	160.04	0.59	6.25	2.85	14.44
2	2742.52	1453.76	1.67	197.83	60.77	0.58	4.42	0.57	7.26
3	700.13	455.16	3.96	47.17	21	0.6	2.6	1.98	2.99
4	159.02	113.08	5.32	17.27	14.58	0.38	2.17	2.68	0.66
5	37.74	28.02	4.9	9.04	5.93	0.37	1.83	2.05	0.23
6	10.19	7.64	4.55	6.44	3.28	0.4	1.85	1.19	0.38
7	3.5	2.65	4.32	5.37	1.84	0.39	2.01	0.8	0.57
8	1.53	1.17	4.05	4.73	1.18	0.31	1.96	0.62	0.6
9	0.98	0.75	3.41	3.91	0.9	0.27	1.82	0.53	0.7
10	0.65	0.51	3.01	3.37	0.81	0.25	1.69	0.51	0.75
11	0.56	0.45	2.27	2.53	0.56	0.22	1.26	0.48	0.87
12	0.48	0.39	1.95	2.16	0.61	0.22	1.34	0.45	0.87
13	0.4	0.34	1.59	1.74	0.49	0.18	1.11	0.48	0.8
14	0.34	0.28	1.41	1.54	0.72	0.16	1.14	0.95	0.86

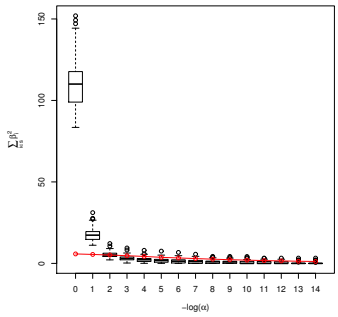
[a] n = 400, m =4000, p = 40, $\sigma_g^2 = 6$, $\sigma_e^2 = 4$, the marginal β_j and σ_j are estimated through LDPE approach.



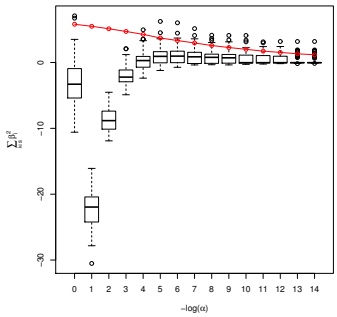
(a) Simple approach



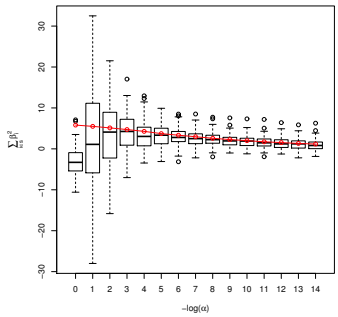
(b) Error approach



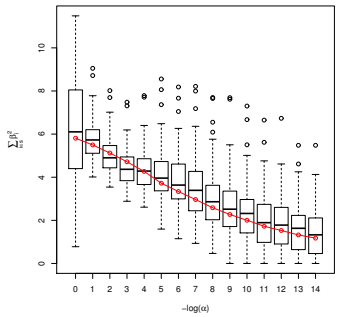
(c) Winner's Curse approach



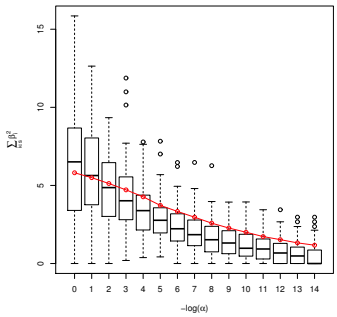
(d) Doubly approach



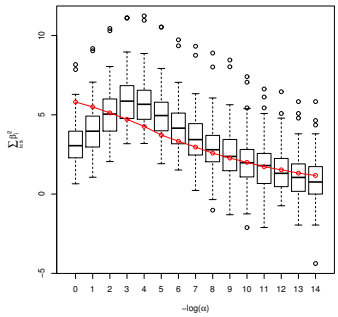
(e) Chen's approach



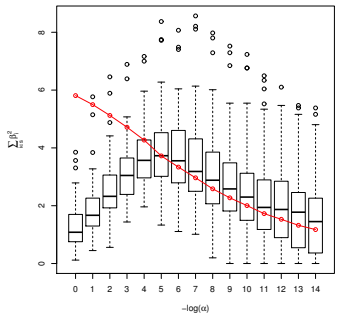
(f) LMM approach



(g) LMM Split approach



(h) scaled Lasso approach



(i) FDE approach

Figure 17: Box plot with varied α for part approaches in Experiment 8

2.4.9 Experiment 9

The first set-up in the second scenario is n = 400, m =600, p = 30, $\sigma_e^2 = 1$, $\tau = 1.5$. $\rho = 0$ in order to compare with the dependence cases later. It can be seen in Table IX that LMM and FDE approaches give the least MSE upon selection. However, in Figure 18, Chen's approach outperforms all the others in the mean estimation. To note that, compared with previous set-ups, FDE and scaled Lasso can follow the selection trend in this scenario. And they are quite similar in the mean figure. Figure 19 shows that Chen's approach covers the truth very well, while the scaled Lasso does not give a very good performance after $-\log(\alpha) \geq 5$.

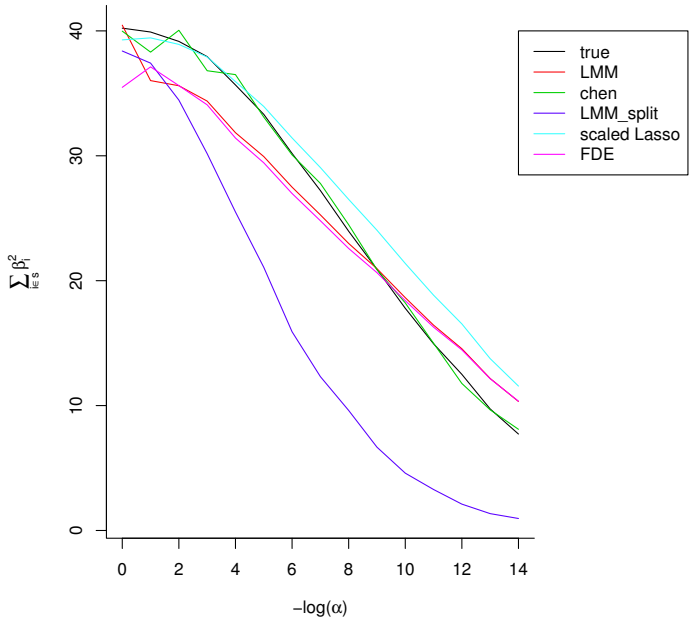
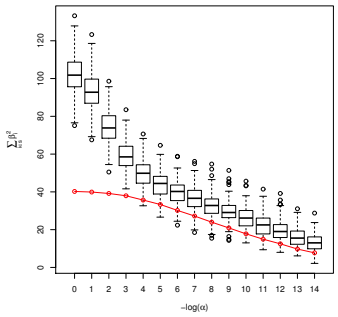


Figure 18: Mean estimation with varied α for part approaches in Experiment 9

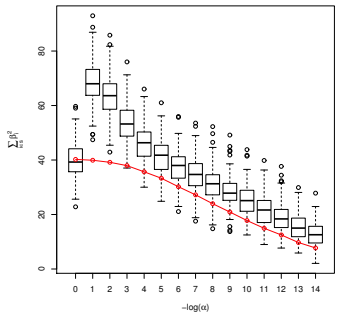
TABLE IX: RESULTS IN EXPERIMENT 9 FOR CURRENT METHODS [a]

$-\log(\alpha)$	simple	error	winner's curse	doubly adjustment	Chen	LMM	LMM-split	scaled Lasso	FDE
0	3989.73	45.09	3989.73	45.09	45.09	8.06	31.37	12.18	23.77
1	2964.71	899.55	315.08	117.13	282.78	18.7	25.75	10.77	10.58
2	1308.26	649.83	51.75	113.31	160.72	16.64	35.24	10.43	16.42
3	518.5	298.21	62.12	133.63	115.29	15.75	76.3	9.94	18.07
4	244.77	151.02	139.61	221.11	57.02	17.56	122.17	9.42	21.18
5	151.55	97.15	199.14	274.31	56.2	13.49	168.87	8.44	17.22
6	120.72	81.49	249.68	317.86	57.34	9.47	218.97	8.99	12.52
7	106.92	74.75	262.11	321.89	34.56	5.94	239.51	9.4	7.98
8	99.04	72.18	255.43	305.3	35.54	3.39	221.12	12.01	4.35
9	90.91	68.83	223.74	262.9	29.12	3.07	218.86	15.38	3.14
10	82.6	64.81	185.39	214.8	23.99	3.39	187.89	17.23	3.04
11	73.27	59.23	151.88	174.07	26.1	4.97	148.4	19.35	4.46
12	62.23	51.42	123.64	140.14	23.18	6.68	118.68	20.21	6.27
13	48.5	41.11	77.66	87.56	17.77	8.37	77.76	19.91	8.2
14	37	31.93	52.38	58.66	12.75	8.92	52.64	17.98	9.01

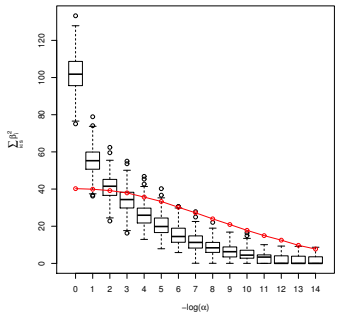
[a] n = 400, m =600, p = 30, $\sigma_e^2 = 1$, $\tau = 1.5$. $\rho = 0$, the marginal β_j and σ_j are estimated through the univariate linear model.



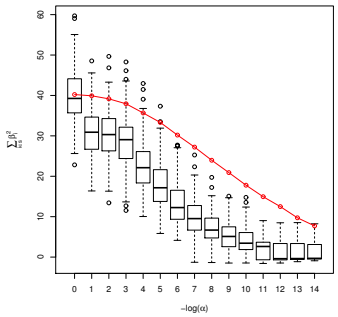
(a) Simple approach



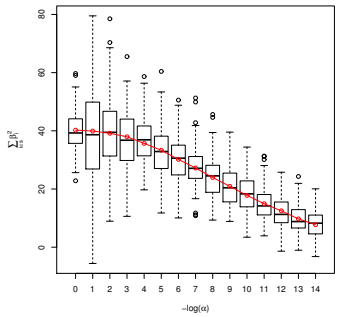
(b) Error approach



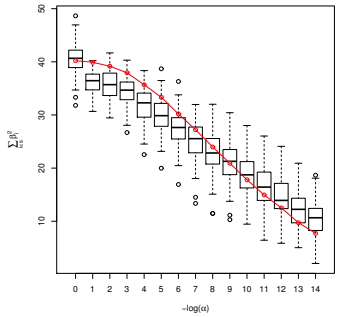
(c) Winner's Curse approach



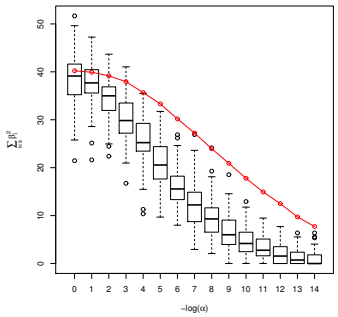
(d) Doubly approach



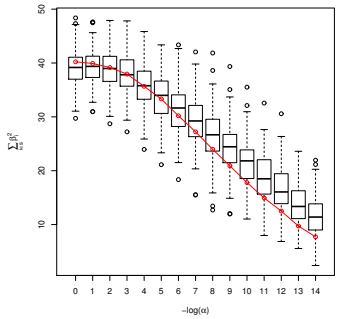
(e) Chen's approach



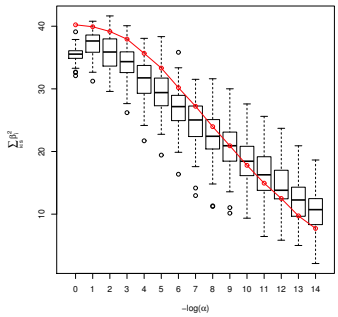
(f) LMM approach



(g) LMM Split approach



(h) scaled Lasso approach



(i) FDE approach

Figure 19: Box plot with varied α for part approaches in Experiment 9

2.4.10 Experiment 10

If the ρ changes from 0 to 0.2, Chen's approach no longer work as the underlying requirement for this approach is independence in X. In fact, LMM, LMM-split, and scaled Lasso are effected by the correlations among X to different degrees. Specifically, scaled Lasso remains high MSE across selections in this case. Although LMM and LMM with split do not performance as bad as others, the MSE suggests their limitation in heritability estimation in correlated data.

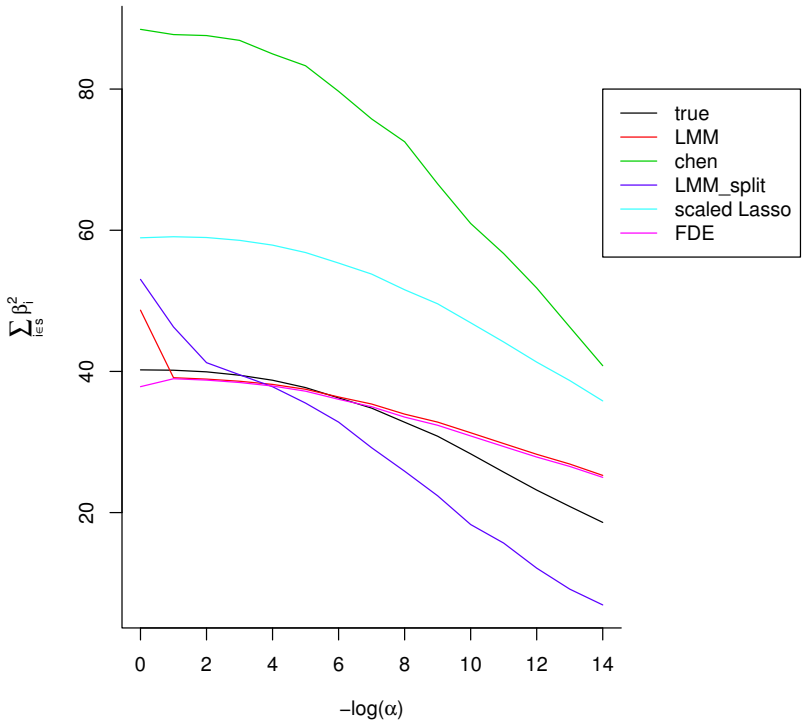
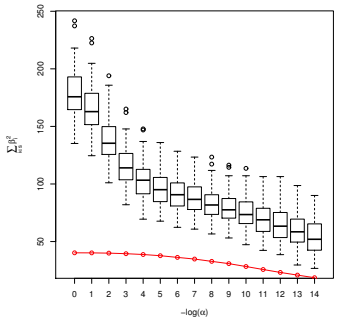


Figure 20: Mean estimation with varied α for part approaches in Experiment 10

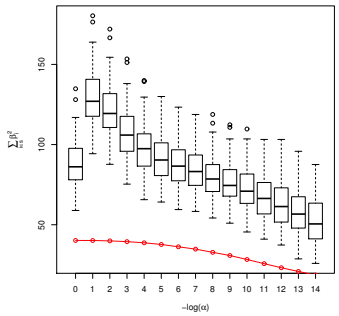
TABLE X: RESULTS IN EXPERIMENT 10 FOR CURRENT METHODS [a]

$-\log(\alpha)$	simple	error	winner's curse	doubly adjustment	Chen	LMM	LMM-split	scaled Lasso	FDE
0	19801.73	2519.43	19801.73	2519.43	2519.43	74.51	181.9	369.15	5.98
1	16243.3	8321.62	5872.13	1674.69	3082.07	2.17	46.34	376.62	2.35
2	9949.6	7008.42	3219.77	1684.54	2724.83	1.9	5.47	381.04	2.19
3	6148.33	4890.74	2397.58	1653.74	2568.96	1.38	5.33	384.05	1.74
4	4373.14	3660.48	1779.32	1349.77	2345.95	1	8	385.22	1.39
5	3559.12	3048.2	1294.08	1010.21	2267.16	0.71	14.66	384.68	0.97
6	3188	2763.99	912.28	710.74	2046.86	1.09	25.22	383.99	1.17
7	2923.08	2556.23	531.42	406.12	1863.39	1.67	49.76	379.71	1.42
8	2707.87	2387.3	304.86	238.84	1740.78	3.18	72.53	372.26	2.45
9	2525.37	2241.06	183.59	158.41	1452.95	6.76	101.8	370.83	5.29
10	2292.07	2048.95	135.49	134.12	1243.04	12.47	134.42	363.39	10.05
11	2054.3	1849.17	123.62	138.65	1107.17	20.97	128.11	358.51	17.71
12	1816.86	1645.65	126.32	150.44	977.79	31.17	151.2	347.76	27.4
13	1588.97	1447	129.88	156.01	789.89	43.42	167.87	337.89	39.43
14	1354.97	1239.85	135.44	162.9	620	52.93	166.92	316.45	49.3

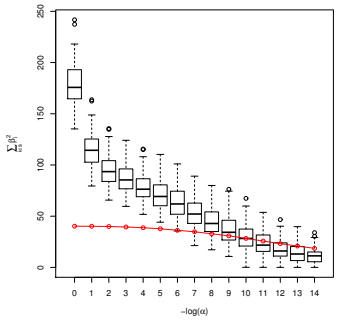
[a] n = 400, m =600, p = 30, $\sigma_e^2 = 1$, $\tau = 1.5$. $\rho = 0.2$, the marginal β_j and σ_j are estimated through the univariate linear model.



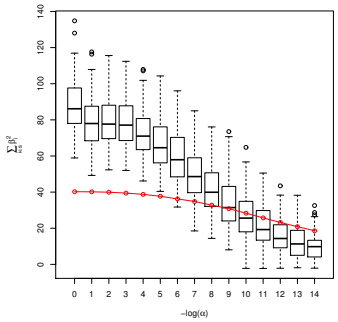
(a) Simple approach



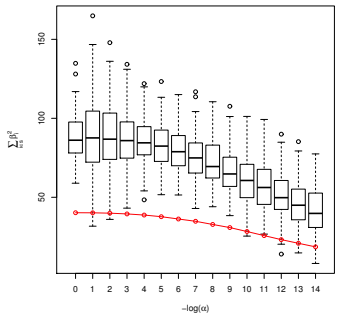
(b) Error approach



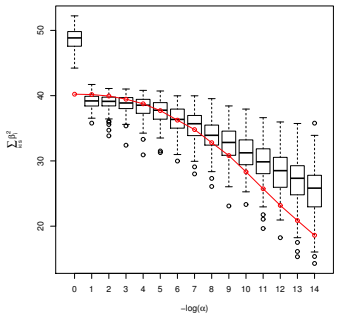
(c) Winner's Curse approach



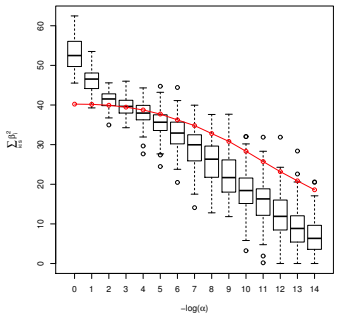
(d) Doubly approach



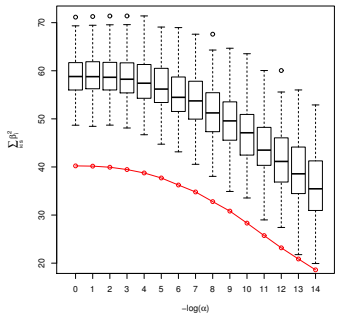
(e) Chen's approach



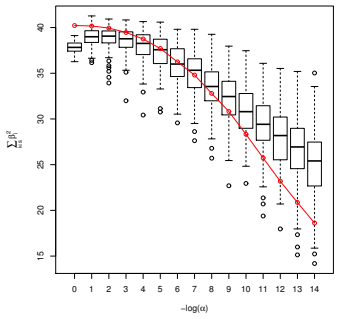
(f) LMM approach



(g) LMM Split approach



(h) scaled Lasso approach



(i) FDE approach

Figure 21: Box plot with varied α for part approaches in Experiment 10

2.5 Summary and Comments

In this chapter, a number of possible solutions to the statistical problem of estimating numerous weak effects are considered and extensive simulation studies are conducted to demonstrate the performance of those approaches under different circumstance. The simple linear marginal regression and the LDPE approach are used for the variable selection in the simulation studies. When there are many weak signals in the data, the LDPE is more biased than the linear regression estimator. As a result, for those approaches that adjust the bias of the individual estimators based on the LDPE, both the bias and the MSE in the total variation estimator are in general larger. For the direct-estimation approaches, the bias in the LDPE appears to not have as much negative influence on the estimator. The explained variation estimator based on the LDPE in the latter case can sometimes have a slightly smaller MSE than the selection based on the marginal regression. When the covariates are correlated, the marginal linear regression induces more bias in the variation estimator.

Most individual-bias-adjustment approaches do not perform well, especially when $-\log(\alpha)$ is small. This is due to the cumulation of errors when the number of selected coefficients are large. Among these methods, Chen's approach gives the best estimator in respect to the bias. One issue with Chen's estimator is the high variability of the estimator in comparison with the direct-estimation approaches.

As discussed in the previous chapter, performance of Lasso-related approach is restricted by the sparsity assumption and the uniform strength assumption. Namely, to achieve the optimal prediction error in Lasso-related approaches, $p \log m \ll \sqrt{n}$ and all the non-zero βs should

be greater than $C\sigma_e\sqrt{2/nlogm}$, where C is some constant greater than $\frac{1}{2}$ (Zhang and Zhang, 2014). Since our simulation set-ups violate the sparsity assumption and the non-zero βs do not satisfy the uniform strength assumption, the scaled Lasso and the FDE approaches are more biased than the LMM approach when the selection criterion is loose. Such discrepancy appears when the selection criterion becomes stricter. However, if we check the MSE of the estimators, the scaled Lasso and FDE approach seem to perform well. This is due to the fact that these two estimators are subject to smaller variation compared with the LMM-related approaches. We also notice that when the covariates are correlated, the FDE approach has a relatively small MSE and the smallest bias among all the approaches.

The LMM approach has the smallest MSE in most simulation set-ups. However, due to the Winner's Curse, the estimator is subject to large bias when the selection criterion is less strict. The LMM approach with sample split can reduce the bias. However, due to the reduction in sample size in both variable selection and variation estimation procedures, the estimator is subject to higher variation when compared with the LMM approach without sampling split. Furthermore, when the selection criterion is strict, LMM approach with split samples seem to under-estimate the true signal strength compared with the LMM approach.

In summary, for researchers planning to estimate the partial heritability, different approaches should be used in different situations. For the marginal estimators, as long as the sparsity assumption holds, the LDPE can provide a good reference for selecting the covariates. On the other hand, if the partial variation estimating approaches involve combining the marginal estimators, the simple linear regression is more appropriate. For a loose selection criterion (e.g.

$-\log(\alpha)$ is small),when the sample size is large, we should use the LMM approach with sample split if the loss of power is not of a major concern. If the sparsity assumption is reasonable, the bias of the scaled Lasso approach is small. When the covariates are correlated, we should use the FDE approach as other approaches no longer work well under this situation. If the variable selection criterion is strict, the LMM approach, Chen's approach, the scaled Lasso, and the FDE approach all can yield good estimates with small MSEs.

CHAPTER 3

ESTIMATING TOTAL EFFECT OF SELECTED VARIABLES: THE CASE OF INDEPENDENT COVARIATES

3.1 The LMM with Sample Split

The simulation results in the previous chapter do not suggest a single approach that outperforms other approaches in all the simulated cases. However, the LMM of Yang et al. provides a robust choice among the approaches. This approach is further studied in this chapter.

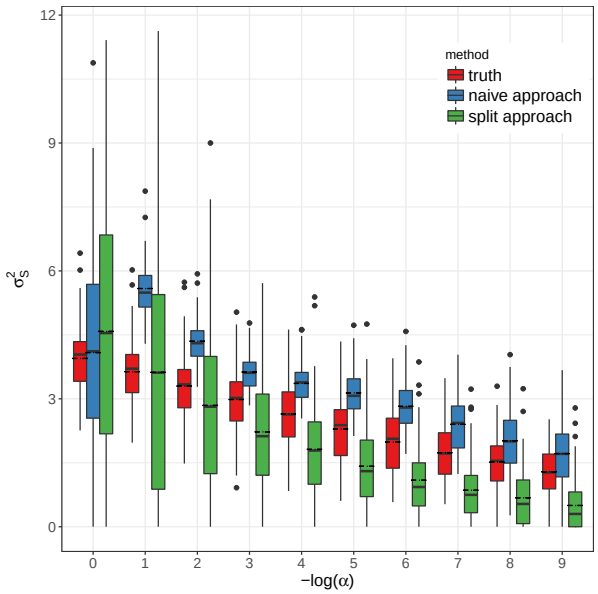


Figure 22: Demonstration of Yang et al.'s approach under variable selection

A simulation study is conducted to compare the naive approach, the sample split approach, and the truth. The data are simulated using steps 1-3 of the algorithm in the simulation section of this chapter with parameters $(n, m, p, \sigma_g^2, \sigma_e^2) = (400, 4000, 40, 4, 6)$. The simulation results with α values varying from $\exp(0)$ to $\exp(-9)$ are shown in Figure 22.

It can be seen from the graph that, the naive approach yields an estimator with a large upward bias when $\alpha < 1$. The bias reduces as α becomes very small. In contrast, the split sample approach has a relatively small downward bias and a large variation. The simulation result suggests that the naive approach can be unsatisfactory and the sample split approach is better. We propose improvements in the following to reduce the bias and the variation of the sample split approach.

3.2 The Proposed Subsampling Approach

Suppose that we have a variable selection approach with tuning parameter λ in place for selecting significant variables to be included in the linear model building. Whether variable x_j is selected depends not only on the underlying signal strength β_j, but also on the tuning parameter values and the sample size. When a subsample is used for selection, the sample size is reduced. As a result, a smaller number of variables are selected if the same selection criterion is applied. To correct the bias due to the sample size reduction, we artificially decrease the tuning parameter value so that the total signal selected remains at the level of the full sample

size. Specifically, for the selection through individual tests, the rejection rule for a given α level is modified to

$$|\frac{\hat{\beta}_{js}}{\hat{\sigma}_{ejs}}| \geq z_{1-\alpha/2}\sqrt{q}, \tag{3.1}$$

where $\hat{\beta}_{js}$ is the parameter estimator for β_j under the subsample, $\hat{\sigma}_{ejs}$ is the variance estimate for $\hat{\beta}_{js}$, and $q \in (0,1)$ is the fraction of the full sample used for variable selection. The rationale for the adjustment is that both $nq\hat{\sigma}^2_{js}$ and $n\hat{\sigma}^2_j$ estimate the same quantity.

To reduce the variability of the split sample estimator, we propose to randomly take subsamples of size $[nq]$ from the observed data for variable selection and use the remaining samples of size $[(1-q)n]$ to estimate the variation captured by the selected covariates using Yang et al's approach. The sample split corresponds to $q = 0.5$. Let $\hat{\sigma}^2_1(q,\alpha), \cdots, \hat{\sigma}^2_T(q,\alpha)$ denote respectively the estimators of the variations captured by the selected covariates for the T random samples drawn from the observed data, where q and α are respectively the proportion of sample used for selection and the test significance level for the selection using the decision rule in Equation 3.1. Let

$$\tilde{\sigma}^2(q,\alpha) = \frac{1}{T}\sum_{t=1}^{T}\hat{\sigma}^2_t(q,\alpha). \tag{3.2}$$

For each $q \in (0,1)$, $\tilde{\sigma}^2(q,\alpha)$ is an estimator of σ_λ^2 in Equation 3.3, where $\lambda = \alpha$.

$$\sigma_\lambda^2 = E\left(\sum_{j \in S_\lambda} \beta_j^2\right). \tag{3.3}$$

To estimate the variance of $\tilde{\sigma}^2(q,\alpha)$, note that subsamples are not independent of each other. The dependence of two subsamples depends on the overlapped individuals. Suppose that the statistic for estimating $\hat{\sigma}^2(q,\alpha)$ can be asymptotically expressed linearly in data from individuals, i.e.,

$$\hat{\sigma}^2(q,\alpha) = \sum_{j=1}^{[n(1-q)]} h(Y_j, X_j),$$

where (Y_j, X_j) is the data from subject j. For another subsample of the same size,

$$\hat{\sigma}_i^2(q,\alpha) = \sum_{j=1}^{[n(1-q)]} h(Y_{i_j}, X_{i_j}).$$

The expected correlation between the statistic from the two subsamples is

$$\rho = \frac{\sum_{l=\max(0,n-2k)}^{[n(1-q)]} l \frac{C(n-k,l)C(k,n-k-l)}{C(n,k)} \sigma^2}{[n(1-q)]\sigma^2} \tag{3.4}$$

where $k = nq$, $C(n,k)$ denotes the combinatory number of taking k from n, and $\sigma^2 = \text{var}\{h(Y_i, X_i)\}$, and l denote the number of overlapped subjects between two random samples.

In Equation 3.4, the numerator can be written as

$$\begin{aligned}
&\frac{(n-k)\sigma^2}{C(n,k)} \sum_{l=\max(0,n-2k)}^{[n(1-q)]} C(n-k-1,l-1)C(k,n-k-l) \\
&= \frac{C(n-1,n-k-1)(n-k)\sigma^2}{C(n,k)} \\
&= \frac{(n-k)^2}{n}\sigma^2
\end{aligned}$$

by binomial theorem. We have

$$\rho = \frac{\frac{(n-nq)^2}{n}\sigma^2}{n(1-q)\sigma^2} = 1-q$$

It can be seen that the variance for $\tilde{\sigma}^2(q,\alpha)$ is approximately

$$\mathrm{Var}\{\tilde{\sigma}^2(q,\alpha)\} = \frac{1}{T}\{1+(T-1)(1-q)\}\sigma^2.$$

Following the derivation, we propose to estimate the variance of $\tilde{\sigma}^2(q,\alpha)$ by

$$v_\sigma(p,\alpha) = \frac{1+(T-1)(1-q)}{T(T-1)} \sum_{t=1}^{T} \left\{\hat{\sigma}_t^2(q,\alpha) - \tilde{\sigma}^2(q,\alpha)\right\}^2.$$

While the methods with different subsampling ratio are all expected to yield estimators with small bias for the variation explained by the selected covariates with the full sample, the variances of the estimators are expected to depend on the sample splitting ratio q. In practice, we need to choose an appropriate q such that the variance of $\tilde{\sigma}^2(q,\alpha)$ is as small as possible. For a fixed set of covariates, a larger sample size for the estimation of the explained variation

using Yang et al's approach yields an estimator with smaller variance. This means we want to have q as small as possible. On the other hand, a smaller sample size used in the variable selection increases the variability of $\tilde{\sigma}^2(q, \alpha)$ because of the uncertainty in the set of selected covariates. We concentrate on four q values: $q = 0.2$, $q = 0.5$, $q = 1 - 1/e$, and $q = 0.8$. The third one, termed bootstrap subsampling, corresponds to the average sample size of a bootstrap sample of size n with repetitions removed. The following algorithm is used to implement the bootstrap subsampling approach:

1. Select n subjects with replacement from the original data, remove repetitions in the sample and denote the resulting sample by A. Let B denote the sample with subjects not selected.
2. For a given α, perform individual tests using sample A by decision rule (Equation 3.1). Denote the selected covariates $x_S = (x_j, j \in S)$.
3. Fit a linear mixed model to sample B with only x_S in the model. Estimate $\hat{\sigma}_t^2(\alpha)$
4. Repeat the steps for $t = 1, \cdots, T$.

The algorithm can be easily adapted to the subsampling approach with a fixed sampling rate.

For the problem of estimating the total variation explained by all the covariates, Jiang et al. (2016) showed that Yang et al's approach yields a consistent estimator under some reasonable assumptions. We adapt their results here for the estimation of the variation explained by a set of selected variables. The following assumptions are made in studying the asymptotic behavior of $\tilde{\sigma}^2(q, \alpha)$ using the hard-thresholding based on the individual tests.

1. X is a random matrix whose entries are independent with mean 0 and variance 1. $(\beta_1, \cdots, \beta_p)$ are independent $N(0, \sigma_g^2/p)$ and $\beta_j = 0, j = p+1, \cdots, m$. β_0 is a fixed number and $\varepsilon \sim N(0, \sigma_\epsilon^2 I_n)$ and independent of X.
2. As $n \to \infty$, $p/m \to c, n/m \to d$, where $c, d \in (0, 1]$.
3. For given $\alpha \in (0,1)$ and $q \in (0,1)$, a random samples of size $[nq]$ from the observed data is drawn and used in selecting variables by the method (Equation 3.1) and the remaining data are used for estimating the variance σ_α^2. Let the estimators be $\hat{\sigma}_t^2(q,\alpha), t = 1, \cdots, T$.

Proposition 1. Under assumptions 1-3, $\hat{\sigma}_t^2(q,\alpha) - \sigma^{*2}(q,\alpha) \to 0$ in probability as $n \to \infty$ for fixed $q \in (0,1)$ and $\alpha \in (0,1)$ and $t = 1, \cdots, T$ where

$$\sigma^{*2}(q,\alpha) = \sum_{j=1}^{p} E\left[\beta_j^2 1_{\{|\sqrt{[nq]}\beta_j/\sigma_{ej} + Z| > \sqrt{q} z_{1-\alpha/2}\}}\right],$$

and Z is a standard normal random variable independent of β_j. As a result,

$$\tilde{\sigma}^2(q,\alpha) - \sigma^{*2}(q,\alpha) \to 0,$$

in probability as $n \to \infty$.

Proof: Let S be a given subset of indices $\{1, \cdots, m\}$. Then

$$y = \beta_0 + \sum_{j \in S} \beta_j x_j + \eta, \tag{3.5}$$

$$\eta = \sum_{j \in S^c} \beta_j x_j + \epsilon, \tag{3.6}$$

where $S^c = \{1, \cdots, m\} - S$, and $\eta \sim N(0, \sigma_E^2)$ with $\sigma_E^2 = \sigma_e^2 + \sum_{j \in S^c \cap \{1, \cdots, p\}} \sigma_g^2 / p$. It can be seen that model (Equation 3.5) satisfies all the conditions required in Jiang et al. (2016) for estimating

$$\sigma_S^2 = E\left(\sum_{j \in S} \beta_j^2\right).$$

The estimator is thus consistent in probability. When S is the set of variables selected from the sample with size $[nq]$, all the arguments hold conditional on the variables being selected. That is,

$$P\left\{|\hat{\sigma}_t^2(q, \alpha) - \sigma_\alpha^{**2}| > \epsilon \mid S_{(q,\alpha)}\right\} \to 0,$$

where $\sigma_\alpha^{**2} = E\left(\sum_{j \in S(q,\alpha)} \beta_j^2\right)$ and

$$S(q, \alpha) = \left\{ j \middle| \frac{|\beta_j \sum_{i=1}^{[nq]} x_{ij}^2 + \sum_{i=1}^{[nq]} x_{ij} e_i|}{\sqrt{\sum_{i=1}^{[nq]} (e_i - \frac{1}{[nq]} \sum_{i=1}^{[nq]} x_{ij} e_i)^2}} > \sqrt{q} z_{1-\alpha/2} \right\}$$

Since the two subsets of data are independent, the conditional probability is the same as the unconditional probability. Since

$$\begin{aligned} \frac{1}{[nq]} \sum_{i=1}^{[nq]} \left(e_i - \frac{1}{[nq]} \sum_{i=1}^{[nq]} x_{ij} e_i\right)^2 &\xrightarrow{P} \sigma_{ej}^2, \\ \frac{\sum_{i=1}^{[nq]} x_{ij} e_i}{\sqrt{\sum_{i=1}^{[nq]} (e_i - \frac{1}{[nq]} \sum_{i=1}^{[nq]} x_{ij} e_i)^2}} &\xrightarrow{W} Z, \end{aligned}$$

uniformly over j, it follows that $\sigma^{**2}(q, \alpha) - \sigma^{*2}(q, \alpha) \to 0$, for any fixed $q, \alpha \in (0, 1)$.

If the full sample is used to select the variables, the variation of the outcome explained by the selected covariates is

$$\sigma_\alpha^2 = E\left(\sum_{j\in S_\alpha}\beta_j^2\right) = \sum_{j=1}^{p} E\left[\beta_j^2 1_{\left\{\frac{|\beta_j\sum_{i=1}^n x_{ij}^2+\sum_{i=1}^n x_{ij}e_i|}{\sqrt{\sum_{i=1}^n(e_i-\frac{1}{n}\sum_{i=1}^n x_{ij}e_i)^2}}>z_{1-\alpha/2}\right\}}\right],$$

which is asymptotically equivalent to

$$\sum_{j=1}^{p} E\left[\beta_j^2 1_{\{|\sqrt{n}\beta_j/\sigma_{ej}+Z|>z_{1-\alpha/2}\}}\right].$$

The bias of the subsampling estimator is therefore

$$\sum_{j=1}^{p} E\left[\beta_j^2\left(1_{\{|\sqrt{[nq]}\beta_j/\sigma_{ej}+Z|>\sqrt{q}z_{1-\alpha/2}\}} - 1_{\{|\sqrt{n}\beta_j/\sigma_{ej}+Z|>z_{1-\alpha/2}\}}\right)\right],$$

which can approximately be rewritten as

$$\sum_{j=1}^{p} E\Bigg[\beta_j^2\Big\{\Phi\{\sqrt{q}(-\sqrt{n}\beta_j/\sigma_{ej}-z_{1-\alpha/2})\}+\Phi\{\sqrt{q}(\sqrt{n}\beta_j/\sigma_{ej}-z_{1-\alpha/2})\}$$
$$-\Phi(-\sqrt{n}\beta_j/\sigma_{ej}-z_{1-\alpha/2})-\Phi(\sqrt{n}\beta_j/\sigma_{ej}-z_{1-\alpha/2})\Big\}\Bigg].$$

From the bias expression, it can be seen that when the signal-to-noise ratio is small, the cutoff value should be less adjusted. On the other hand, if the signal-to-noise ratio is large, more adjustment my be considered. Based on this observation, we also proposed a weighted estimator. When the signal is small, more weight will be put on the unadjusted estimator, and

vice versa. Since we do not know the truth signal strength in practice, we use an empirical weight based on the ratio between variation estimation without variable selection and with variable selection at $\alpha = 0.01$. The weighted estimator is as follows,

$$\sigma_w^2(\alpha, q) = \sigma_a^2(\alpha, q) + \Big(\sigma_u^2(\alpha, q) - \sigma_a^2(\alpha, q)\Big)e^{-c\delta} \tag{3.7}$$

where $c = 10$, and $\delta = \hat{\sigma}_u^2(0.01, q)/\hat{\sigma}_u^2(1, q)$

3.3 Simulation Study and Method Comparison

In this section, we perform simulation studies to compare the proposed estimators for σ_λ^2 with estimators that have good performance without variable selection. Those estimators include the naive approach, subsampling approaches with different sampling ratios, the bootstrap subsampling estimator, and the EigenPrism estimators with and without subsampling. The EigenPrism is an approach to estimating the total variation attributable to all covariates, recently proposed by Janson et al. (2017). Compared with Yang et al.'s approach, EigenPrism does not impose restrictions on the coefficients. Instead, it assumes a normally distributed covariates as in Dicker (2014). Like the LMM, this method does not require sparsity or the information on the noise level in estimating the total variation explained by all the covariates. EigenPrism also provides a variance estimate so that inference can be performed on the total variation explained, often in terms of confidence intervals. We compare this method with other methdos in the simulation to see how well it performs in comparison to Yang et al.'s approach to estimating the variation explained by a set of non-randomly selected covariates.

The basic idea of the EigenPrism approach is as follows. Note that, by the singular value decomposition, we have $X = UDV'$ where U is a $n \times n$ orthonormal matrix, D is a $n \times n$ diagonal with non-increasing and non-negative entries in diagonal, and V is $m \times n$ orthonormal matrix. Let $z = U'y$, then z becomes a $n \times 1$ vector. For the simple linear model in Equation 2.1 under the assumption that $X \sim N(0, I)$ and $\epsilon \sim N(0, \sigma_e^2)$, it can be seen that

$$E(z_i^2|d) = d_i^2\sigma_g^2/m + \sigma_e^2 \tag{3.8}$$

where z_i is the ith element in z, and d_i is the ith diagonal element in D. Let $w \in R^n$ be a $n \times 1$ vector of non-negative weights, conditional on the matrix D, the expectation of $\sim_{i=1}^n w_i z_i^2$ conditional on D can be written as

$$E(\sum_{i=1}^{n} w_i z_i^2|D) = \sigma_g^2 \sum_{i=1}^{n} w_i d_i^2/m + \sigma_e^2 \sum_{i=1}^{n} w_i \tag{3.9}$$

Therefore, by finding the set of w satisfying $\sum_{i=1}^n w_i = 0$ and $\sum_{i=1}^n w_i d_i^2/m = 1$ will make $\sum_{i=1}^n w_i z_i^2$ an unbiased estimator of σ_g^2. Since the upper bound of $var(\sum_{i=1}^n w_i z_i^2|D)$ is $2(\sigma_g^2 + \sigma_e^2)^2 \max(\sum_{i=1}^n w_i, \sum_{i=1}^n w_i^2(d_i^2/m)^2)$, it can be formed as convex optimization to minimize the upper bound of the $var(\sum_{i=1}^n w_i z_i^2|D)$ subject to the $\sum_{i=1}^n w_i = 0$ and $\sum_{i=1}^n w_i d_i^2/m = 1$. Therefore, with unbiased estimator of σ_g^2, EigenPrism can also yield an estimator of $\hat{\sigma}_g^2$ variance. This method can be easily adapted to cases with selected covariates or subsampling approach. However, when variables are non-randomly selected, bias can be introduced into the variation

estimator. It is of interest to see how well the EigenPrism performs in comparison to the Yang et al's approach with non-randomly selected covariates.

The simulated data are generated in the following way. The covariates are the bi-allelic genetic markers taking values $0, 1, 2$. The simulated outcome is a quantitative trait. Specifically,

1. $\mathbf{X}$: Generate a random number π following $Uniform(0.05, 0.5)$ distribution. Generate n independent random numbers, each follows $Binomial(2, \pi)$ distribution. Standardized to have mean 0 and variance 1. Repeat m times to generate the covariate matrix.
2. $\boldsymbol{\beta}$: Generate $\beta_j, j = 1, \cdots, p \overset{iid}{\sim} N(0, \sigma_g^2/p)$, and set $\beta_j = 0, j = p+1, \cdots, m$.
3. $\boldsymbol{Y}$: Generate ϵ following $N(0, \sigma_e^2 \mathbf{I})$ and set $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$.
4. For $q \in \{0.2, 0.5, 0.8\}$ randomly sample from the observed data $[nq]$ subjects without replacement. Denote the select sample as A and the rest of sample as B. Select variables using sample A with $-\log \alpha \in \{1, \cdots, 9\}$ by (Equation 3.1).
5. For selected variables in the previous step, using sample B to estimate the attributable variation of Y explained by the selected covariates by Yang et al's approach.
6. Repeat the previous two steps 100 times to obtain the mean and variance of the estimator.

We replicate the algorithm N times and compute the mean squared error (MSE) to measure the performance of the estimator,

$$MSE = \frac{1}{N} \sum_{l=1}^{N} (\hat{\sigma}_{l\alpha}^2 - \sigma_\alpha^2)^2. \tag{3.10}$$

In the following simulation studies, N is set to 100. We specify parameters $(n, m, p, \sigma_g^2, \sigma_e^2)$ for each setting taking the sparsity of the non-zero SNP effects into consideration.

3.3.1 Experiment 1

The parameters were set as follow: $n = 400, m = 4000, p = 40, \sigma_g^2 = 4, \sigma_e^2 = 6$. The simulation results are shown in Table XI-XII and Figure 23. From Figure 23, the subsampling approaches have small bias while the naive approach is subject to large bias. From Table XI, all the subsampling approaches outperform the naive approach with respect to the MSE for most α values. The bootstrap subsampling approach and the subsampling with $q = 0.5, 0.8$ give slightly smaller MSE compared with subsampling with $q = 0.2$. Weighted approach yields the smallest MSE compared with other subsampling approaches. EigenPrism approach works well when no selection is conducted. With variable selection, EigenPrism yields higher MSE compared with compared with Yang et al.'s approach. If we incorporate the EigenPrism with subsampling approach, the MSE is reduced compared with using EigenPrism directly. In Table XII, the 90% confidence intervals were computed through normal approximation using the proposed variance estimate for our subsampling approach. In comparison, I also obtain the 90% confidence intervals for the EigenPrism approach with and without subsampling using the approach proposed in Janson et al. (2017). The confidence intervals and their length were obtained. It can be seen that the confidence interval of EigenPrism does not cover the truth well when variables are selected. With subsampling approach, Yang et al.'s approach yields confidence intervals with better coverage and shorter length. EigenPrism approach with subsampling seems to cover

the truth well, however, the confidence intervals are much wider compared with Yang et al.'s approach using subsampling.

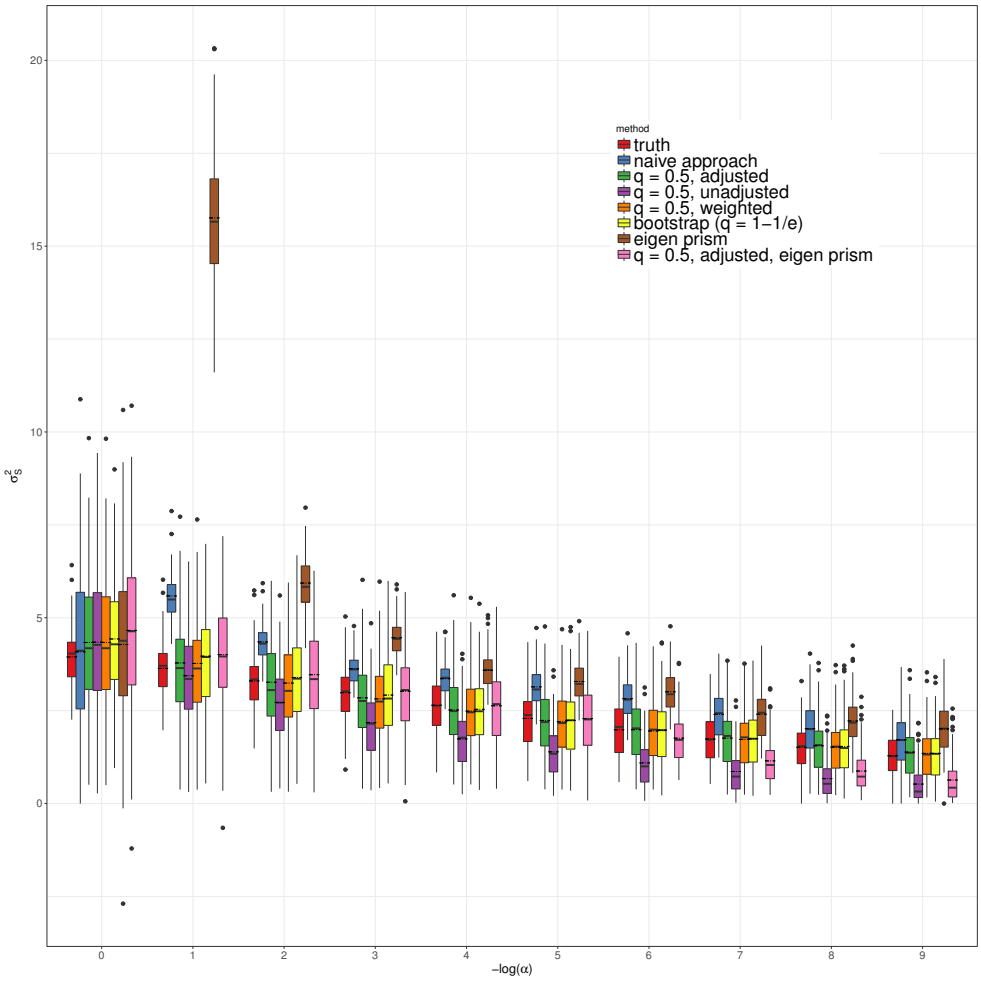


Figure 23: Box plot of subsampling approach with varied α in Experiment 1

TABLE XI: MSE IN EXPERIMENT 1 FOR SUBSAMPLING APPROACH [a]

	Yang et al.'s approach							EigenPrism	
$-\log(\alpha)$	Naive	$q = 0.2$	$q = 0.5$, adjusted	$q = 0.5$, unadjusted	$q = 0.5$, weighted	bootstrap	$q = 0.8$	Naive	$q = 0.5$
0	4.92	4.5	3.1	3.06	3.09	2.33	1.66	4.86	5.1
1	4.15	2.35	2.05	1.75	2.02	2.24	2.41	148.67	2.58
2	1.45	1.62	1.56	1.24	1.53	1.67	1.81	7.42	1.79
3	0.78	1.06	0.97	1.15	0.95	0.97	1.06	2.55	0.92
4	0.83	0.8	0.67	1.19	0.65	0.6	0.67	1.21	0.56
5	0.95	0.58	0.44	1.1	0.43	0.43	0.39	1.22	0.41
6	0.89	0.52	0.36	1.08	0.35	0.34	0.31	1.14	0.33
7	0.58	0.38	0.26	1.02	0.25	0.23	0.23	0.71	0.60
8	0.37	0.32	0.17	0.91	0.17	0.16	0.16	0.52	0.61
9	0.3	0.27	0.16	0.8	0.15	0.14	0.13	0.47	0.65

[a] The Experiment 1 is set up as n = 400, m =4000, p = 40, $\sigma_g^2 = 4$, $\sigma_e^2 = 6$

TABLE XII: 90% CONFIDENCE INTERVAL BY SUBSAMPLING APPROACH IN EXPERIMENT 1

$-\log(\alpha)$	EigenPrism			Subsampling Approach using LMM			Subsampling Approach using EigenPrism		
	coverage [a]	90% CI [b]	length [c]	coverage	90% CI	length	coverage	90% CI	length
0	0.93	$(0.57, 7.99)$	7.42	0.91	$(0.84, 7.83)$	6.99	1	$(0, 11.54)$	11.54
1	0	$(13.29, 18.23)$	4.94	0.94	$(0.89, 6.68)$	5.79	1	$(0, 9.26)$	9.26
2	0.09	$(4.10, 7.76)$	3.66	0.92	$(0.99, 5.54)$	4.55	1	$(0, 7.61)$	7.61
3	0.60	$(2.79, 6.13)$	3.34	0.92	$(1.05, 4.63)$	3.58	1	$(0, 6.38)$	6.38
4	0.99	$(1.11, 6.06)$	4.95	0.91	$(1.09, 3.94)$	2.85	1	$(0, 5.45)$	5.45
5	1	$(0, 7.10)$	7.10	0.90	$(1.08, 3.39)$	2.31	1	$(0, 4.77)$	4.77
6	1	$(0, 8.44)$	8.44	0.89	$(1.04, 2.94)$	1.91	1	$(0, 7.88)$	7.88
7	1	$(0, 10.42)$	10.42	0.90	$(0.96, 2.56)$	1.59	1	$(0, 10.39)$	10.39
8	1	$(0, 12.07)$	12.07	0.91	$(0.87, 2.24)$	1.36	1	$(0, 12.09)$	12.09
9	1	$(0, 13.04)$	13.04	0.90	$(0.78, 1.99)$	1.22	1	$(0, 10.07)$	10.07

[a] Coverage: the percentage of confidence interval covers the truth based on simulation.

[b] CI: Confidence interval.

[c] Length: the length of each confidence interval

3.3.2 Experiment 2

We increase the number of non-zero βs to $p = 400$ which means the signals are far from sparse. All other parameters remain the same. Therefore, the individual signal strength is weaker compared with Experiment 1. The simulation results are shown in Figure 24 and on the Table XIII-XIV. Figure 24 shows that the naive approach and EigenPrism approach is subject to very large upward bias while the bias of the subsampling approaches are smaller in comparison. However, the bias of the subsampling approaches is larger than those observed in Experiment 1. On the other side, the bias of the subsampling approach without adjustment yields smaller bias compared with the results in Experiment 1. This is consistent with what we expect before. From Table XIII, we see that the subsampling approaches have substantially smaller MSE in comparison to the Yang et al.'s approach and EigenPrism approach. The bootstrap subsampling approach and the subsampling with $q = 0.5, 0.8$ are slightly better than the subsampling with $q = 0.2$. For $q = 0.5$, the estimator from unadjusted selection criterion performs better that that from adjusted approach in respect to MSE. According, the weighted approach yields an estimator closer to the unadjusted estimator. In Table XIV, the 90% confidence interval obtained by sampling approach using Yang et al.'s approach cover the truth quite well for different αs. EigenPrism gives a good coverage when $\alpha = 1$. With variable selection, this approach no longer works well. The subsampling approach using EigenPrism consistently give a wider confidence interval compared with the subsampling approach using Yang et al.'s method.

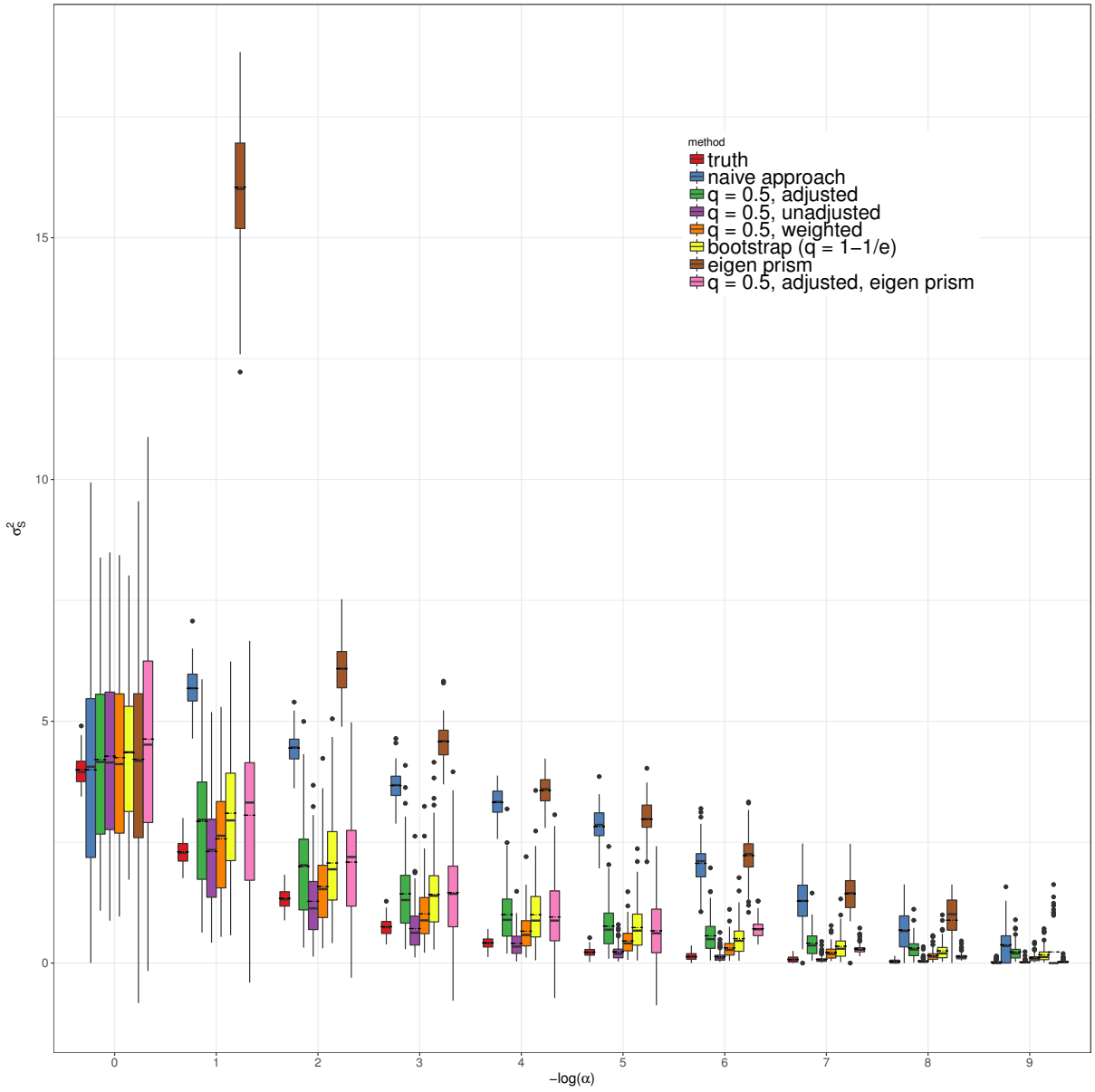


Figure 24: Box plot of subsampling approach with varied α in Experiment 2

TABLE XIII: MSE IN EXPERIMENT 2 FOR SUBSAMPLING APPROACH [a]

	Yang et al.'s approach							EigenPrism	
$-\log(\alpha)$	Naive	$q = 0.2$	$q = 0.5$, adjusted	$q = 0.5$, unadjusted	$q = 0.5$, weighted	bootstrap	$q = 0.8$	Naive	$q = 0.5$
0	5.42	4.79	3.16	3.13	3.1	2.12	1.33	5	4.96
1	11.58	2.75	1.93	1.09	1.3	2.28	3.69	190.3	2.66
2	9.76	2.03	1.35	0.52	0.7	1.54	2.66	22.83	1.85
3	8.62	1.93	0.99	0.19	0.36	1.02	1.94	14.82	1.35
4	8.51	1.67	0.65	0.07	0.19	0.68	1.12	10.02	0.84
5	6.87	1.33	0.47	0.02	0.11	0.44	0.53	7.7	0.56
6	3.89	1.08	0.29	0.01	0.06	0.23	0.26	4.57	0.36
7	1.66	0.81	0.17	0.01	0.04	0.13	0.13	2.01	0.06
8	0.58	0.63	0.11	0	0.02	0.08	0.07	0.97	0.01
9	0.25	0.47	0.07	0	0.01	0.04	0.03	0.26	0

[a] The Experiment 2 is set up as n = 400, m =4000, p = 400, $\sigma_g^2 = 4$, $\sigma_e^2 = 6$

TABLE XIV: 90% CONFIDENCE INTERVAL BY SUBSAMPLING APPROACH IN EXPERIMENT 2

$-\log(\alpha)$	EigenPrism			Subsampling Approach using LMM			Subsampling Approach using EigenPrism		
	coverage [a]	90% CI [b]	length [c]	coverage	90% CI	length	coverage	90% CI	length
0	0.9	$(0.48, 7.95)$	7.47	0.91	$(0.75, 7.67)$	6.93	1	$(0, 11.57)$	11.57
1	0	$(13.54, 18.55)$	5.01	0.96	$(0.18, 5.68)$	5.49	1	$(0, 8.35)$	8.35
2	0	$(4.24, 7.94)$	3.70	0.94	$(0, 3.99)$	3.99	1	$(0, 6.26)$	6.26
3	0	$(2.9, 6.26)$	3.36	0.93	$(0, 2.90)$	2.90	1	$(0, 4.85)$	4.85
4	0.01	$(1.09, 6.05)$	4.96	0.92	$(0, 2.09)$	2.09	1	$(0, 3.80)$	3.80
5	0.99	$(0, 6.95)$	6.95	0.90	$(0, 1.63)$	1.63	1	$(0, 3.17)$	3.17
6	1	$(0, 8.54)$	8.54	0.90	$(0, 1.21)$	1.21	1	$(0, 7.39)$	7.39
7	1	$(0, 11.01)$	11.01	0.92	$(0, 0.92)$	1.92	1	$(0, 10.58)$	10.58
8	1	$(0, 10.56)$	10.56	0.91	$(0, 0.70)$	0.70	1	$(0, 9.52)$	9.52
9	1	$(0, 2.69)$	2.69	0.92	$(0, 0.54)$	0.54	0.99	$(0, 3.12)$	3.12

[a] Coverage: the percentage of confidence interval covers the truth based on simulation.

[b] CI: Confidence interval.

[c] Length: the length of each confidence interval

3.3.3 Experiment 3

We increase the magnitude of variances to $\sigma_g^2 = 20, \sigma_e^2 = 30$ while keeping other parameters the same as in Experiment 1. The simulation results are shown in Figure 25, Table XV-XVI.

From Figure 25, we see the EigenPrism estimator is much more biased than other approaches when the α level is relatively large. For the subsampling approach, $q = 0.5$ and $q = 1 - 1/e$ yield similar results in respect to the bias and variation of the estimators. Since the signal strength is relatively large, we see that the adjusted estimator is less biased compared with unadjusted estimator. And weighted approach is closer to the adjusted approach here. Although subsampling approach can reduce the large upward bias in the EigenPrism approach, the bias from the EigenPrism using subsampling approach is more biased compared with subsampling approach using Yang et al.'s method.

From Table XV, we see that the relative merits of the approaches with respect to their MSEs are similar to that observed in Experiment 1. Still, EigenPrism approach yields a larger MSE compared with Yang et al.'s approach. The estimators from $q = 0.5$ and bootstrap give a smaller MSE compared with $q = 0.2$ or $q = 0.8$.

Table XVI shows that the confidence interval by Yang et al.'s approach under subsampling perform much better in terms of coverage and length of the confidence interval compared with EigenPrism approach with or without subsampling.

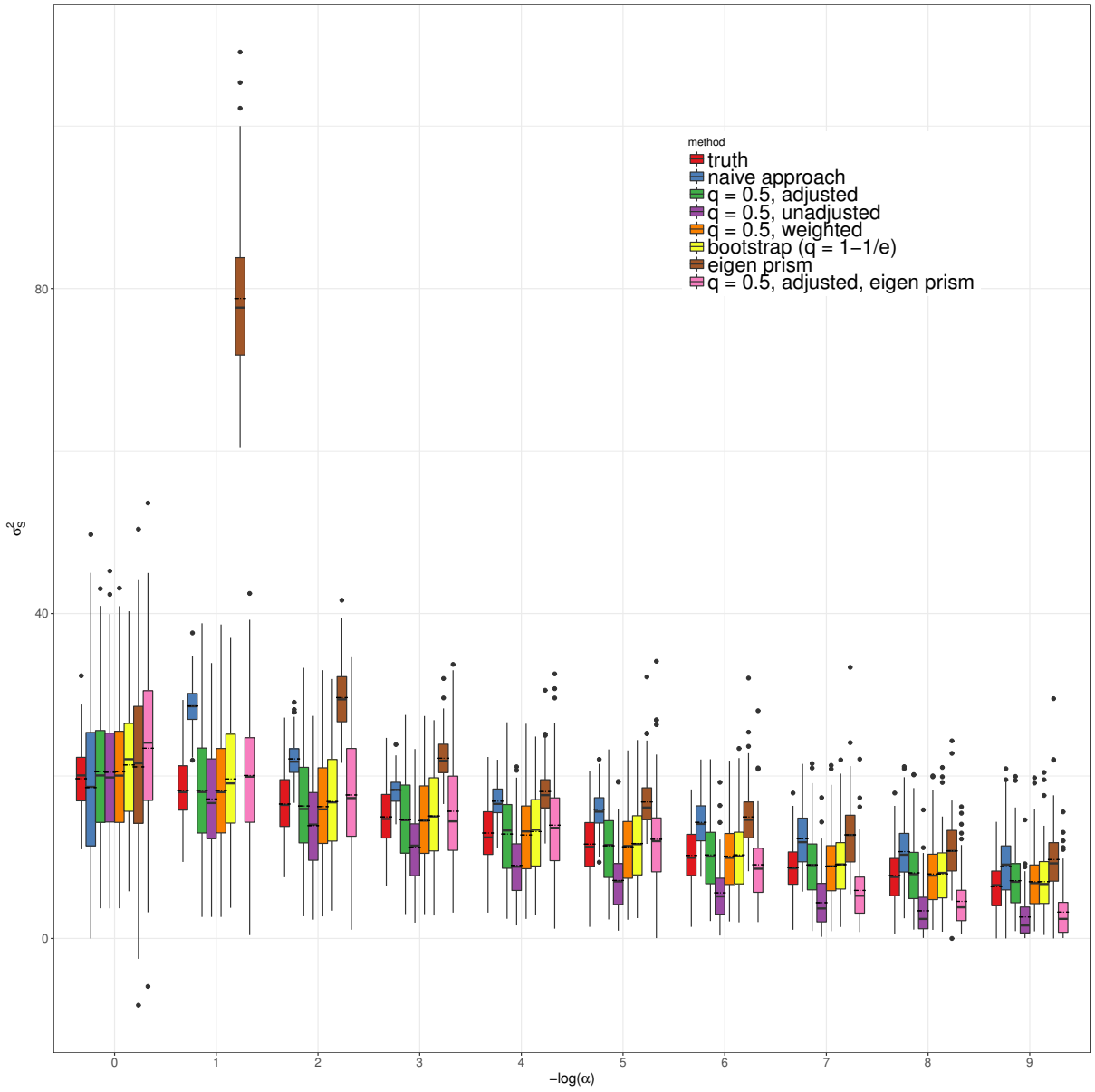


Figure 25: Box plot of subsampling approach with varied α in Experiment 3

TABLE XV: MSE IN EXPERIMENT 3 FOR SUBSAMPLING APPROACH [a]

	Yang et al.'s approach							EigenPrism	
$-\log(\alpha)$	Naive	$q = 0.2$	$q = 0.5$, adjusted	$q = 0.5$, unadjusted	$q = 0.5$, weighted	bootstrap	$q = 0.8$	Naive	$q = 0.5$
0	109.55	100.29	70.35	69.46	70.19	56.96	36.66	117.02	126.03
1	115.24	55.86	50.14	44.98	49.82	55.52	54.92	3773	81.65
2	38.92	37.48	35.7	28.7	35.11	38.99	44.42	200.75	64
3	20.21	25.86	23.29	26.54	22.83	26.29	24.54	74.32	57.41
4	22.34	16.19	13.72	25.1	13.44	13.5	15.72	48.81	47.68
5	24.03	13.33	9.22	28.33	9.11	8.92	9.86	50.19	44.64
6	20.06	10.27	7.08	30.5	7.01	6.99	7.1	50.65	31.98
7	15.54	8.18	5.9	26.86	5.78	5.19	6.21	46.01	33.58
8	11.86	8.32	4.99	27.28	4.91	4.99	5.86	36.11	36.73
9	9.37	8.49	4.22	20.41	4.01	4.27	5.06	42.36	33.73

[a] The Experiment 3 is set up as n = 400, m =4000, p = 40, $\sigma_g^2 = 20$, $\sigma_e^2 = 30$

TABLE XVI: 90% CONFIDENCE INTERVAL BY SUBSAMPLING APPROACH IN EXPERIMENT 3

$-\log(\alpha)$	EigenPrism			Subsampling Approach using LMM			Subsampling Approach using EigenPrism		
	coverage [a]	90% CI [b]	length [c]	coverage	90% CI	length	coverage	90% CI	length
0	0.93	$(2.58, 39.67)$	37.09	0.92	(3.21, 37.81)	34.6	1	$(0, 57.92)$	57.92
1	0	$(66.4, 91.16)$	24.76	0.93	(3.66, 32.82)	29.16	1	$(0, 46.33)$	46.33
2	0.20	$(20.5, 38.8)$	18.30	0.90	(4.82, 27.79)	22.97	1	$(0, 38.34)$	38.34
3	0.60	$(13.82, 30.56)$	16.74	0.91	(5.58, 23.63)	18.05	0.97	$(0, 32.43)$	32.43
4	0.95	$(5.77, 30.4)$	24.63	0.95	(5.72, 19.98)	14.27	0.94	$(0, 28.03)$	28.03
5	1	$(0, 35.66)$	35.66	0.94	(5.60, 17.22)	11.62	0.94	$(0, 24.64)$	24.64
6	1	$(0, 42.01)$	42.01	0.92	(5.50, 15.04)	9.54	1	$(0, 39.55)$	39.55
7	1	$(0, 51.76)$	51.76	0.91	(4.99, 13.11)	8.12	1	$(0, 52.06)$	52.06
8	0.99	$(0, 58.39)$	58.39	0.90	(4.58, 11.56)	6.98	1	$(0, 61.27)$	61.27
9	0.96	$(0, 64.24)$	64.24	0.90	(4.07, 10.17)	6.10	0.98	$(0, 50.09)$	50.09

[a] Coverage: the percentage of confidence interval covers the truth based on simulation.

[b] CI: Confidence interval.

[c] Length: the length of each confidence interval

3.3.4 Experiment 4

We increase the number of variables in the experiment and keep the ratio between m and p the same as in the previous experiment. The parameters are $n = 400, m = 10000, p = 100, \sigma_g^2 = 20, \sigma_e^2 = 30$. The simulation results are shown in Table XVII-XVIII, and Figure 26.

From Figure 26, we see that the subsampling approaches have smaller bias compared with the naive approach. EigenPrism approach yields higher bias when variable selection is existed compared with previous experiments. Although the naive approach gives an estimator with small bias in some α values, the relative bias of the estimator in general increases as α decreases. Under this setting-up, the adjusted approach yields smaller bias compared with unadjusted approach.

The MSE in Table XVII shows that the naive approach has large MSEs compared with the subsampling approaches. The EigenPrism approach can give similar MSE as naive approach when no selection is conducted. Otherwise, EigenPrism approach is consistently yield larger MSE than the naive approach and its corresponding subsampling approach also perform less well compared with subsampling approach with Yang et al.'s method in respect to MSE.

From Table XVIII, we also observe that confidence intervals from the subsampling approach using Yang et al.'s approach performs well. And the subsampling approach using EigenPrism consistently yields a coverage probability around 1, meaning that the confidence interval is too wide for the true signal estimators.

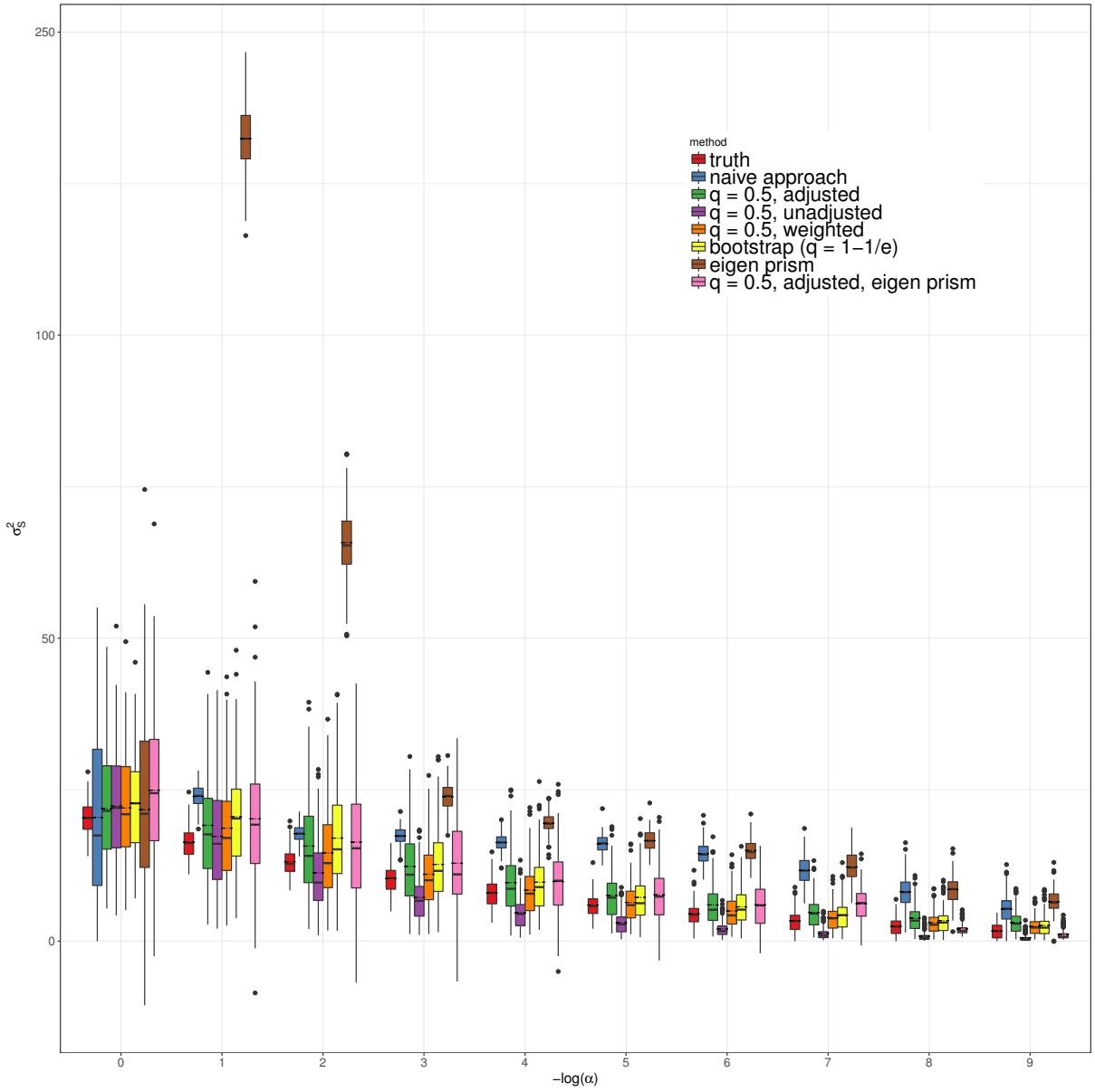


Figure 26: Box plot of subsampling approach with varied α in Experiment 4

TABLE XVII: MSE IN EXPERIMENT 4 FOR SUBSAMPLING APPROACH [a]

	Yang et al.'s approach							EigenPrism	
$-\log(\alpha)$	Naive	$q = 0.2$	$q = 0.5$, adjusted	$q = 0.5$, unadjusted	$q = 0.5$, weighted	bootstrap	$q = 0.8$	Naive	$q = 0.5$
0	224.43	177.32	95.61	99.87	95.57	66.68	38.62	255.14	218.27
1	61.75	100.26	82.35	70.11	77.05	91.67	100.1	32930.03	144.85
2	26.18	63.74	70.62	39.51	56.84	82.66	89.37	2795.17	98.56
3	53.9	49.22	43.28	25.32	31.49	44.02	67.53	189.11	65.45
4	73.5	38.93	31.44	21.05	22.04	32.05	42.32	137.72	41.53
5	105.08	35.29	18.46	13.29	12.04	17.81	20.98	117.07	24.8
6	103.3	31.24	14.11	9.7	8.96	11.76	11.91	113.72	17.35
7	74.12	26.64	7.65	6.21	4.3	5.92	5.87	84.95	14.31
8	38.6	22.52	5.5	3.8	2.88	4.15	3.86	40.98	1.37
9	16.91	20.95	4.17	2.31	2.07	2.82	2.55	25.77	1.27

[a] The Experiment 4 is set up as n = 400, m =10000, p = 100, $\sigma_g^2 = 20$, $\sigma_e^2 = 30$

TABLE XVIII: 90% CONFIDENCE INTERVAL BY SUBSAMPLING APPROACH IN EXPERIMENT 4

$-\log(\alpha)$	EigenPrism			Subsampling Approach using LMM			Subsampling Approach using EigenPrism		
	coverage [a]	90% CI [b]	length [c]	coverage	90% CI	length	coverage	90% CI	length
l0	0.96	$(0, 49.86)$	49.86	0.93	(3.16, 36.56)	33.40	1	$(0, 76.25)$	76.25
1	0	$(179.11, 215.24)$	36.13	0.94	(2.74, 32.70)	29.96	0.99	$(0, 59.39)$	59.39
2	0	$(53.54, 77.94)$	24.4	0.90	(2.66, 27.37)	24.71	1	$(0, 46.93)$	46.93
3	0.02	$(14.71, 32.98)$	18.27	0.90	(2.14, 22.55)	20.42	1	$(0, 37.06)$	37.06
4	0.12	$(10.92, 28.02)$	17.10	0.90	(1.77, 17.90)	16.13	1	$(0, 29.33)$	29.33
5	0.88	$(3.53, 29.65)$	26.12	0.88	(1.64, 14.37)	12.73	1	$(0, 23.71)$	23.71
6	1	$(0, 35.73)$	35.73	0.89	(1.47, 11.63)	10.16	1	$(0, 19.72)$	19.72
7	1	$(0, 43.53)$	43.53	0.92	(1.27, 9.48)	8.21	1	$(0, 18.48)$	18.48
8	1	$(0, 57.99)$	57.99	0.90	(1.08, 7.66)	6.58	1	$(0, 53.74)$	53.74
9	1	$(0, 62.64)$	62.64	0.88	(0.93, 6.33)	5.40	1	$(0, 51.05)$	51.05

[a] Coverage: the percentage of confidence interval covers the truth based on simulation.

[b] CI: Confidence interval.

[c] Length: the length of each confidence interval

3.4 Application to Real Data

In this section, we apply the proposed approaches to detecting the expression quantitative trait locus (eQTL) in a gene-expression study of human brain tissues. In this study, a total of 155 postmortem cerebellum samples were collected from the Stanley Medical Research Institute (SMRI), with 99 males and 56 females. All of them were of European ancestry. The Affymetrix Genome-wide Human SNP 5.0 Array was used for SNP genotyping. The outcome we use here is the gene expression of a probe that hybridizes to a specific genomic region in the chromosome. Correspondingly, the covariates we use are all the SNPs located in that chromosome. For illustrative purpose, we use a probe located at Chromosome 21 which has the smallest number of SNPs among all chromosomes except sex chromosomes. The probe with ID 8069448 was randomly chosen for the analysis. A more comprehensive analysis will be presented elsewhere.

Before analyzing the data, a standard quality checking (QC) procedure was performed. SNPs with $MAFs < 0.01$ and HWE $P < 0.001$ were excluded. Because Yang et al's approach does not work when with missing data, subjects were removed if there is any missing value in the SNPs. After the QC process, 130 samples are remained, and the number of SNPs used is 23,862.

The σ_α^2 estimates by the naive approach, the subsampling approach with different sampling ratios, and the bootstrap subsampling approach are shown in Table XIX. We also include 90% confidence interval obtained by using proposed variance of the estimators. Without variable selection, the bootstrap subsampling approach and subsampling approach with $q = 0.5$ yield estimators with variances lower than the naive approach. On the other hand, the subsampling

approach with $q = 0.2$ yields similar estimates of σ_α^2 to those of the naive approach. The naive approach consistently yields higher estimates compared with the other methods with variable selection. The results are compatible with those observed in simulation studies.

The results suggest that the selection effectively reduces the scope in searching for the causal SNPs. For example, when $\alpha = \exp(-4)$, the tests select 566 SNPs which account for 2.4% of the total SNPs. However, they explain about 50% of the total variation in the $23,862$ SNPs.

TABLE XIX: RESULTS OF THE ANALYSIS OF BRAIN TISSUE eQTL DATA

	Naive		$q = 0.2$		$q = 0.5$		bootstrap		$q = 0.8$	
$-\log(\alpha)$	SNPs	$\hat{\sigma}_S^2$	$\hat{\sigma}_S^2$	90% CI	$\hat{\sigma}_S^2$	90% CI	$\hat{\sigma}_S^2$	90% CI	$\hat{\sigma}_S^2$	90% CI
0	23862	0.55	0.5	(0.06,0.94)	0.43	(0,0.89)	0.45	(0,0.92)	0.48	(0.11,0.85)
1	8970	0.54	0.47	(0.03,0.91)	0.42	(0,0.89)	0.41	(0.01,0.81)	0.51	(0.14,0.88)
2	3524	0.42	0.34	(0,0.74)	0.27	(0,0.64)	0.34	(0,0.71)	0.31	(0,0.64)
3	1328	0.46	0.3	(0,0.70)	0.27	(0,0.60)	0.28	(0,0.56)	0.39	(0.06,0.72)
4	566	0.43	0.26	(0, 0.59)	0.17	(0,0.45)	0.26	(0,0.54)	0.29	(0.01,0.57)
5	220	0.48	0.23	(0, 0.51)	0.15	(0,0.38)	0.16	(0,0.39)	0.23	(0,0.46)
6	95	0.45	0.19	(0,0.47)	0.08	(0,0.24)	0.1	(0,0.26)	0.15	(0,0.38)
7	40	0.47	0.21	(0,0.49)	0.12	(0,0.28)	0.09	(0,0.25)	0.07	(0,0.23)
8	12	0.3	0.17	(0,0.4)	0.06	(0.06,0.06)	0.08	(0,0.24)	0.05	(0,0.17)
9	4	0.08	0.12	(0,0.35)	0.05	(0.05,0.05)	0.05	(0,0.21)	0.03	(0,0.10)

CHAPTER 4

ESTIMATING TOTAL EFFECT OF SELECTED VARIABLES: THE CASE OF CORRELATED COVARIATES

4.1 Modified Subsampling Approach for Correlated Covariates

In the previous chapter, we have concentrated on studying the subsampling approach under the assumption that the covariates having effects on the outcome are independent. For GWASs, if the causal SNPs are typed and in loci fairly apart, this assumption may be reasonable. However, if the causal SNPs are clustered or some of the causal SNPs are not typed and the tagged SNPs are correlated with other causal SNPs or tagged SNPs, this requirement fails to hold. In this chapter, we extend the proposed subsampling approach to the case where covariates with non-zero effects are correlated. Since we do not know the set of covariates that have effects, we consider correlated covariates in respect to whether they have an effect or not.

More specifically, let y be the outcome which is a quantitative trait in GWAS, let $x_1, \cdots, x_m$ be covariates denoting the typed SNP values in GWAS. Under the linear model

$$Y = \beta_0 + X\beta + \varepsilon, \tag{4.1}$$

with correlated covariates, i.e., $var(X) = \Sigma$, the variance of the outcome Y is

$$\text{var}(Y) = \beta^t \Sigma \beta + \sigma_\epsilon^2, \tag{4.2}$$

It is known that Yang et al.'s approach may not work well in estimating the total variation of the outcome attributable to the covariates when $\Sigma \neq I$. We first need to resolve this issue before applying it to estimating the variation explained by a set of non-randomly selected covariates.

Our proposed approach is a transformation to disentangle the correlation among covariates. Let the inverse of Σ be T. Let the new set of covariates Z be

$$Z = XT^{\frac{1}{2}}$$

After the transformation, $var(Z) = var(XT^{\frac{1}{2}}) = I$ and the linear model becomes

$$Y = \gamma_0 + Z\gamma + \varepsilon, \tag{4.3}$$

where $\gamma = (T)^{-\frac{1}{2}}\beta$. The total variation explained by the new covariate Z is

$$\text{var}(Z\gamma) = \gamma^t\gamma = \beta^t T^{-\frac{1}{2}} T^{-\frac{1}{2}}\beta \tag{4.4}$$

$$= \beta^t\Sigma\beta = \text{var}(X\beta). \tag{4.5}$$

This means the variation explained by the covariates are invariant under transformation. In fact, this property holds under any invertible linear transformation. In general, for any invertible matrix compatible with X. Let $W = XU$ and $\delta = U^{-1}\beta$, it follows that

$$X\beta = XUU^{-1}\beta = W\delta.$$

As a result, $\text{var}(W\delta) = \text{var}(X\beta)$. The benefit of transforming X into Z is that Yang et al's approach is now appropriate for calculating the variation explained by Z.

To carry out the transformation, we need to calculate the inverse of Σ, which can be challenging in high dimensional setting. Friedman et al. (2007) proposed an algorithm called graphic Lasso to calculate the inverse covariance matrix. Graphic Lasso, or glasso, uses the l^1 penalty in the estimating equations for the inverse of Σ to increase its sparsity. By extending the block-wise coordinate descent algorithm in Banerjee et al. (2007) and applying coordinate descent algorithms to solve individual penalty problems, glasso is able to compute the Σ^{-1} in remarkably faster speed compared with other competing methods. Later in 2015, Friedman and his colleagues have also developed a package called "glasso" to implement the glasso algorithm in R. The computing time of glasso is approximately $O(m^3)$, when Σ is an $m \times m$ matrix. It can be seen that the length of computing Σ^{-1} is mainly depended on the number of covariates. If the m is greater than 1000, which is quite normal in GWAS, the computing burden can be very intensive for a single calculation of the inverse of Σ. For the subsampling approach, we need to compute the Σ^{-1} multiple times. From the computation complexity point of view, using glasso in subsampling approach may not be efficient.

We propose an alternative approach based on the consecutive regressions. Start from variable x_1, let $z_1 = x_1$; next regress x_2 on x_1, let the residual be z_2; regress x_3 on x_1 and x_2, let the residual be z_3; for $i \geq 4$, regress x_i on $x_{i-1}, \cdots, x_1$, let the residual be z_i. When the number of variables is close to the sample size, the regression analysis becomes impractical. In those cases, a penalized regression analysis may be considered to obtain the residual as the transformation.

Alternatively, a parametric regression may be applied. For example, an autoregression model with a fixed order may be considered. In the simulation study, we consider an order of 3 in the obtaining the residuals.

To estimate the variation of the outcome explained by a set of randomly selected covariates, let S denote the set selected. The linear model links the outcome to the set of selected covariates is

$$Y = \beta_0 + X_S\beta_S^* + \varepsilon^*, \tag{4.6}$$

where X_S is the selected set of covariates and β_S^* is the set of regression coefficients and $E(\varepsilon^*) = 0$. Note that in general $\beta_S^* \neq \beta_S = (\beta_j, j \in S)$ because of the correlation among covariates. The variation of Y explained by the set S can be written as

$$\sigma_S^2(\lambda) = \beta_S^{*t} var(X_S)\beta_S^* = \beta^t cov(X, X_S) var^{-1}(X_S) cov(X_S, X)\beta \tag{4.7}$$

where X is the full set of covariates, and β is the true coefficients associated with X. In practice, as the true signal β is unknown, the second part of Equation 4.7 cannot be used directly for estimation. We may use the first part of Equation 4.7 to estimate $\sigma_S^2(\lambda)$ as we did when all the covariates are involved. However, our regression models for disentangle the correlation need to be applied to the set of covariates X_S instead. When the set of covariates are non-randomly selected, we use the subsampling approach again to circumvent the selection bias. The application of the approach remains the same as in the previous chapter once the

correlation is removed. For the simulation, we can use the Equation 4.7 to calculate the true signal captured by variable selection.

The following algorithm is used to implement the bootstrap subsampling approach for correlated cases,

1. For $q = 1 - 1/e$ randomly sample from the observed data $[n]$ subjects with replacement, then remove the repeat subject. Denote the select sample as A and the rest of sample as B.
2. Use sample A to select variables with varying tuning parameter values. Denote the selected covariate indices by S.
3. Use sample B to obtain the z_j for $j = 1, \ldots, s \in S$. This is done by regressing x_j on x_{j-1}, x_{j-2}, and x_{j-3} if $j \geq 3$. For $j < 3$, regression x_j on all the available x_k with $k < j$.
4. Fit a linear mixed model to $z_j, j \in S$. Estimate $\hat{\sigma}_t^2(\lambda)$.
5. Repeat the 1 - 4 steps for $t = 1, \cdots, T$.

4.2 Simulation Designs and Data Generation

Note that if we generate β as random vector whose components are either 0 or normally distributed with mean 0 as assumed in Yang et al's LMM approach, it follows that,

$$E(\beta'\Sigma\beta) = \sum_{j=1}^{m} E(\beta_j^2) + \sum_{j=1}^{m} \sum_{k \neq j} E(\beta_j) E(\beta_k) \rho_{jk}.$$

Since $E(\beta_j) = 0$, we can eliminate the second item on the right hand side of the equation and obtain

$$E(\beta'\Sigma\beta) = \sum_{j=1}^{m} E(\beta_j^2).$$

This suggests that correlations among covariates do not affect the LMM estimator. We expect using X or Z in the LMM approach yield similar results. In practice, however, the normality of the β with mean 0 is questionable. When β is not symmetrically distributed around 0, correlation among covariates may induce bias in the LMM estimator. Using X or Z in Yang et al's approach leads to different results and using Z is better. In the simulation, non-zero β components are simulated either following a normal distribution with mean 0 or appropriately fixed.

We consider two types of variable selection approaches. One continues to use the individual tests with varying significance levels by running a marginal regression. The other applies Lasso variable selection with varying tuning parameter values. The Lasso minimizes

$$||y - \beta_0 - X\beta||^2 + \lambda||\beta||_1 \tag{4.8}$$

The domain of β is restricted by the inequality $||\beta||_1 \leq t$, where t is a consistent associated with the choice of $\lambda \in [0, C]$. In the simulation studies, we choose $\lambda \in [0, 3.2]$.

To compare the proposed estimator for σ_λ^2 with other frequently used approaches in practice, we also implement the subsampling approach without transforming X. The procedure is similar to the above algorithm. The only difference is in step 3. Instead of calculating Z matrix, for

selected variables in the previous step, we directly use the subset of sample B to estimate the attributable variation of Y explained by the selected covariates by Yang et al's approach. In addition to subsampling approach, we also include the naive approach using X and Z for estimation in the simulation. To compare the efficiency between Lasso approach and individual tests in selecting variables, we include the individual tests for variable selection as we did for the independent cases in the previous chapter. Since the EigenPrism approach does not perform nearly as well as the LMM approach for non-randomly selected covariates when covariates are independent, we expect it does not perform well either in the case of correlated covariates. This is because the proposed de-correlation approaches relies on the corresponding estimation approaches for independent covariates. For this reason, we do not include this approach in our simulation study for comparison here.

The simulated covariates are the bi-allelic genetic markers taking values $0, 1, 2$. The simulated outcome is a quantitative trait. The covariance matrix of covariates follows the autoregression model: the diagonal elements $\Sigma_{jj} = 1$and the off-diagonal elements $\Sigma_{jk} = \rho^{|j-k|}$. Specifically,

1. $\mathbf{X}$: Generate a random number π following $Uniform(0.05, 0.5)$ distribution. Generate n independent random numbers, each follows $Binomial(2, \pi)$ distribution. Repeat m times to generate the initial covariant matrix. Let the new $X_1 = X_1$, for $i = 2, \ldots, m$, generate a random number $r \sim U(0, 1)$, if $r < \rho$ then the new $X_i = X_{i-1}$ otherwise the new $X_i = X_i$. Standardized X_i to have mean 0 and variance 1. This makes the new X has

an auto-regressive covariance-variance structure with each element equaling $\rho^{|i-j|}$, where i and j are the index of two columns of X.

2. $\boldsymbol{\beta}$: In order to verify when Yang's approach collapses in practice, we generate β in two ways. In the first way, non-zero βs will be random, e.g., $\beta_j, j = 1, \cdots, p \overset{iid}{\sim} N(0, \sigma_g^2/p)$. In another way, non-zero βs will be fixed, e.g., $\beta_j^2 = \sigma_g^2/p$ for $j = 1, \cdots, p$. Set $\beta_j = 0, j = p+1, \cdots, m$.

3. $\boldsymbol{Y}$: Generate ϵ following $N(0, \sigma_e^2\mathbf{I})$ and set $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$.

4. In each iteration, to calculate the truth , first apply the Lasso approach with $\lambda \in [0, 3.2]$ on the full data, then use the formula Equation 4.7 to calculate the truth.

We replicate the algorithm N times and compute the mean squared error (MSE) to measure the performance of the estimator,

$$MSE = \frac{1}{N}\sum_{l=1}^{N}(\hat{\sigma}_{l\lambda}^2 - \sigma_\lambda^2)^2. \tag{4.9}$$

4.3 Simulation Results

In the following experiments, we specify parameters $(n, m, p, \sigma_g^2, \sigma_e^2, \rho)$ for each setting taking the sparsity of the non-zero SNP effects into consideration. Bootstrap subsampling with $q = 1 - 1/e$ is used here as it gives the smallest MSE in different experiment in the last chapter for independent covariates. Four approaches are compared here, the naive approach using X, the naive approach using Z, the subsampling approach using X, and the subsampling approach using Z.

4.3.1 Experiment 1

The parameters were set as follow: $n = 1000, m = 4000, p = 80, \sigma_g^2 = 20, \sigma_e^2 = 30, \rho = 0.5$, and we generate the βs by following the random normal distribution. Individual tests with adjustment and Lasso approach are used for variable selection.

Table XX and Figure 27 present the results for using Lasso for variable selection, where we chose to vary the λ in order to select varied size of X in the data. Figure 27 is the box plot with varied λs. It can be seen that overall subsampling approach win over the naive approach with smaller bias and acceptable variance. Although the estimator from subsampling approach using X is slightly downward biased compared with subsampling approach using Z, both of them work fine as the boxes of them are highly overlapped with the truth box across different λs. In other words, the results using Z and X are very similar.

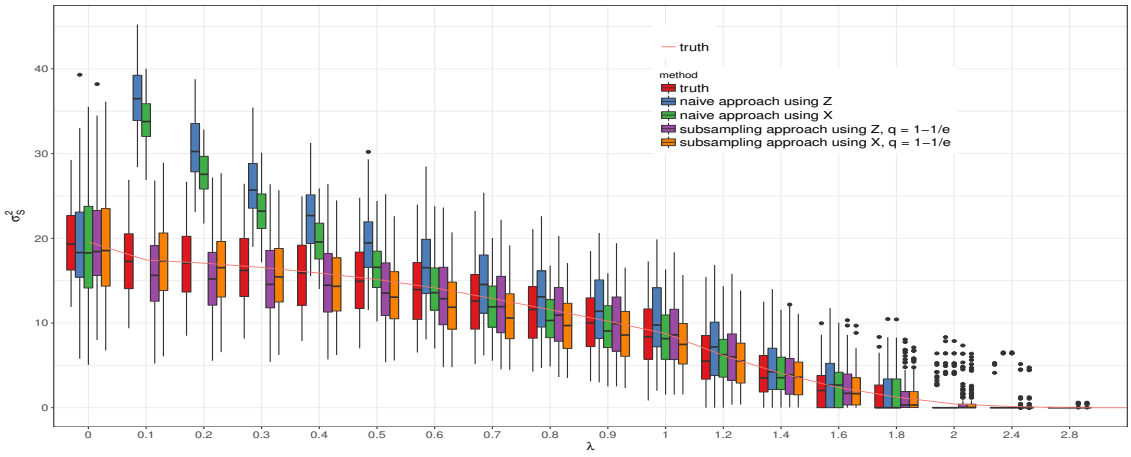


Figure 27: Box plot with varied λ in Experiment 1

Table XX gives the MSE results obtained through simulation. It shows that subsampling approach using Z as the covariants yields the smallest MSE of the estimators among four methods. Naive approach starts with much larger MSE compared with subsampling approach, and follows with closer MSE with subsampling approach as λ level becomes larger, meaning that the selection criterion is stricter.

TABLE XX: MSE IN EXPERIMENT 1 BY LASSO [a,b]

λ [c]	naive with Z	naive with X	subsampling with Z	subsampling with X
0	16.1	14.84	17.81	15.74
0.1	376.94	278.43	6.69	4.36
0.2	185.12	118.6	6.61	4.18
0.3	93.93	50.26	5.58	4.15
0.4	45.96	19.67	4.77	5.06
0.5	21.45	7.59	4.47	7.07
0.6	9.35	4.7	3.63	7.83
0.7	5.4	5.59	3.04	7.56
0.8	3.94	4.78	2.61	6.45
0.9	3.53	3.6	2.29	5.15
1	3.42	2.65	1.79	3.57
1.2	2.46	1.99	1.62	2.41
1.4	1.54	1.13	1.24	1.53
1.6	1.11	0.81	1.17	1.18
1.8	0.75	0.56	0.97	0.9
2	0.43	0.38	0.5	0.47
2.4	0.09	0.09	0.05	0.04
2.8	0	0	0.01	0.01
3.2	0	0	0	0

[a] n = 1000, m = 4000, p = 80, $\sigma_g^2 = 20$, $\sigma_e^2 = 30$, $\rho = 0.5$.
[b] βs is generated by following the random normal distribution.
[c] λ stands for tuning parameters in Lasso regression.

Table XXI and Figure 28 present the same statistics using individual tests for variable selection. It gives similar results in respect to the performance of subsampling approach and naive approach. In Figure 28, when $-\log(\alpha)$ becomes larger, the total variation explained by selected variables decreases slower compared with that using Lasso regression. Table XXI also shows that subsampling with matrix Z can yield the smallest MSE compared with using X in subsampling approach.

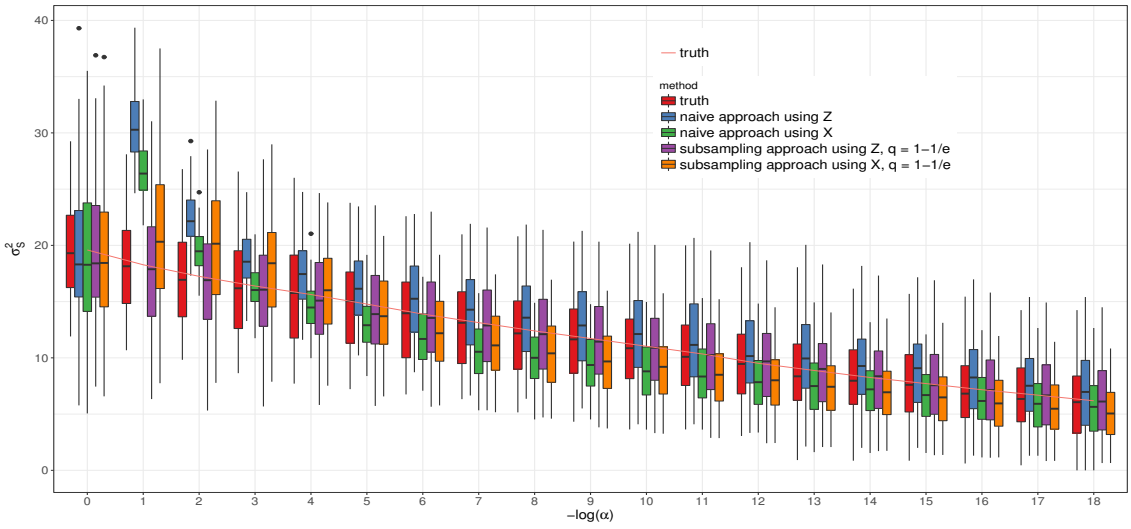


Figure 28: Box plot with varied α in Experiment 1

TABLE XXI: MSE IN EXPERIMENT 1 BY INDIVIDUAL TESTS [a,b]

$-\log(\alpha)$ [c]	naive with Z	naive with X	subsampling with Z	subsampling with X
0	16.1	14.84	14.91	15.13
1	155.91	76.07	9.61	19.99
2	31.03	11.99	6.56	15.38
3	11.07	7.71	4.92	6.76
4	7.05	8.5	4.18	3.09
5	5.08	9.51	3.3	3.62
6	4.13	10.07	2.52	5.05
7	3.13	10.51	2.49	6.66
8	2.84	9.86	2.1	6.94
9	2.63	9.19	2.45	7.51
10	2.67	7.88	1.87	6.73
11	2.74	6.48	1.79	6.48
12	2.65	5.3	1.91	5.01
13	2.71	4.72	1.89	4.83
14	2.69	3.9	2.14	4.23
15	2.6	3.64	1.83	3.85
16	2.41	3.34	1.65	3.31
17	2.16	3.2	1.63	2.99
18	1.97	2.71	1.75	3.04

[a] n = 1000, m = 4000, p = 80, $\sigma_g^2 = 20$, $\sigma_e^2 = 30$, $\rho = 0.5$.
[b] βs is generated by following the random normal distribution.
[c] α stands for tuning parameters in individual tests.

Table XXII shows the efficiency of using lasso approach and individual tests for variable selection. We list the selected sample size at each selection level, the total variation explained by the selected covariants, and the average signal strength for each coefficient under certain selection level. It can be seen that Lasso approach can yield a smaller subset with similar total signal strength than individual test, indicating more efficient than individual test.

TABLE XXII: EFFICIENCY COMPARISON IN EXPERIMENT 1

α	selected size	$\hat{\sigma}_g^2$	$\frac{\hat{\sigma}_g^2}{n}$	λ	selected size	$\hat{\sigma}_g^2$	$\frac{\hat{\sigma}_g^2}{n}$
0	4000	19.7	0	0	4000	19.63	0
1	1919.98	18.02	0.01	0.1	485.68	15.84	0.03
2	972.92	16.9	0.02	0.2	346.44	15.5	0.04
3	510.29	16.1	0.03	0.3	236.94	15.19	0.06
4	275.56	15.24	0.06	0.4	153.64	14.68	0.1
5	154.61	14.36	0.09	0.5	94.04	13.97	0.15
6	91.15	13.69	0.15	0.6	54.96	13.14	0.24
7	56.86	12.89	0.23	0.7	31.52	12.17	0.39
8	37.7	12.18	0.32	0.8	18.82	11.05	0.59
9	26.73	11.53	0.43	0.9	12.23	9.86	0.81
10	20.16	10.83	0.54	1	8.55	8.65	1.01
11	15.94	10.14	0.64	1.2	4.64	6.23	1.34
12	13.08	9.52	0.73	1.4	2.45	4.12	1.68
13	11.03	8.92	0.81	1.6	1.22	2.55	2.09
14	9.43	8.37	0.89	1.8	0.56	1.42	2.55
15	8.19	7.79	0.95	2	0.23	0.7	3.08
16	7.14	7.31	1.02	2.4	0.04	0.17	4.62
17	6.25	6.76	1.08	2.8	0	0.02	4.01
18	5.52	6.32	1.14	3.2	0	0	NA

4.3.2 Experiment 2

In this experiment, we keep all the parameter settings the same, but non-zero βs are simulated by assuming that all of them are fixed at the level $\sqrt{\sigma_g^2/p}$. We also compare Lasso regression and individual tests for the variable selection. For Lasso approach, as the signal is reduced to 0 when $\lambda > 1.6$, we do not include the results of $\lambda > 1.6$ here. Instead, we add more λs between 0 and 0.1, as the upward bias are more severe when λ within this range. Table XXIII and Figure 29 present the results for Lasso selection.

Table XXIII shows that subsampling approach can reduce the MSE of estimator when variables are selected. Additionally, subsampling approach using Z seems to have much less MSE compared with that using X when λ level is relatively large. This is also consistent with what we expect before simulation. Figure 29 shows that subsampling approach with Z performs best with small bias and relatively small variance. Although the naive approach using matrix Z can work fine when λ level is relatively large, it collapses by yielding a large upward bias for small λs.

TABLE XXIII: MSE IN EXPERIMENT 2 BY LASSO [a,b]

λ [c]	naive with Z	naive with X	subsampling with Z	subsampling with X
0	72.64	295.74	66.07	634.8
0.02	282.02	95.07	22.97	14.03
0.04	249.18	107.92	22.42	7.72
0.06	214.96	130.55	21.36	11.3
0.08	179.8	159.86	21.32	18.03
0.1	149.73	192.28	19.28	26.89
0.2	45.1	393.98	17.38	110.82
0.3	16.7	578.91	16	229.83
0.4	11.73	740.29	13.92	348.43
0.5	10.95	843.6	13.52	438.57
0.6	11.05	871.19	12.26	483.27
0.7	11.89	832.05	11.58	483.01
0.8	13.64	743.52	11.96	448.31
0.9	14.48	637.04	12.97	397.64
1	13.36	520.76	11.6	333.75
1.2	17.59	286.21	13.05	207.35
1.4	32.64	96.46	10.4	93.45
1.6	60.45	9.29	9.58	19.91

[a] n = 1000, m = 4000, p = 80, $\sigma_g^2 = 20$, $\sigma_e^2 = 30$, $\rho = 0.5$.

[b] Non-zero βs is assumed to be fixed at the level $\sqrt{\sigma_g^2/p}$.

[c] λ stands for tuning parameters in Lasso regression.

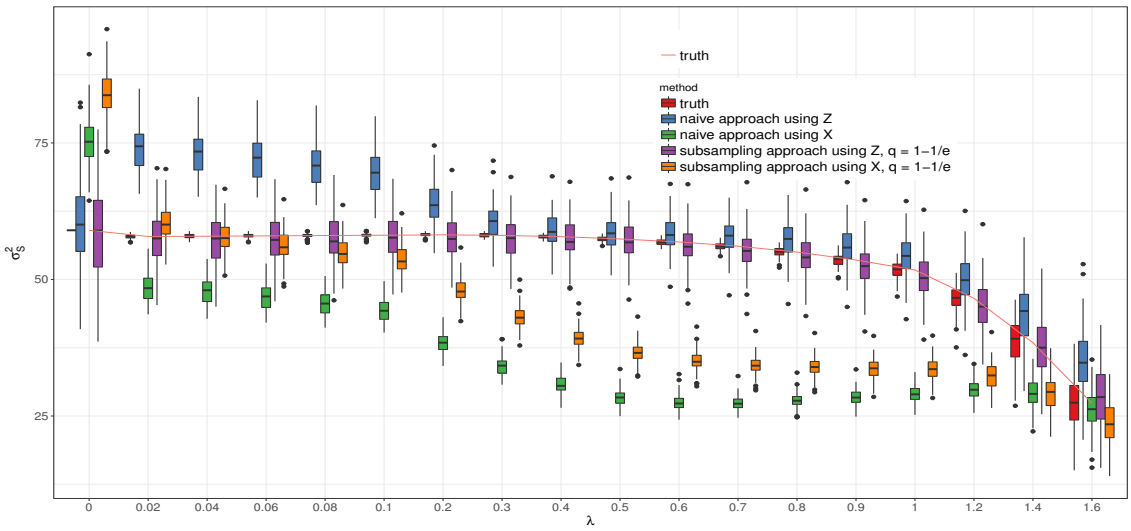


Figure 29: Box plot with varied λ in Experiment 2

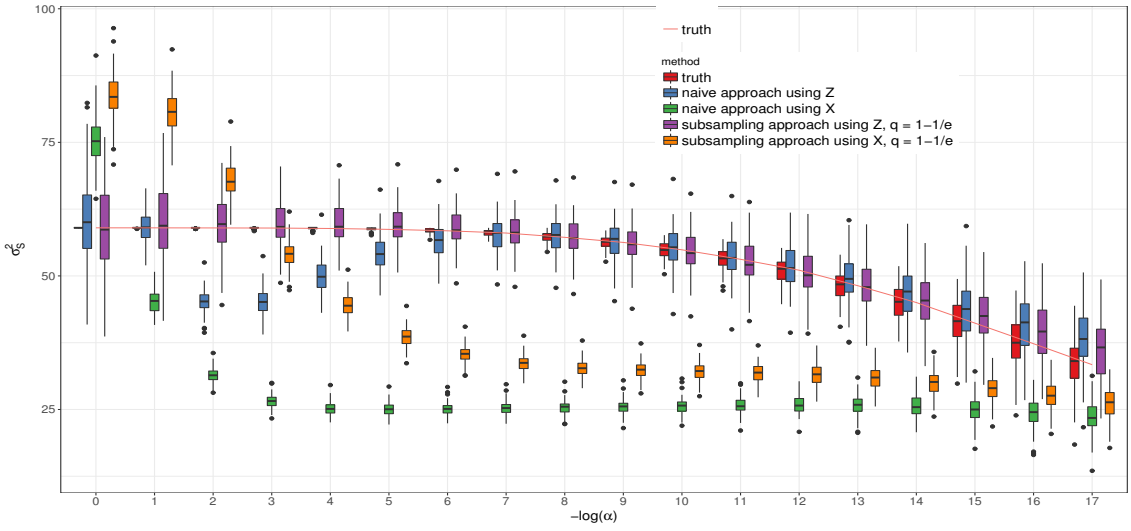


Figure 30: Box plot with varied α in Experiment 2

Table XXIV and Figure 30 display the results when using individual tests for variable selection. The overall trend is similar with that using Lasso regression for variable selection. In Table XXIV, we can see that subsampling approach using Z consistently wins over other approaches by giving the smallest MSE across different levels of α.

TABLE XXIV: MSE IN EXPERIMENT 2 BY INDIVIDUAL TESTS [a,b]

$-\log(\alpha)$ [c]	naive with Z	naive with X	subsampling with Z	subsampling with X
0	72.64	295.74	72.06	634.83
1	8.95	197.53	50.76	482.25
2	193.37	761.09	30.28	90.77
3	200.03	1054.46	19.54	30.08
4	88.06	1138.65	15.38	211.73
5	30.79	1138.69	12.61	410.53
6	13.26	1115.15	11.56	540.65
7	9.75	1076.97	11.24	601.27
8	9.38	1017.8	11.5	603.27
9	9.72	951.51	11.33	575.36
10	10.11	858.8	11.15	523.56
11	9.25	752.2	10.98	457.82
12	9.11	640.68	10.62	390.6
13	9.3	505.81	9.43	303.35
14	10.39	386.52	8.99	230.83
15	13	270.93	8.3	156.52
16	19.3	175.58	12.01	101.09
17	27.67	102.65	15.86	60.81

[a] n = 1000, m = 4000, p = 80, $\sigma_g^2 = 20$, $\sigma_e^2 = 30$, $\rho = 0.5$.

[b] Non-zero βs is assumed to be fixed at the level $\sqrt{\sigma_g^2/p}$.

[c] α stands for tuning parameters in individual tests.

We also compare the efficiency of both selection methods in Table XXV. It gives similar results as Experiment 1.

TABLE XXV: EFFICIENCY COMPARISON OF TWO SELECTION METHODS IN EXPERIMENT 2

α	selected size	$\hat{\sigma}_g^2$	$\frac{\hat{\sigma}_g^2}{n}$	λ	selected size	$\hat{\sigma}_g^2$	$\frac{\hat{\sigma}_g^2}{n}$
0	4000	58.68	0.01	0	4000	58.4	0.01
1	1938.77	59.88	0.03	0.02	692.83	57.59	0.08
2	1002.25	59.66	0.06	0.04	613.98	57.38	0.09
3	546.65	59.66	0.11	0.06	565.98	57.46	0.1
4	317.89	59.59	0.19	0.08	527.9	57.48	0.11
5	200.74	59.25	0.3	0.1	494.3	57.59	0.12
6	139.26	58.74	0.42	0.2	359.44	57.55	0.16
7	105.87	58.05	0.55	0.3	256.84	57.48	0.22
8	86.76	57.1	0.66	0.4	180.62	57.33	0.32
9	74.71	55.91	0.75	0.5	127.83	56.86	0.44
10	66.4	54.43	0.82	0.6	93.9	56.13	0.6
11	59.53	52.59	0.88	0.7	73.35	55.23	0.75
12	53.66	50.5	0.94	0.8	60.73	54.08	0.89
13	48.22	48.09	1	0.9	52.22	52.39	1
14	43.29	45.31	1.05	1	45.81	50.51	1.1
15	38.4	42.39	1.1	1.2	35.18	45.06	1.28
16	33.8	39.29	1.16	1.4	25.7	37.57	1.46
17	29.53	35.93	1.22	1.60	17.29	28.40	1.64

4.3.3 Experiment 3

In this experiment, we increase the p from 80 to 800, therefore, the data becomes less sparse, and many weak signals are associated in the data. β are generated by following normal distribution in this experiment.

Table XXVI and Figure 31 show results using Lasso approach for variable selection. From Table XXVI, we see that when the individual signal strength is weak, using Lasso regression with $\lambda ge1.4$ will cannot select any X from the data, therefore, the MSE becomes to be zero as there is no covariates included in the analyses when λ is relatively large. In other words, it

can be seen compared with Experiment 1, the strength of signal reduces faster. Additionally, it shows that the subsampling approach can give smaller MSE compared with naive approach.

TABLE XXVI: MSE IN EXPERIMENT 3 BY LASSO [a,b]

λ [c]	naive with Z	naive with X	subsampling with Z	subsampling with X
0	28.76	18.81	30.03	21.95
0.1	691.94	589.76	10.88	9.2
0.2	449.96	368.26	8.95	7.72
0.3	315.52	255.92	6.07	5.27
0.4	226.25	183.71	4.15	3.82
0.5	156.01	126.97	2.41	2.3
0.6	90.52	74	1.47	1.5
0.7	39.43	32.68	1.04	1.09
0.8	13.11	11.11	0.81	0.84
0.9	3.67	3.25	0.51	0.52
1	1.09	1	0.27	0.27
1.2	0.11	0.11	0.04	0.04
1.4	0	0	0	0
1.6	0	0	0	0
1.8	0	0	0	0
2	0	0	0	0
2.4	0	0	0	0
2.8	0	0	0	0
3.2	0	0	0	0

[a] n = 1000, m = 4000, p = 800, $\sigma_g^2 = 20$, $\sigma_e^2 = 30$, $\rho = 0.5$.
[b] βs is generated by following the random normal distribution.
[c] λ stands for tuning parameters in Lasso regression.

Figure 31 shows that subsampling approach yields less biased estimator compared with naive approach, although the estimator gives larger downward bias compared with that in Experiment 1. For subsampling approach using Z and X, there is not much difference between the two results.

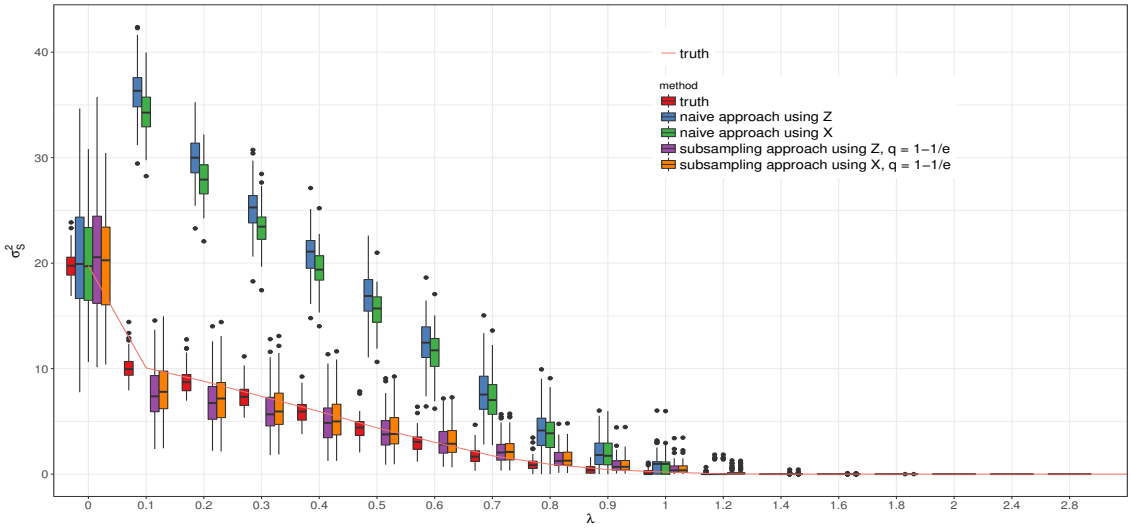


Figure 31: Box plot with varied λ in Experiment 3

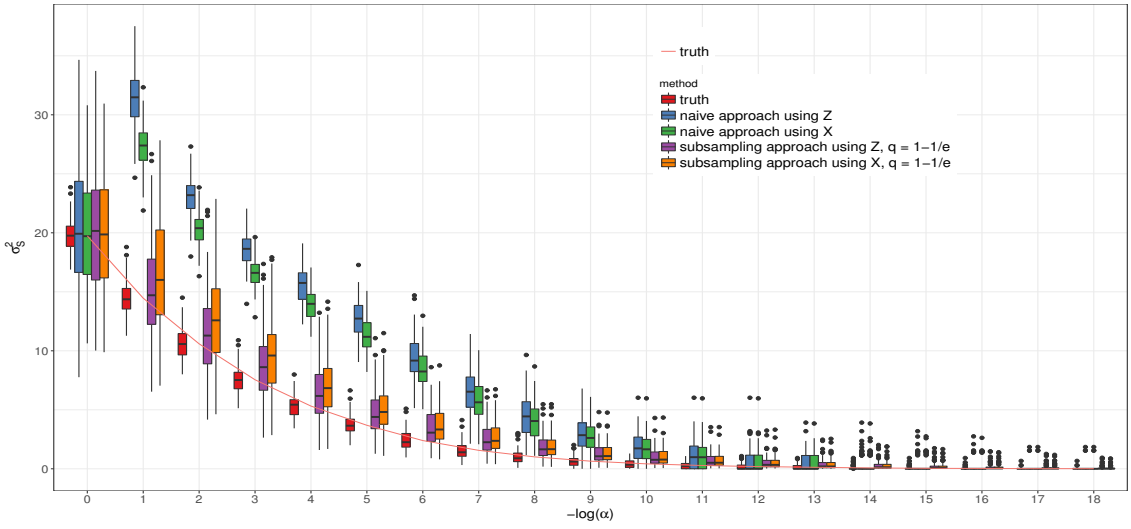


Figure 32: Box plot with varied α in Experiment 3

Table XXVII and Figure 32 give the results when the individual tests are used for variable selection. It also shows the superiority of subsampling approach when the data is correlated. Table XXVIII shows that when the data is not sparse and the signals are weak, the efficiency of lasso approach and individual tests are similar.

TABLE XXVII: MSE IN EXPERIMENT 3 BY INDIVIDUAL TESTS [a,b]

$-\log(\alpha)$ [c]	naive with Z	naive with X	subsampling with Z	subsampling with X
0	28.76	18.81	28.3	19.75
1	288.39	169.15	17.43	22.85
2	160.55	97.12	11.74	16.7
3	123.91	82.96	7.6	11.36
4	106.26	74.55	5.58	7.55
5	82.33	59.07	3.56	4.56
6	51.72	37.46	2.55	3.02
7	26.4	19.31	1.84	2.06
8	13.49	10.36	1.29	1.36
9	6.49	5.2	0.88	0.9
10	3.13	2.66	0.64	0.64
11	1.65	1.48	0.42	0.41
12	1.05	0.94	0.27	0.26
13	0.56	0.51	0.17	0.17
14	0.38	0.33	0.1	0.09
15	0.23	0.2	0.06	0.06
16	0.15	0.14	0.04	0.04
17	0.05	0.05	0.03	0.03
18	0.05	0.05	0.01	0.01

[a] n = 1000, m = 4000, p = 800, $\sigma_g^2 = 20$, $\sigma_e^2 = 30$, $\rho = 0.5$.
[b] βs is generated by following the random normal distribution.
[c] α stands for tuning parameters in individual tests.

TABLE XXVIII: EFFICIENCY COMPARISON OF TWO SELECTION METHODS IN EXPERIMENT 3

α	selected size	$\hat{\sigma}_g^2$	$\frac{\hat{\sigma}_g^2}{n}$	λ	selected size	$\hat{\sigma}_g^2$	$\frac{\hat{\sigma}_g^2}{n}$
0	4000	20.51	0.01	0	4000	20.81	0.01
1	1969.6	15.25	0.01	0.1	500.44	7.65	0.02
2	1025	11.44	0.01	0.2	371.8	6.75	0.02
3	548.93	8.59	0.02	0.3	268.51	5.9	0.02
4	300.35	6.44	0.02	0.4	186.16	5	0.03
5	167.5	4.74	0.03	0.5	122.09	3.97	0.03
6	94.41	3.45	0.04	0.6	74.8	3.1	0.04
7	54.29	2.54	0.05	0.7	41.84	2.23	0.05
8	31.57	1.86	0.06	0.8	20.97	1.51	0.07
9	18.68	1.35	0.07	0.9	9.42	0.93	0.1
10	11.13	1	0.09	1	3.83	0.53	0.14
11	6.71	0.73	0.11	1.2	0.5	0.12	0.23
12	4.1	0.53	0.13	1.4	0.05	0.01	0.26
13	2.54	0.39	0.15	1.6	0	0	0.29
14	1.59	0.28	0.18	1.8	0	0	0
15	0.99	0.19	0.2	2	0	0	NA
16	0.63	0.14	0.23	2.4	0	0	NA
17	0.41	0.1	0.24	2.8	0	0	NA
18	0.26	0.06	0.24	3.2	0	0	NA

4.3.4 Experiment 4

In this experiment, β are generated in the fixed way, and all the parameters are set the same as Experiment 3. It can been seen from the Table XXIX that subsampling approach still works much better compared with naive approach. However, using Z or X make no difference in the MSE of estimators. This might due to the fact that the fixed effect of β is very small. Figure 33 also shows that the subsampling approach using X or Z work well compared with naive approach.

TABLE XXIX: MSE IN EXPERIMENT 4 BY LASSO [a,b]

λ [c]	naive with Z	naive with X	subsampling with Z	subsampling with X
0	73.06	272.1	79.69	607.5
0.02	429.46	59.5	110.83	38.54
0.04	441.09	73.13	122.14	47.82
0.06	436.77	74.85	126.28	52.32
0.08	425.4	72.24	123.52	52.97
0.1	396.69	65.36	126.73	57.92
0.2	309.72	51.69	111.2	59.51
0.3	287.48	62.2	87.29	52.2
0.4	297.5	89.61	64.07	42.35
0.5	344.39	142.73	39.7	27.99
0.6	394.01	205.94	22.35	17.22
0.7	426.37	262.77	11.86	10.32
0.8	412.62	283.71	8.36	8.6
0.9	346.32	255.9	7.6	8.31
1	236.91	184.41	8.41	9.08
1.2	52.72	43.84	5.45	5.65
1.4	5.37	4.8	1.48	1.52
1.6	0.53	0.53	0.27	0.27

[a] n = 1000, m = 4000, p = 800, $\sigma_g^2 = 20$, $\sigma_e^2 = 30$, $\rho = 0.5$.

[b] Non-zero βs is assumed to be fixed at the level $\sqrt{\sigma_g^2/p}$.

[c] λ stands for tuning parameters in Lasso regression.

We also include the results by using individual tests for variable selection in Table XXX and Figure 34. From Figure 34, we can see that subsampling approach using Z can still reduce the bias a lot compared with other approaches especially when the α level is relatively large. However, such superiority disappears when α becomes small and all the approaches do not perform well compared with the truth.

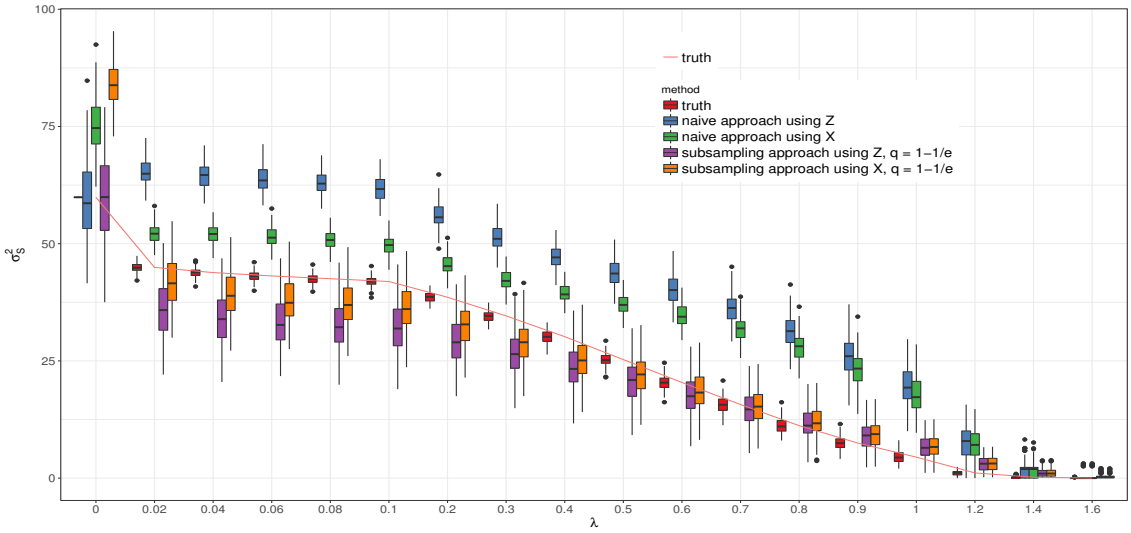


Figure 33: Box plot with varied λ in Experiment 4

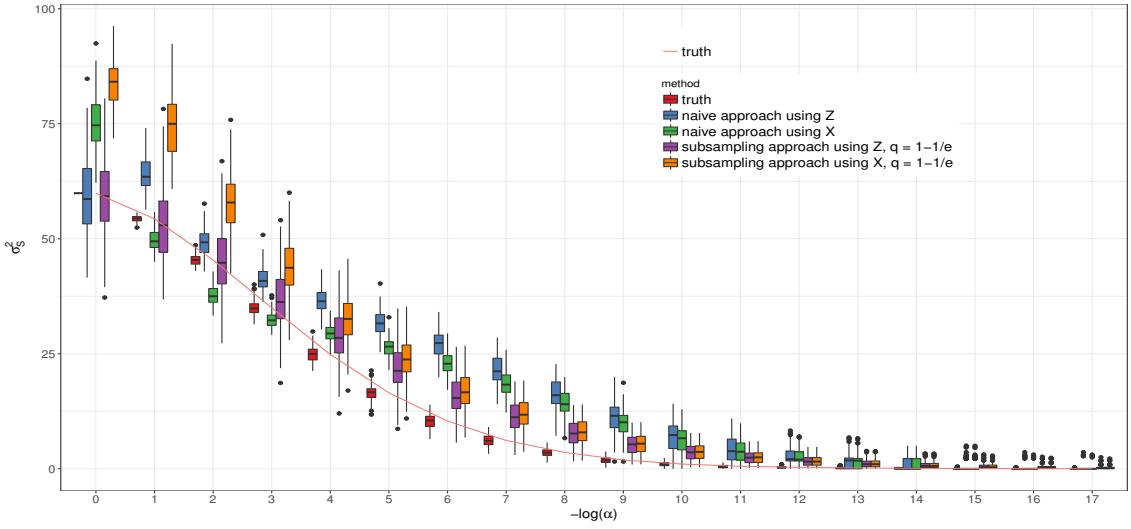


Figure 34: Box plot with varied α in Experiment 4

From Table XXX, it can be seen that subsampling approach with matrix Z works the best in respect to the MSE. Table XXXI shows that under the existence of weak signals, individual tests and Lasso approach performs the similar in respect of the selection efficiency.

TABLE XXX: MSE IN EXPERIMENT 4 BY INDIVIDUAL TESTS [a,b]

$-\log(\alpha)$ [c]	naive with Z	naive with X	subsampling with Z	subsampling with X
0	73.06	272.1	81.89	599.83
1	111.84	25.85	72.83	454.18
2	21.19	62.17	50.65	193.56
3	42.94	8.82	44.63	109.38
4	136.94	24.74	42.37	81.46
5	237.55	100.28	42.9	65.65
6	284.61	161.33	39.94	51.84
7	243.85	158.36	33.73	39.44
8	168.75	117.35	22.56	24.99
9	92.25	68.72	15.06	16.03
10	44.82	34.81	9.28	9.61
11	19.92	15.92	5.26	5.34
12	9.62	7.94	2.98	3
13	4.94	4.22	1.69	1.68
14	2.61	2.47	0.98	0.97
15	1.52	1.45	0.58	0.56
16	0.58	0.52	0.36	0.35
17	0.39	0.33	0.21	0.2

[a] n = 1000, m = 4000, p = 800, $\sigma_g^2 = 20$, $\sigma_e^2 = 30$, $\rho = 0.5$.

[b] Non-zero βs is assumed to be fixed at the level $\sqrt{\sigma_g^2/p}$.

[c] α stands for tuning parameters in individual tests.

TABLE XXXI: EFFICIENCY COMPARISON OF TWO SELECTION METHODS IN EXPERIMENT 4

α	selected size	$\hat{\sigma}_g^2$	$\frac{\hat{\sigma}_g^2}{n}$	λ	selected size	$\hat{\sigma}_g^2$	$\frac{\hat{\sigma}_g^2}{n}$
0	4000	59.63	0.01	0	4000	59.89	0.01
1	2104.73	53.54	0.03	0.02	710.89	36.21	0.05
2	1185.09	45.16	0.04	0.04	634.67	34.31	0.05
3	691.52	36.81	0.05	0.06	592.52	33.35	0.06
4	413.94	28.78	0.07	0.08	560.48	32.78	0.06
5	251.49	21.7	0.09	0.1	533.35	32.17	0.06
6	154.7	15.72	0.1	0.2	426.12	29.41	0.07
7	96.02	11.24	0.12	0.3	342.37	26.71	0.08
8	59.9	7.77	0.13	0.4	273.02	23.74	0.09
9	37.44	5.36	0.14	0.5	214.95	20.76	0.1
10	23.45	3.72	0.16	0.6	166.52	17.75	0.11
11	14.7	2.53	0.17	0.7	126.18	14.67	0.12
12	9.22	1.75	0.19	0.8	93.15	11.74	0.13
13	5.78	1.2	0.21	0.9	66.28	9.07	0.14
14	3.67	0.85	0.23	1	45.3	6.69	0.15
15	2.29	0.59	0.26	1.2	17.89	3.09	0.17
16	1.44	0.4	0.28	1.4	5.3	1.15	0.22
17	0.92	0.28	0.3	1.6	1.2	0.35	0.29

Based on the experiment results, we can see that our proposed subsampling approach with "de-correlated" strategy can effectively reduce the bias in the estimation of total variation of selected variables. Furthermore, by comparing the average strength of signal explained by the subset of variables, it indicates that Lasso can select a smaller subset of variables with comparable signal strength than the individual test does. Besides the efficiency in terms of variable selection, Lasso approach also wins over individual tests in terms of computational time. On average, the procedure using Lasso approach is 2 times faster than the same procedure using individual tests for variable selection.

We also calculated the 90% CI using the proposed variance estimate in Chapter 3. Results are shown in Table XXXII for all the experiment. Results on other subsampling approaches demonstrate similar patterns and therefore suppressed to save space. From the table, when the βs are random, the coverage is much less than the expected. This is due to the underestimation of the true variance. But for fixed βs, the coverage is relatively small. Possible reasons for the coverage not performing as good as that in independent cases could be the problem with the asymptotic linear approximation for the estimator and/or the boundary effect, i.e., variation estimators are kept positive.

4.4 Application to Real Data

For demonstration purposes, we apply the adaptive subsampling approach to the same eQTL data we used in Chapter 3. In this study, a total of 155 postmortem cerebellum samples were collected from the Stanley Medical Research Institute (SMRI), with 99 males and 56 females. All of them were of European ancestry. The Affymetrix Genome-wide Human SNP 5.0 Array was used for SNP genotyping. The outcome we use here is the gene expression of a probe that hybridizes to a specific genomic region in the chromosome. Correspondingly, the covariates we use are all the SNPs located in that chromosome. For illustrative purpose, we use a probe located at Chromosome 21 which has the smallest number of SNPs among all chromosomes except sex chromosomes. The same probe with ID 8069448 was randomly chosen for the analysis.

TABLE XXXII: 90% CONFIDENCE INTERVAL BY SUBSAMPLING APPROACH IN CORRELATED CASES

Experiment 1			Experiment 2			Experiment 3			Experiment 4		
coverage	90% CI	length	coverage	90% CI	length	coverage	90% CI	length	coverage	90% CI	length
1	(8.58,30.67)	22.09	0.97	(38.99,77.82)	38.83	0.96	(9.76,31.87)	22.11	0.95	(40.28,79.49)	39.21
0.89	(11.55,20.13)	8.58	0.93	(48.7,66.47)	17.77	0.67	(3.59,11.71)	8.12	0.56	(26.84,45.57)	18.73
0.9	(11.76,19.23)	7.48	0.91	(48.99,65.77)	16.78	0.69	(3.21,10.28)	7.07	0.47	(25.38,43.25)	17.87
0.84	(11.98,18.4)	6.42	0.94	(49.3,65.63)	16.33	0.74	(2.88,8.92)	6.04	0.41	(24.68,42.03)	17.35
0.8	(11.87,17.48)	5.61	0.9	(49.64,65.32)	15.68	0.76	(2.48,7.52)	5.04	0.4	(24.39,41.17)	16.78
0.71	(11.56,16.38)	4.82	0.91	(49.9,65.28)	15.38	0.78	(1.91,6.03)	4.12	0.36	(24.07,40.28)	16.21
0.75	(10.93,15.36)	4.43	0.84	(50.92,64.18)	13.26	0.81	(1.51,4.69)	3.18	0.34	(22.15,36.67)	14.52
0.78	(10.19,14.16)	3.97	0.84	(51.64,63.33)	11.69	0.81	(1.01,3.45)	2.44	0.38	(20.11,33.3)	13.19
0.72	(9.24,12.85)	3.61	0.86	(51.98,62.67)	10.69	0.85	(0.61,2.4)	1.79	0.45	(17.81,29.66)	11.85
0.71	(8.18,11.54)	3.37	0.84	(51.96,61.76)	9.8	0.79	(0.3,1.55)	1.25	0.53	(15.46,26.06)	10.6
0.78	(7.04,10.27)	3.22	0.83	(51.52,60.74)	9.22	0.84	(0.12,0.94)	0.82	0.7	(12.95,22.56)	9.61
0.7	(4.77,7.68)	2.91	0.86	(50.78,59.69)	8.91	0.99	(0,0.27)	0.27	0.81	(10.44,18.89)	8.45
0.78	(2.85,5.39)	2.54	0.81	(49.75,58.41)	8.66	1	(0,0.04)	0.04	0.76	(8.09,15.39)	7.3
0.82	(1.51,3.6)	2.09	0.76	(48.2,56.59)	8.39	1	(0,0.01)	0.01	0.74	(5.96,12.18)	6.22
0.9	(0.63,2.22)	1.59	0.77	(46.33,54.69)	8.36	1	(0,0)	0	0.61	(4.09,9.29)	5.2
0.92	(0.19,1.2)	1.01	0.79	(40.91,49.22)	8.31	1	(0,0)	0	0.44	(1.45,4.73)	3.28
1	(0.02,0.33)	0.31	0.81	(33.4,41.75)	8.35	1	(0,0)	0	0.61	(0.29,2.01)	1.72
1	(-0.03,0.06)	0.06	0.82	(24.2,32.6)	8.4	1	(0,0)	0	0.89	(0,0.76)	0.76

Before analyzing the data, a standard quality checking (QC) procedure was performed. SNPs with $MAFs < 0.01$ and HWE $P < 0.001$ were excluded. Because Yang et al's approach does not work when with missing data, subjects were removed if there is any missing value in the SNPs. After the QC process, 130 samples are remained, and the number of SNPs used is 23,862.

We first checked the correlation matrix of all the covariants in the data. It shows that the maximum absolute correlation between two columns is greater than 0.9. Therefore, using the de-correlated Z matrix should be more suitable here. We use the Lasso approach to select the variables, the tuning parameter λ is varied between 0 and 0.1. Table XXXIII displays the results based on our proposed approach. From the table, we can see that when all the SNPs are used for variance estimation, directly using X as the covariates will yield an estimator larger than using Z as the covariates. With variable selection, the estimator from Z is consistently lower than that from X. Furthermore, we can also see that by using Lasso approach for variable selection, the number of X is rapidly decreased.

TABLE XXXIII: SUMMARY OF ANALYSES RESULTS IN eQTL DATA

		Naive		subsampling using X		subsampling using Z	
λ	SNPs	$\hat{\sigma}^2_S using X$	$\hat{\sigma}^2_S using Z$	$\hat{\sigma}^2_S$	90% CI	$\hat{\sigma}^2_S$	90% CI
0	23862	0.55	0.16	0.48	(0.11,0.85)	0.15	(0,0.38)
0.02	125	0.35	0.17	0.28	(0.05,0.51)	0.13	(0,0.29)
0.04	80	0.34	0.17	0.26	(0.10,0.42)	0.11	(0,0.27)
0.06	50	0.37	0.12	0.16	(0,0.32)	0.02	(0,0.18)
0.08	18	0.26	0.12	0.14	(0,0.30)	0.01	(0,0.17)
0.1	12	0.26	0.15	0.03	(0,0.09)	0.01	(0,0.17)

CHAPTER 5

CONCLUSION AND DISCUSSION

Estimating total variation explained by a subset of non-randomly selected covariates in high-dimension regression is an important problem. This problem is more challenging when many weak signals are involved. Yang et al.'s approach, widely used in the estimation of total variation, does not work well when covariates are non-randomly selected. In this thesis, we examined many existing approaches that may provide a possible solution to this problem. The comprehensive simulation studies performed in this research demonstrate that, for those Lasso-related approaches, their performance is mostly constrained by the sparsity and uniform signal strength assumption. In other words, when many non-zero coefficients exist or the coefficient is not strong enough, Lasso-related approaches do not provide a reasonably good estimate of the total variation explained by the select covariates. For approaches that are based on adjusting the bias in individual estimates, their performance is not satisfactory because the accumulative errors and/or the variance of the final estimator can be very large when individual adjusted estimators and combined.

To tackle this problem, we propose subsampling approaches with adjustment to cutoff values for estimating the variation explained by a set of non-randomly selected covariates. We start from the cases with the independent covariates. Simulation results demonstrate that the proposed approach can effectively reduce the bias and the mean squared error over the naive approach. The proposed approach is computationally simple to implement and can be easily

adapted to different selection methods. As a byproduct, the subsampling approach also provides variance estimates for the proposed estimators, which can give a 90% confidence interval with satisfying coverage rate. We also compare the variance estimates from proposed approach with those from EigenPrism approach (Janson et al., 2017). The variance estimates from our proposed approach outperforms EigenPrism approach in yielding an estimate of variance closer to the empirical variance based on simulation. The probable reason for EigenPrism not working well is that this method is mainly reserved for high dimensional setting. Upon variable selection, m ¡ n, and the upper bound of variance estimator will no longer hold for this approach.

The proposed approach is also extended to deal with correlated covariates. The covariate matrix is transformed to remove the correlations among covariates before the total variation is estimated by Yang et al.'s approach. Other variable selection approaches, e.g., l^1-penalized approaches such as Lasso, are also incorporated into the procedure. Simulation study shows that the modified proposed approach can also reduce the bias of the estimators compared with the naive approach. Our simulation study also indicates that Lasso approach is more efficient in selecting covariates than the individual tests when the covariates are correlated.

The research is subject to a number of limitations. First, although we had conducted a simple theoretical analysis of the proposed approach, our conclusions are mainly drawn based on the simulation results. A comprehensive theoretical analysis of the proposed approach is lacking. The limiting spectral distribution theory of a random matrix used by Jiang et al. (2016) for proving the consistency of Yang et al's approach should be useful for further theoretical analysis, which may shed light on whether an optimal subsample ratio exists to achieve the smallest mean

squared error and on further extensions. Second, the proposed approaches substantially reduce the bias of the naive estimator. However, the estimators themselves are still subject to some bias, how to further reduce the bias is of interest. Third, the proposed approach is applied to the real data. But the estimated variations explained by the selected covariates is very small. A better demonstration of the usefulness of the proposed approach may be to data such as the human's height data in Yang et al. (2010). Such an analysis can generate more useful information on causal SNPs responsible for the trait. We have dealt with the quantitative trait in genomic studies. Extension of the proposed approaches to binary traits are of particular interest in practice.

APPENDICES

Appendix A

VALIDATION OF LOW-DIMENSIONAL PROJECTION ESTIMATOR

We check the performance of LDPE under various simulation scenarios. The experiment uses the setting of (Zhang and Zhang, 2014) and includes some extensions based upon that. In (Zhang and Zhang, 2014), n = 200, m = 3000. X follows $N(0, \Sigma)$ with $\Sigma = \rho^{|i-j|}_{m\times m}$; for $j = 1500, 1800, \ldots, 3000$, $\beta_j = 3\lambda_1$, all the other $\beta_j = 3\lambda_1/j^\alpha$ $\lambda_1 = \sqrt{2/n \log m}$. Besides $\alpha = c(1, 2)$ and $\rho = c(0.2, 0.8)$ to evaluate the performance in different strength of the β and dependent levels of X in the original paper, we try smaller $\alpha = (0.5, 0.1)$ to see if LDPE can perform well when lots of small βs present.

TABLE XXXIV: SUMMARY STATISTICS FOR LDPE

	(α, ρ)					
	$(2, 1/5)$	$(1, 1/5)$	$(2, 4/5)$	$(1, 4/5)$	$(0.5, 1/5)$	$(0.1, 1/5)$
bias	0.0354	-0.057	0.0211	-0.0384	-0.114	-0.275
sd	0.142	0.109	0.192	0.239	0.288	2.130
median abs error	0.0964	0.109	0.143	0.157	0.227	1.569
all β_j coverage	0.96	0.96	0.97	0.980	0.97	0.97
maximal β_j coverage	0.94	0.97	0.96	0.99	0.96	0.97

The above Table XXXIV gives the same summary statistics simulation table presented in Zhang and Zhang (2014). For the set-ups of $\alpha = c(1, 2)$ and $\rho = c(0.2, 0.8)$, we obtain

Appendix A (Continued)

similar results showing the superiority of LDPE method in the paper. However, when the signal strengths are weak and the number of signals is large, the bias, standard deviation, and the median absolute error are all increased. Although the coverage rate for all the βs and the maximal β_j suggest that the confidence interval has good coverage, the increase bias indicate the $var(\beta_j)$ is inflated, therefore enlarging the confidence interval to get a good coverage probability.

Appendix B

FUNCTIONAL DE-BIASED ESTIMATOR

To verify the computational procedures of our code and to demonstrate the possible limitations discussed above, simulations are conducted with similar set-up in the (Guo et al., 2016) and additional cases. According to (Guo et al., 2016), n = 400, m = 600, p varies from 25 to 600, all the non-zero βs equal with a flat rate τ, $X \sim N(0, (0.8)^{|i-j|})$, $\epsilon \sim N(0, 1)$, and $Y = X\beta + \epsilon$.

They used the mean squared error (MSE) of the estimated $||\hat{\beta}||_2^2$. Results are displayed in Table XXXV. It can be seen that FDE method works well when size of non-zero βs is small. With the increase of signal strength, the MSE increases adequately under sparsity situation. However, It worsens quickly when the p increases. Specially, when p is large relative to n, the MSE of FDE method is exploded.

TABLE XXXV: MSE OF FDE WITH VARIOUS SETTINGS

	Strength Parameters, τ			
Sparsity Parameter, p	0.1	0.2	0.3	0.4
25	0.032	0.073	0.073	0.081
50	0.118	0.290	0.323	0.364
100	0.470	1.501	2.087	2.136
200	2.190	12.260	31.338	55.194
400	12.658	154.364	730.431	2562.352
600	26.781	455.299	2160.898	6286.643

CITED LITERATURE

Alexander, E., Collins, L., and et al.: Twin and family studies reveal strong environmental and weaker genetic cues explaining heritability of eosinophilic esophagitis. Journal of Allergy and Clinical Immunology, 134:1084–1092, 2014.

Antoniadis, A.: Comments on: l1-penalization for mixture regression models by n. standler, p. buhlmann and s. van de geer. Test, 19:257258, 2010.

Banerjee, O., Ghaoui, L. E., and D'Aspremont, A.: Model selection through sparse maximum likelihood estimation. Journal of Machine Learning Research, 9:485–516, 2007.

Belloni, A., Chernozhukov, V., and et al.: Square-root lasso: Pivotal recovery of sparse signals via conic programming. Biometrika, 98:791806, 2011.

Benjamini, Y. and Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, 57(1):289–300, 1995.

Benjamini, Y. and Hochberg, Y.: On the adaptive control of the false discovery rate in multiple testing with independent statistics. Journal of Educational and Behavioral Statistics, 25:60–83, 2000.

Benjamini, Y., Krieger, A., and Yekutieli, D.: Adaptive linear step-up procedures that control the false discovery rate. Biometrika, 93:491507, 2006.

Benjamini, Y. and Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Annals of Statistics, 29(4):11651188, 2001.

Bickel, P., Ritov, Y., and Tsybakov, A.: Simultaneous analysis of lasso and dantzig selector. Annals of Statistics, 37(4):17051732, 2009.

Bin, R., Janitza, S., Sauerbrei, W., and Boulesteix, A.: Subsampling versus bootstrapping in resampling-based model selection for multivariate regression. Biometrics, 72:272–280, 2006.

Breiman, L.: Better subset regression using the nonnegative garrote. Technometrics, 37(4):373–384, 1995.

Capen, E. C., Clapp, R. V., and Campbell, W. M.: Competitive bidding in high-risk situations. Journal of Petroleum Technology, 23:641–653, 1971.

Chen, C., Scurrah, K., and et al.: Heritability and shared environment estimates for myopia and associated ocular biometric traits: the genes in myopia (gem) family study. Human Genetics, 121:511520, 2007.

Chen, H. Y.: Estimation of strength of detected signals in high-dimensional data. 2016.

Deloukas, P., K., S., and et al.: Large-scale association analysis identifies new risk loci for coronary artery disease. Nature Genetics, 45(1):25–33, 2012.

Dicker, L.: Variance estimation in high-dimensional linear models. Biometrika, 101:269–284, 2014.

Dunn, O.: Estimation of the medians for dependent variables. Annals of Mathematical Statistics, 30(1):192–197, 1959.

Efron, B.: Microarrays, empirical bayes and the two groups model. Statistical Science, 23:1–22, 2008.

Efron, B., Johnstone, I., and et al.: Least angle regression. Annals of Statistics, 32(2):407499, 2004.

Fan, J. and Li, R.: Variable selection via nonconcave penalised likelihood and its oracle properties. Journal of American Statistical Association, 96:1348–1360, 2001.

Friedman, J., Hastie, T., and et al.: Pathwise coordinate optimization. Annals of Applied Statistics, 1(2):302–332, 2007.

Friedman, J., Hastie, T., and et al.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1):1–22, 2010.

Fu, W.: The bridge versus the lasso. Journal of Computational and Graphical Statistics, 7(3):397–416, 1998.

Galton, F.: Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute, 15:246–263, 1886.

Garner, C.: Upward bias in odds ratio estimates from genome-wide association studies. Genetic Epidemiology, 31:288–295, 2007.

Geer, S. V. D.: High dimensional generalized linear models and the lasso. Annals of Statistics, 36:614–645, 2008.

Geer, S. V. D., Buhlmann, P., Ritov, Y., and Dezeure, R.: On asymptotically optimal confidence regions and tests for high-dimensional models. The Annals of Statistics, 42(3):1166–1202, 2014.

Ghosh, A., Zou, F., and Wright, F.: Estimating odds ratios in genome scans: An approximate conditional likelihood approach. The American Journal of Human Genetics, 82(5), May 2008.

Goeman, J.: L-1 penalized estimation in the cox proportional hazards model. Biometrical Journal, 52(2):70–84, 2010.

Griffiths, A., Miller, J., Suzuki, D., and et al.: An introduction to genetic analysis, 7th edition. New York: W. H. Freeman, 2000.

Guo, Z., Wang, W., Cai, T. T., and Li, H.: Optimal estimation of co-heritability in high-dimensional linear models. arXiv preprint, 2016. arXiv:1605.07244.

Jang, K.: The behavioral genetics of psychopathology. Mawah, New Jersey: Lawrence Erlbaum Associates, 2005.

Janson, L., Barber, R. F., and Candes, E.: Eigenprism: inference for high dimensional signal-to-noise ratios. J. R. Statist. Soc. B, 79(4):1017 – 1065, 2017.

Jiang, J., Li, C., Paul, D., Yang, C., and Zhao, H.: High-dimensional genome-wide association study and misspecified mixed model analysis. The Annals of Statistics, 10 2016.

Lee, S., DeCandia, T., Ripke, S., and et al.: Estimating the proportion of variation in susceptibility to schizophrenia captured by common snps. Nature Genetics, 44:247–250, 2012.

Listgarten, J., Lippert, C., and et al.: Improved linear mixed models for genome-wide association studies. Nature Methods, 9:525526, 2012.

Meinshausen, N. and Buhlmann, P.: Stability selection (with discussion). J. R. Statist. Soc. B, 72:417–473, 2010.

Naomi, R. and Visscher, P.: Estimating trait heritability. Nature Education, 1(1):29, 2008.

Poderman, T., Benyamin, B., and et al.: Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nature Genetics, 47(7):702–709, 2015.

Politis, D. D. and Romano, J. P.: Large sample confidence regions based on subsamples under minimal assumptions. Annals of Statistics, 22:2031–2050, 1994.

Render, R., Plomin, R., and Vandenberg, S.: Who discovered the twin method? Behavior Genetics., 20(2):277285, 1990.

Shao, J.: The efficiency and consistency of approximations to the jackknife variance estimators. J. Amer. Statist. Assoc., 84:114–119, 1989.

Shao, J. and Wu, J.: A general theory for jackknife variance estimation. Annals of Statistics, 17:1176–1197, 1989.

Simes, R.: An improved bonferroni procedure for multiple tests of significance. Biometrika, 73(3):751754, 1986.

Stadler, N., Buhlmann, P., and van de Geer, S.: l1-penalization for mixture regression models (with discussion). Test, 19:209285, 2010.

Storey, J.: A direct approach to false discovery rates. Journal of the Royal Statistical Society, 64(3):479498, 2002.

Strachan, T. and Read, A.: Human molecular genetics (4th ed.). page 467495, 2010.

Sun, T. and Zhang, C.: Scaled sparse linear regression. Biometrika, 99:879–898, 2012. MR2999166.

Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B(58):267288, 1996.

Tibshirani, R.: The lasso method for variable selection in the cox model. Statistics in Medicine, 16:385395, 1997.

Tibshirani, R., Michael, S., and et al.: Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society, 67(1):91–108, 2005.

Vattikuti, S., Guo, J., and Chow, C. C.: Heritability and genetic correlations explained by common snps for metabolic syndrome traits. PLoS Genetics, 8(e1002637), 2012.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., and Nyholt, D. R.: Common snps explain a large proportion of the heritability for human height. Nature Genetics, 42:565–569, 6 2010.

Yang, J., Manolio, T., and et al: Genome partitioning of genetic variation for complex traits using common snps. Nature Genetics, 43(6):519–525, 6 2011.

Yuan, M. and Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, 68(1):49–69, 2006.

Zhang, C. and Zhang, S.: Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society, Series B Statistical Methodology, 78:217–242, 2014.

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

Zhong, H. and Prentice, R. L.: Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. Biostatistics, 9(4):621–634, 2008.

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

Zuo, H. and Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B, 67:301–320, 2005.

VITA

NAME: Yaru Shi

EDUCATION: Bachelor of Science, Biotechnology, Beijing Institute of Technology, 2011

Master of Science, Biostatistics, Georgetown University, 2012

Doctor of Philosophy, Biostatistics, University of Illinois at Chicago, 2018

PUBLICATIONS: Shi, Y., Kim, Y., Kostygina, G., Emery, S.: Efficient Sampling Strategy for SVM Through Semi-Supervised Active Learning. JSM Proceedings, Section on Statistical Learning and Data Science, 2016.

Kostygina, G., Tran, H., Shi, Y., Kim, Y., Emery, S.: Sweeter Than a Swisher: amount and themes of little cigar and cigarillo content on Twitter. Tobacco Control. 25: 75-82, 2016.

Kim, Y., Kornfield, R., Shi, Y., Vera, L., Daubresse, M., Alexander, G., Emery, S.: Effects of Televised Direct-to-Consumer Advertising for Varenicline on Prescription Dispensing in the United States, 2006-2009. Nicotine and Tobacco Research. 18(5): 1180-1187, 2015.

Sun, B., Wang, Y., Kota, K., Shi, Y., Zheng, Y.: Telomere length variation: A potential new telomere biomarker for lung cancer risk. Lung Cancer, 88: 297-303, 2015.

Oppong, B., Makambi, K., Shi, Y.: Stage of Presentation at Initial Breast Cancer Diagnosis: Does Race Remain a Factor? The Breast Journal, 21(4), 445-446, 2015.