ABSTRACT

MARTINEZ, KARA ANN. Robust Methods in Kernel Association Testing. (Under the direction of Arnab Maity.)

In genetic data analyses and genome-wide association studies, testing for relationships between a trait and set of genetic markers, such as SNPs, is a common task. One class of methods, kernel association tests, models the genetic marker set through a nonparametric function and applies kernel machine regression techniques. As described in Chapter 2, the kernel machine literature is wide-ranging with many application areas; however, many kernel-based hypothesis tests incorporate least squares estimates or asymptotic distributions that rely upon Gaussian errors. These normality assumptions may be inappropriate for certain types of traits, such as antibody levels, that often have skewed conditional distributions. Unfortunately, the area of robust kernel association testing is relatively unexplored. In this thesis, we apply ideas from robust regression methods to develop kernel-based hypothesis tests that do not require any specific error distribution.

In Chapter 3, we propose a general and robust hypothesis test to detect associations between genetic marker sets and a univariate, continuous response. In developing our robust kernel association test (RobKAT), we construct a score-type statistic with a flexible choice of loss function when estimating parameters. Rather than requiring the use of least squares, RobKAT allows alternative loss functions that are not tailored to a normal distribution. We evaluate RobKAT across various data distributions through simulation study. When errors are normally distributed, RobKAT controls type I error and shows comparable power to normality-based methods, such as the sequence kernel association test (SKAT). In all other distributional settings investigated, our robust test has similar or greater power than SKAT. Finally, we apply our robust testing method to data from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) study to detect associations between selected genes including the Major Histocompatibility Complex (MHC) region on chromosome six and neurotropic herpesvirus antibody levels in schizophrenia patients. RobKAT detected significant association with four SNP-sets (*HIST1H2BJ*, MHC, *POM121L2*, and *SLC17A1*), three of which were undetected by SKAT.

In Chapter 4, we extend the RobKAT methodology from testing a univariate, quantitative trait to a multivariate response that contains continuous, binary, discrete, or mixed phenotypes. We develop

this multivariate robust kernel association test (MrKAT) by implementing general loss functions to appropriately weight skewed error distributions and influential observations. We observe through simulation study that the MrKAT approach adequately controls type I error and tends to be more powerful than ordinary least squares-based multivariate methods when marginal responses are heavy-tailed or skewed. Simulation results also suggest that MrKAT is powerful when the data contain outlying observations of high influence or leverage. Using data donated by schizophrenia patients in the CATIE study, we test possible associations between SNP-sets in the MHC region and three neurotropic herpesvirus antibody levels. For these continuous and skewed antibody levels, MrKAT detected five significant gene regions (*BTN2A1, BTN2A2*, MHC, *POM121L2, SLC17A1*), while ordinary least squares-based methods failed to detect any significant genes.

While RobKAT and MrKAT test for associations between genes and phenotypes, researchers may also wish to consider gene-environment interactions in these tests. Many methods for overall genetic marker effects, however, depend on the assumption of normally distributed errors. For example, the composite kernel likelihood ratio test (CKLRT) requires normality to construct a likelihood. To address this issue, the robust composite kernel association test (RoCKAT) developed in Chapter 5 modifies the robust strategies of previous chapters to gene-environment interaction models. Like our previous methods, RoCKAT uses a variance component score test statistic that estimates parameters with a general, robust loss function. Unlike the CKLRT, we propose a permutation-based algorithm that approximates the null distribution by permuting within exposure groups. By doing so, we create a method that does not rely on known error distributions and constructs a permutation test without violating the exchangability condition.

As detailed in the simulation studies of Chapter 5, the CKLRT has inflated type I error when the distributions are skewed, while RoCKAT controls the type I error. When the errors follow a Gaussian distribution, the CKLRT is more powerful overall; however, the power of RoCKAT is similar to that of the CKLRT when a genetic main effect is present. When applying our RoCKAT software to the CATIE study, we analyze the association between cognitive scores and the MHC region, while accounting for interactions with herpesvirus exposures. We find that RoCKAT detects a significant relationship between cognitive reasoning, a skewed trait, and the MHC region, which literature shows has links to schizophrenia. The only significant association detected with the CKLRT is the *SLC17A3* gene.

As discussed in Chapter 6, our strategy of applying robust regression techniques to kernel association tests has the potential to solve a variety of analysis problems. While this thesis has developed three methods, we believe that there are many additional avenues to explore in the future.

Robust Methods in Kernel Association Testing

by
Kara Ann Martinez

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

Statistics

Raleigh, North Carolina

2020

APPROVED BY:

Jung-Ying Tzeng	Wenbin Lu
Daowen Zhang	Arnab Maity Chair of Advisory Committee

DEDICATION

To my wonderful family for their encouragement, love, and support.

BIOGRAPHY

Kara Martinez was born in southeast Michigan to her parents, Mark and Laurie, who always stressed the importance of academics and hard work. During her undergraduate years at the University of Michigan, she had the time to explore these academic interests further. Scribbling proofs while on the bus and asking too many questions during office hours, Kara began to discover her love of mathematics and statistics; however, what to do with these interests were unclear until she spent a summer in North Carolina at the Summer Institute for Biostatistics (SIBS) program. With the Carolina heat, statistics seminars, and community of students passionate about the same things, Marie Davidian inspired us all to become statisticians.

It was then in August 2015 that Kara began the crazy journey that is the statistics PhD, and these five years quickly became a strange collection of both joyous moments and stress-induced breakdowns. Kara made wonderful friends over Zumba classes, fell in love on the 4th floor of SAS, played Catan until 3am, and sampled every flavor of Howling Cow ice cream. In addition to these wonderful memories, Kara somehow was paid to study the subjects that she loved most. At the same time, graduate school was filled with days that felt like a stressful decent into madness. There were many late nights, buggy code, indecipherable papers, high-stress exams, and every step of the PhD was both a mental and intellectual challenge.

Kara has accepted a clinical statistician position at GlaxoSmithKline in Collegeville, Pennsylvania. Although the future is always uncertain, she is excited to explore this new chapter of her life.

ACKNOWLEDGEMENTS

Firstly, I have an immense amount of gratitude for my wonderful advisor, Arnab Maity, who has helped me grow into the statistician that I am today. I'm thankful for his kind patience in answering my questions, working through research problems, and giving me the freedom to name my methods with silly acronyms. I also would like to thank my advisory committee: Jung-Ying Tzeng, Wenbin Lu, and Daowen Zhang. My dissertation would not be the same without their thoughtful comments and advice. I am thankful, too, for the collaboration with my co-authors of "Robust Kernel Association Testing (RobKAT)," Patrick Sullivan and Robert Yolken.

I am grateful to the entire NC State statistics department, for both teaching and supporting me throughout my PhD journey. In particular, I would like to thank Spencer Muse for providing funding through his NIH training grant, Biostatistics Training in the Omics Era. I'm also thankful to the statistics department moms, Lanakila Alexander and Alison McCoy, for providing support and keeping us all sane.

I am also incredibly lucky to have my amazing parents and family, who have always encouraged me to make education a priority. I am thankful for their love, support, and unwavering belief that I would make it to this point, despite the hundreds of times that I doubted myself throughout this process.

To my wonderful boyfriend, David Hu, I am so thankful to have had you by my side throughout this journey. You have always supported and believed in me through my many stress-related breakdowns, even when it meant sleeping under a desk until I finished writing my analysis report. I'm so happy to have you as my best friend, and I love you so much.

I also need to thank my friends and especially, Katherine Allen, for always lending an ear to listen and for going through this crazy, grad school adventure with me. Additionally, I send a huge "thank you" to all the people at NC State who kept me going. You know who you are.

TABLE OF CONTENTS

LIST OF TABLES . . . . **vii**

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

Chapter 1 Nontechnical Introduction **1**

Chapter 2 Technical Introduction and Literature Review **5**
- 2.1 Introduction 5
- 2.2 Kernel Machine Regression 7
- 2.3 Hypothesis Testing with a Univariate Response 9
  - 2.3.1 Least Squares Kernel Machine 9
  - 2.3.2 Quantile Regression Kernel Machine 11
  - 2.3.3 Logistic and GLM Kernel Machines 13
- 2.4 Hypothesis Testing with a Multivariate Response 14
  - 2.4.1 Normality-Based Methods 15
  - 2.4.2 Robust Methods 16
  - 2.4.3 Methods for Rare Variant Covariates 17
  - 2.4.4 Methods for Pleiotropic Relationships 17
- 2.5 Hypothesis Testing with Gene-Environment Interaction 18
- 2.6 Other Applications in KMR Hypothesis Testing 20
  - 2.6.1 Microbiome Data 20
  - 2.6.2 Copy Number Variation Data 21
- 2.7 Discussion 22

Chapter 3 Robust Kernel Association Testing **24**
- 3.1 Introduction 24
- 3.2 Methodology 27
  - 3.2.1 Model Specification and Testing Problem 27
  - 3.2.2 Test Statistic 28
  - 3.2.3 Null Distribution of the Test Statistic 31
- 3.3 Simulation Study 31
  - 3.3.1 Simulation Study Design 32
  - 3.3.2 Simulation Results 33
- 3.4 Application to CATIE Herpesvirus Data 36
- 3.5 Discussion 39

Chapter 4 Multivariate Robust Kernel Association Testing with Continuous, Binary, and Mixed Responses **43**
- 4.1 Introduction 43
- 4.2 Methods 46
  - 4.2.1 Univariate Continuous Response 46
  - 4.2.2 Multivariate Continuous Response 48
  - 4.2.3 Multivariate Binary Responses 50
  - 4.2.4 Multivariate Mixed Responses 51
- 4.3 Simulation Study 51
  - 4.3.1 Non-normal Responses 51

4.3.2 High Leverage Observations . . . 53
4.4 Application to CATIE Herpesvirus Data . . . 56
4.5 Discussion . . . 59

Chapter 5 Robust Composite Kernel Association Testing . . . 63
5.1 Introduction . . . 63
5.2 Methods . . . 66
5.2.1 Model and Notation . . . 66
5.2.2 Modeling Functions with Kernels . . . 66
5.2.3 RoCKAT Statistic . . . 68
5.2.4 Null Distribution and Inference . . . 70
5.3 Simulation Study . . . 72
5.3.1 Simulation Study Design . . . 72
5.3.2 Simulation Results . . . 73
5.4 Application to CATIE Cognitive Scores Data . . . 75
5.5 Discussion . . . 79

Chapter 6 Future Research . . . 82

References . . . 85

APPENDICES . . . 94
Appendix A Supplemental Information for Chapter 3 . . . 95
A.1 Type I Error Results for Additional Kernels and Loss Functions with n=100 . . . 96
A.2 Type I Error Results for Additional Kernels and Loss Functions with n=300 . . . 97
A.3 Power Results for Additional Kernels and Loss Functions with n=100 . . . 98
A.4 Power Results for Additional Kernels and Loss Functions with n=300 . . . 100
A.5 Full CATIE Data Analysis . . . 102
A.6 RobKAT Naive χ^2 Mixture Null Distribution . . . 105
A.7 High-Dimensional SNP-set Simulation with with n=300 . . . 107
Appendix B Supplemental Information for Chapter 4 . . . 109
B.1 Power Results for the Standard and Robust Settings with Nonlinear h . . . 110
B.2 CATIE Data Analysis on Binary Responses . . . 111
B.3 Influence and Leverage Measurements . . . 112
B.4 MrKAT Time of Computation . . . 113
B.5 Simulation Results for the Non-normal Setting with 200 Responses . . . 114
B.6 Simulation Results for the Non-normal Setting with 25 Responses . . . 116
B.7 Percent Contribution Exploration . . . 118
B.8 Simulation Results in the Non-normal Setting with p=787 SNPs . . . 120
Appendix C Supplemental Information for Chapter 5 . . . 121
C.1 CATIE Data Analysis Results (HSV-1) . . . 122
C.2 CATIE Data Analysis Results (CMV) . . . 123
C.3 Power Results with Continuous E . . . 124

LIST OF TABLES

Table 3.1 Simulation study for type I error of RobKAT (Huber loss), QRKM (LAD loss), and SKAT. The simulation includes IBS kernel, sample size $n = 100$, and 9 simulated SNPs per sample. Estimates show proportion of $10,000$ iterations with p-value below α cutoff. 33
Table 3.2 Simulation study for type I error of RobKAT (Huber loss), QRKM (LAD loss), and SKAT. The simulation includes IBS kernel, sample size $n = 100$, and 787 simulated SNPs per sample. Estimates show proportion of $10,000$ iterations with p-value below α cutoff. 34
Table 3.3 Data analysis of CATIE data using RobKAT (huber loss), QRKM (LAD loss), and SKAT. The significance threshold was determined via a Bonferroni correction based on 12 gene analyses, 3 kernels, and 3 methods ($\alpha = 0.05/(12*3*3) = 0.00046$). Significant results are shown in bold. 39

Table 4.1 Simulation study for type I error of MrKAT and DKAT the non-normal setting. The simulation includes four continuous responses, various error distributions, IBS genetic kernel, and a sample size of 100. Estimates show the proportion of 10,000 iterations with p-value below the α cutoff. 53
Table 4.2 Simulation study for the type I error rate of MrKAT and DKAT in the Standard (no high leverage points) and Robust (1 high leverage point) settings. The simulation includes IBS genetic kernel and sample size of 100. Estimates show the proportion of 10,000 iterations with p-value below the α cutoff. . . . 55
Table 4.3 Applying MrKAT and DKAT to the continuous form of HSV-1, HSV-2, and CMV antibody levels. The bold values highlight significant p-values after a Bonferroni correction ($0.05/(2*3*12) = 0.00069$) for the 2 methods, 3 kernels, and 12 genes. 58

Table 5.1 Type I error simulation showing the proportion of tests with p-value less than $\alpha = 0.05$. The simulation uses IBS genetic marker kernel, sample size $n = 100$, and $1,000$ iterations at each setting of β. RoCKAT used 500 permutations per test, CKLRT used $N = 5000$ generated variables for the empirical null distribution of the LRT. Data generated using the MHC gene ($p = 787$ SNPs). Bold numbers represent values that fall 1.96 standard errors or more above the $\alpha = 0.05$ target. 73
Table 5.2 Data analysis of CATIE cognitive score data using RobKAT (Huber loss) and the CKLRT, while incorporating possible gene-environment interaction with HSV-2 exposure. The significance threshold was determined via a Bonferroni correction based on 12 gene analyses and 2 methods (significance level = $0.05/24 = 0.00208$). Significant results are shown in bold. 79

Table A.1 Simulation study for type I error that includes sample size $n = 100$, 9 SNPs, and IBS kernel (mirroring Table 1) as well as linear and quadratic kernels. The study compares RobKAT (Huber loss), QRKM (LAD loss), and SKAT as well as tests with redescending loss functions, RobKAT (Bisquare loss) and RobKAT (Hampel loss). Type I errors are identical for any $h(\boldsymbol{Z}_i)$ function. Estimates show proportion of $10,000$ iterations with p-value below α cutoff. 96

Table A.2 Simulation study for type I error with the larger sample size of $n = 300$. The simulation includes IBS kernel and 9 SNPs (mirroring Table 1) as well as linear and quadratic kernels. Methods tested include RobKAT (Huber loss), QRKM (LAD loss), and SKAT as well as tests with redescending loss functions, RobKAT (Bisquare loss) and RobKAT (Hampel loss). Estimates show proportion of $10,000$ iterations with p-value below α cutoff. 97

Table A.3 RobKAT analysis on CATIE antibodies, mirroring Section 4 and Table 3 in the main text. Here, we include redescending loss functions (Hampel, Bisquare) in addition to Huber loss, LAD loss, and SKAT. The significance threshold was determined via a Bonferroni correction based on 12 gene analyses, 3 kernels, and 5 methods ($\alpha = 0.05/(12*3*5) = 0.00028$). Significant results are shown in bold. 102

Table A.4 RobKAT analysis on CATIE antibodies as described in Section 4, but instead using log-transformed antibody responses and including redescending loss functions (Hampel, Bisquare). The significance threshold was determined via a Bonferroni correction based on 12 gene analyses, 3 kernels, and 5 methods ($\alpha = 0.05/(12*3*5) = 0.00028$). Significant results are shown in bold. 103

Table A.5 Type I error simulations naively using the LSKM null distribution with the RobKAT statistic. The LSKM null distribution of the test statistic for the overall genetic marker effect follows a Chi-squared mixture, which is approximated using a Satterthwaite approximation (Liu et al. 2007). The simulation includes IBS kernel, Huber loss, sample sizes $n \in \{25, 50, 100\}$, $p = 9$ SNPs, and $1,000$ iterations at each setting of c. Data was generated from $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + c * h(\boldsymbol{Z}_i) + \epsilon_i$ with ϵ_i iid from t_3, χ_1^2, $N(0,1)$, and $Cauchy(0,1)$. The linear case is $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \mathbf{1}_9$, and the nonlinear case is $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{p} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{p} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$. When using this naive approach with incorrect null distribution, Type I error is inflated. 106

Table A.6 Simulation study for type I error of RobKAT (Huber loss), QRKM (LAD loss), and SKAT in the setting of 787 simulated SNPs. This mirrors the simulation in Section 3.1 for $p = 787$ in the main text (Table 2) but with larger sample size ($n = 300$). The simulation includes IBS kernel, and estimates show proportion of $10,000$ iterations with p-value below α cutoff. 108

Table B.1 Applying MrKAT and DKAT to the binary form of HSV1, HSV2, and CMV antibody levels. The bold values highlight significant p-values after a Bonferroni correction ($0.05/(2*3*12) = 0.00069$) for the 2 methods, 3 kernels, and 12 genes. 111

Table B.2 Approximate influence of added outlier to simulation study described in Section 4.3.2. Simulated Cook's Distance and leverage were based on 1000 iterations mirroring simulation study settings of nonlinear $h_j(\boldsymbol{Z}_i)$ for $j \in \{1,2,3,4\}$, four responses, and $c = 0.5$. Leverage is only dependent on covariate matrix, $\boldsymbol{X}_i$. 112

Table B.3 Maximum influence (measured by Cook's Distance to linear or generalized linear model) and maximum leverage for the CATIE antibody data analyzed in Section 4.4. Since leverage values sum to q = # Covariates = 2 and the expected leverage value would be $q/n = 0.0028$, we do not see evidence of extreme leverage observations in this data set. These metrics also do not show strong evidence of highly influential observations. 112

Table B.4 Simulation study for type I error exploring the testing properties from 200 continuous, multivariate responses that have errors that are symmetric ($N(0,1)$), skewed (χ^2_1), heavy-tailed (t_3), and substantially heavy-tailed (Cauchy). Of the 200 responses, five were associated ($Y_{ij} = \boldsymbol{X}_i\beta + c * h_j(\boldsymbol{Z}_i) + \epsilon_{ij}$) and 195 were not associated ($Y_{ij} = \epsilon_{ij}$) with the covariates. Data was generated with IBS kernel, sample size $n = 100$, $\beta = -5\mathbf{1}_4$, and estimates averaged over 1000 iterations. 115

Table B.5 Simulation study for type I error exploring the testing properties from 4 continuous, multivariate responses that have symmetric ($N(0,1)$), skewed (χ^2_1), heavy-tailed (t_3), and substantially heavy-tailed (Cauchy) error distributions. Of the 4 responses, all were generated with $Y_{ij} = \boldsymbol{X}_i\beta + c * h_j(\boldsymbol{Z}_i) + \epsilon_{ij}$. Data was generated with IBS kernel, sample size $n = 25$, $\beta = -5\mathbf{1}_4$, and estimates averaged over 10,000 iterations. 117

Table B.6 Simulated T_j/T for each of $j = 1,2,3$ continuous responses, averaged over 1000 iterations. Data was generated from $Y_j = \boldsymbol{X}_i\beta + c_j h_j(\boldsymbol{Z}_i) + \epsilon_{ij}$ for $i = 1,2,...,n$ samples and $\boldsymbol{Z}_i$ a randomly sampled 9-SNP from the *SLC17A1* gene. Independent errors represent $\epsilon_{ij} \sim N(0,1)$, and correlated ϵ_{ij} is the $(i,j)^{th}$ entry of E^*, where $E_{ij} \sim N(0,1)$ and the errors have the correlation structure . 118

Table B.7 Simulated T_j/T for each of $j = 1,2,3$ continuous responses, averaged over 1000 iterations. Data was generated from $Y_j = \boldsymbol{X}_i\beta + c_j h(\boldsymbol{Z}_i) + \epsilon_{ij}$ for $i = 1,2,...,n$ samples, $X_{ij} \sim N(0,1)$, $\beta = -5\mathbf{1}_2$, $\boldsymbol{Z}_i$ a randomly sampled 9-SNP from the *SLC17A1* gene, $h_1(\boldsymbol{Z}_i) = h_2(\boldsymbol{Z}_i) = h_3(\boldsymbol{Z}_i) = 2z_1 + 3z_2 + z_3 + 2z_4 + z_5$, independent errors from $\epsilon_{ij} \sim N(0, \sigma^2_j)$, and an IBS kernel. To vary spreads of the different responses, $\sigma_1 = 1$, $\sigma_2 = 5$, and $\sigma_3 = 10$. 119

Table C.1 Data analysis of CATIE cognitive score data using RobKAT (Huber loss) and the CKLRT, while incorporating possible gene-environment interaction with HSV-1 exposure. The significance threshold was determined via a Bonferroni correction based on 12 gene analyses and 2 methods ($\alpha^* = 0.05/24 = 0.00208$). Significant results are shown in bold. 122

Table C.2 Data analysis of CATIE cognitive score data using RobKAT (Huber loss) and the CKLRT, while incorporating possible gene-environment interaction with CMV exposure. The significance threshold was determined via a Bonferroni correction based on 12 gene analyses and 2 methods ($\alpha^* = 0.05/24 = 0.00208$). Significant results are shown in bold. 123

Table C.3 Power simulations of ROCKAT and the CKLRT. The number of causal SNPs (k) is 20, and no covariates. Data generated with $E_i \sim N(0,1)$ using the MHC gene ($p = 787$ SNPs) with $h_G(\boldsymbol{G}_i) = \sum_{i=1}^{k} G_{ij}$, and $h_{GE}(\boldsymbol{G}_i, E_i) = \sum_{i=1}^{k} G_{ij} E_i$. The simulation tests use $E^*_i = I\{E_i \leq med(\boldsymbol{E})\}$, IBS genetic marker kernel, sample size $n = 100$, and $1,000$ iterations at each setting of β. ROCKAT used 500 permutations per test, CKLRT used $N = 5000$ generated variables for the empirical null distribution of the LRT. There is low power within this β window, as the K_{GE} built on the binary representation does not model the data with continuous E_i well. 124

LIST OF FIGURES

Figure 1.1 Histograms of data generated from various distributions. 2

Figure 3.1 Power simulations of RobKAT (black solid line), SKAT (blue dashed line), and LAD (red dotted line). The simulation uses IBS kernel, sample size $n = 100$, $p = 9$ SNPs, and $1,000$ iterations at each setting of c. The top row corresponds to the linear case $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \mathbf{1}_9$, and the bottom row corresponds to the nonlinear case $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{p} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{p} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$. 35

Figure 3.2 Power simulations with bimodal errors using RobKAT (black solid line), SKAT (blue dashed line), and LAD (red dotted line). The simulation uses IBS kernel, sample size $n = 100$, $p = 9$ SNPs, and $1,000$ iterations at each setting of c. Figures (a) and (b) show the linear case $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \mathbf{1}_9$, and figures (c) and (d) display nonlinear case $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{p} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{p} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$. 35

Figure 3.3 Power simulations of RobKAT (black solid line), SKAT (blue dashed line), and LAD (red dotted line). The simulation uses IBS kernel, sample size $n = 100$, $1,000$ iterations at each setting of c, 787 simulated SNPs per sample, and 2 causal SNPs. The top row corresponds to the linear case $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_{i1} + \boldsymbol{Z}_{i2}$, and the bottom row corresponds to the nonlinear case $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{2} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{2} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$. 36

Figure 3.4 Power simulations of RobKAT (black solid line), SKAT (blue dashed line), and LAD (red dotted line). The simulation uses IBS kernel, sample size $n = 100$, $1,000$ iterations at each setting of c, 787 simulated SNPs per sample, and two causal SNPs. Figures (a) and (b) show the linear case $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_{i1} + \boldsymbol{Z}_{i2}$, and figures (c) and (d) display nonlinear case $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{2} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{2} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$. 36

Figure 4.1 Power simulation results for four continuous, independent, multivariate responses that are symmetric $N(0,1)$, skewed (χ_1^2), heavy-tailed (t_3), and substantially heavy-tailed (Cauchy). MrKAT is shown in black, and DKAT is shown in red. The simulation implements IBS kernel, n=100, $h_1(\boldsymbol{Z}_i) = h_2(\boldsymbol{Z}_i) = h_3(\boldsymbol{Z}_i) = h_4(\boldsymbol{Z}_i) = 2Z_{i1} + 3Z_{i2} + Z_{i3} + 2Z_{i4} + Z_{i5}$, and $\boldsymbol{\beta} = -5\mathbf{1}_4$. 52

Figure 4.2 Power simulation results from the standard setting (no high leverage points). MrKAT is shown in black, and DKAT is shown in red. This simulation included 4 Responses, and IBS kernel, n=100, $h_1(\boldsymbol{Z}_i) = h_2(\boldsymbol{Z}_i) = h_3(\boldsymbol{Z}_i) = h_4(\boldsymbol{Z}_i) = 2Z_{i1} + 3Z_{i2} + Z_{i3} + 2Z_{i4} + Z_{i5}$, and $\boldsymbol{\beta} = -5\mathbf{1}_4$ 55

Figure 4.3 Power simulation results from the robust setting (1 high leverage points). MrKAT is shown in black, and DKAT is shown in red. This simulation included 4 Responses, and IBS genotype kernel, sample size $n = 100$, $h_1(\boldsymbol{Z}_i) = h_2(\boldsymbol{Z}_i) = h_3(\boldsymbol{Z}_i) = h_4(\boldsymbol{Z}_i) = 2Z_{i1} + 3Z_{i2} + Z_{i3} + 2Z_{i4} + Z_{i5}$, and $\boldsymbol{\beta} = -5\mathbf{1}_4$. For the influential observation, we set $Y_{100} \sim N(10,1)$ or $Y_{100} = 1$, $X_{100} \sim N(3,1)$. . 56

Figure 5.1 Power simulations of RoCKAT (1: black solid line) and CKLRT (2: red dashed line). The simulation uses IBS genetic marker kernel, sample size $n = 100$, and $1,000$ iterations at each setting of β. RoCKAT used 500 permutations per test, CKLRT used $N = 5000$ generated variables for the empirical null distribution of the LRT. The number of causal SNPs (k) is 20, and $\boldsymbol{X}_i \sim N(0, \mathbb{I}_2)$. Data generated using $E_i \in \{0,1\}$ and the MHC gene ($p = 787$ SNPs) with $h_G(\boldsymbol{G}_i) = \boldsymbol{G}_i^T \mathbf{1}_k$, and $h_{GE}(\boldsymbol{G}_i, \boldsymbol{E}_i) = \sum_i^k G_{ik} E_i$. 74
Figure 5.2 Power simulations of ROCKAT (1: black solid line) and CKLRT (2: red dashed line). The simulation uses IBS genetic marker kernel, sample size $n = 100$, and $1,000$ iterations at each setting of β. ROCKAT used 500 permutations per test, CKLRT used $N = 5000$ generated variables for the empirical null distribution of the LRT. The number of causal SNPs (k) is 20, and $\boldsymbol{X}_i \sim N(0, \mathbb{I}_2)$. Data generated using $E_i = [E_i^{(2)}\ E_i^{(3)}]^T$ as a 3-level factor variable and the MHC gene ($p = 787$ SNPs) with $h_G(\boldsymbol{G}_i) = \boldsymbol{G}_i^T \mathbf{1}_k$, and $h_{GE}(\boldsymbol{G}_i, \boldsymbol{E}_i) = \sum_i^k G_{ik}(E_i^{(2)} + 2E_i^{(3)})$. 75
Figure 5.3 Power simulations of ROCKAT (black) and CKLRT (grey). The simulation uses IBS genetic marker kernel, sample size $n = 100$, and $1,000$ iterations at each combination of parameters. ROCKAT used 500 permutations per test, CKLRT used $N = 5000$ generated variables for the empirical null distribution of the LRT. The number of causal SNPs (k) is 20. Data generated using the MHC gene ($p = 787$ SNPs) and $\boldsymbol{E}_i$ is a scalar in $\{0,1\}$. 76
Figure 5.4 Power simulations of ROCKAT (black) and CKLRT (grey). The simulation uses IBS genetic marker kernel, sample size $n = 100$, and $1,000$ iterations at each combination of parameters. ROCKAT used 500 permutations per test, CKLRT used $N = 5000$ generated variables for the empirical null distribution of the LRT. The number of causal SNPs (k) is 20. Data generated using the MHC gene ($p = 787$ SNPs) and $\boldsymbol{E}_i$ is a 3-level factor variable. 77
Figure 5.5 Distribution of cognitive functioning standardized scores, separated by the five domains explored in the CATIE study. 78

Figure A.1 Power simulations with linear and quadratic kernels of RobKAT Huber (1:black solid line), SKAT (2:blue dashed line), and LAD (3:red dotted line) as well as tests with redescending loss functions, RobKAT Bisquare (4:magenta dot-dashed line) and RobKAT Hampel (5: green long-dash line). The simulation uses sample size $n = 100$, $p = 9$ SNPs, and $1,000$ iterations at each setting of c. The linear case is $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \mathbf{1}_9$, and the nonlinear case is $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^p \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^p \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$. 98
Figure A.2 Power simulations with bimodal errors that include linear and quadratic kernels of RobKAT Huber (1:black solid line), SKAT (2:blue dashed line), and LAD (3:red dotted line), as well as tests with redescending loss functions, RobKAT Bisquare (4: magenta dot-dashed line) and RobKAT Hampel (5: green long-dash line). The simulation uses sample size $n = 100$, $p = 9$ SNPs, and $1,000$ iterations at each setting of c. The linear case is $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \mathbf{1}_9$, and the nonlinear case is $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^p \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^p \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$. 99

Figure A.3 Power simulations with linear and quadratic kernels of RobKAT Huber (1:black solid line), SKAT (2:blue dashed line), and LAD (3:red dotted line) as well as tests with redescending loss functions, RobKAT Bisquare (4:magenta dot-dashed line) and RobKAT Hampel (5: green long-dash line). The simulation uses sample size $n = 300$, $p = 9$ SNPs, and $1,000$ iterations at each setting of c. The linear case is $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \boldsymbol{1}_9$, and the nonlinear case is $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{p} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{p} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$. 100
Figure A.4 Power simulations with bimodal errors that include linear and quadratic kernels of RobKAT Huber (1:black solid line), SKAT (2:blue dashed line), and LAD (3:red dotted line) as well as tests with redescending loss functions, RobKAT Bisquare (4: magenta dot-dashed line) and RobKAT Hampel (5: green long-dash line). The simulation uses sample size $n = 300$, $p = 9$ SNPs, and $1,000$ iterations at each setting of c. The linear case is $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \boldsymbol{1}_9$, and the nonlinear case is $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{p} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{p} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$. 101
Figure A.5 Distribution of the continuous responses from the CATIE study. Before log-transformation (a-c), all distributions are notably skewed. After log-transformation (d-f), all distributions appear bimodal. This suggests a need for a robust testing method. 104
Figure A.6 Power simulations naively using the LSKM null distribution with the RobKAT statistic. The LSKM null distribution of the test statistic for the overall genetic marker effect follows a Chi-squared mixture, which is approximated using a Satterthwaite approximation (Liu et al. 2007). The simulation includes IBS kernel, Huber loss, sample sizes $n \in \{25, 50, 100\}$, $p = 9$ SNPs, and $1,000$ iterations at each setting of c. Data was generated from $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + c * h(\boldsymbol{Z}_i) + \epsilon_i$ with ϵ_i iid from t_3 (1: black), χ_1^2 (2: red), $N(0,1)$ (3: green), and $Cauchy(0,1)$ (4: blue). The top row corresponds to the linear case $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \boldsymbol{1}_9$, and the bottom row corresponds to the nonlinear case $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{p} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{p} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$. When using this naive approach with incorrect null distribution, testing power is suboptimal compared to the permutation approach. 105
Figure A.7 Power simulations of RobKAT (black solid line), SKAT (blue dashed line), and LAD (red dotted line) for the setting of 787 simulated SNPs. This mirrors the simulation in Section 3.1 for $p = 787$ in the main text (Figure 3) but with a larger sample size ($n = 300$). The simulation uses IBS kernel, $1,000$ iterations at each setting of c, and 2 causal SNPs. The top row corresponds to the linear case $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_{i1} + \boldsymbol{Z}_{i2}$, and the bottom row corresponds to the nonlinear case $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{2} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{2} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$. 107
Figure A.8 Power simulations with bimodal errors using RobKAT (black solid line), SKAT (blue dashed line), and LAD (red dotted line) for the setting of 787 simulated SNPs. This mirrors the simulation in Section 3.1 for $p = 787$ in the main text (Figure 4) but with a larger sample size ($n = 300$). The simulation uses IBS kernel, $1,000$ iterations at each setting of c, and 2 causal SNPs. Figures (a) and (b) show the linear case $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_{i1} + \boldsymbol{Z}_{i2}$, and figures (c) and (d) display nonlinear case $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{2} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{2} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$. 108

Figure B.1 Power simulation results from the standard setting (no high leverage points). MrKAT is shown in black, and DKAT is shown in red. 4 Responses; IBS kernel, n=100; $h_1 = h_2 = h_3 = h_4 = h(z) = 2(z_1 - z_2)^2 + z_2 z_3 + 3sin(2z_3)z_4 + z_5^2 + 2cos(z_4)z_5$; $\beta = -5\mathbf{1}_4$ 110

Figure B.2 Power simulation results from the robust setting (1 high leverage point). MrKAT is shown in black, and DKAT is shown in red. 4 Responses; IBS kernel, n=100; $h_1 = h_2 = h_3 = h_4 = 2(z_1 - z_2)^2 + z_2 z_3 + 3sin(2z_3)z_4 + z_5^2 + 2cos(z_4)z_5$; $\beta = -5\mathbf{1}_4$; $Y_{100} \sim N(10,1)$ or $Y_{100} = 1$, X_{100} $N(3,1)$ 110

Figure B.3 Estimated average times for one MrKAT test from 1000 simulated iterations. For $i = 1,2,...,n$ and $j = 1,2,...,r$, each simulation setting included data generated from $Y_{ij} = \boldsymbol{X}_i^T \boldsymbol{\beta} + h_j(\boldsymbol{Z}_i) + \epsilon_{ij}$, where $X_i \sim N(0,1)$, $\boldsymbol{\beta} = -5\mathbf{1}_5$, $\boldsymbol{\epsilon}_i \sim N_r(0, I_r)$, $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i \mathbf{1}_p$, and $\boldsymbol{Z}_i$ contained $p = 200$ SNPs with overall minor allele frequencies matching the *SLC17A1* gene. For Plot (a), times were estimated for $n = 300$ and $r = 3, 10, 50, 100, 200, 300$ responses. For Plot (b), times were estimated for $n = 50, 100, 300, 500$. Times were calculated on a 2.9 GHz Intel Core i5 machine with 8 GB 1867 MHz memory. 113

Figure B.4 Power simulation results for 200 continuous, multivariate responses that have errors that are symmetric ($N(0,1)$), skewed (χ_1^2), heavy-tailed (t_3), and substantially heavy-tailed (Cauchy). Of the 200 responses, five were associated ($Y_{ij} = \boldsymbol{X}_i \beta + c * h_j(\boldsymbol{Z}_i) + \epsilon_{ij}$) and 195 were not associated ($Y_{ij} = \epsilon_{ij}$) with the covariates. MrKAT is shown in black, and DKAT is shown in red. Implementation included an IBS genotype kernel, sample size of $n = 100$, 1000 iterations at each value of c, $\beta = -5\mathbf{1}_4$, and $h_1 = h_2 = h_3 = h_4 = 2z_1 + 3z_2 + z_3 + 2z_4 + z_5$. 114

Figure B.5 Power simulation results for 4 continuous, multivariate responses that have errors that are symmetric ($N(0,1)$), skewed (χ_1^2), heavy-tailed (t_3), and substantially heavy-tailed (Cauchy). Of the 4 responses, all were generated with $Y_{ij} = \boldsymbol{X}_i \beta + c * h_j(\boldsymbol{Z}_i) + \epsilon_{ij}$. MrKAT is shown in black, and DKAT is shown in red. Implementation included an IBS genotype kernel, sample size of $n = 25$, 1000 iterations at each value of c, $\beta = -5\mathbf{1}_4$, and $h_1 = h_2 = h_3 = h_4 = 2z_1 + 3z_2 + z_3 + 2z_4 + z_5$. 116

Figure B.6 Power simulation results for four continuous, independent, multivariate responses that are symmetric ($N(0,1)$), skewed (χ_1^2), heavy-tailed (t_3), and substantially heavy-tailed (Cauchy). MrKAT is shown in black, and DKAT is shown in red. Genetic marker sets ($p = 787$ SNPs) were randomly sampled from the MHC region. This simulation utilized an IBS kernel, sample size $n = 100$, 1000 iterations at each value of c, $\beta = -5\mathbf{1}_4$, and the linear form $h_1(\boldsymbol{Z}_i) = ... = h_4(\boldsymbol{Z}_i) = 2z_1 + 3z_2 + z_3 + 2z_4 + z_5$; 120

CHAPTER

1

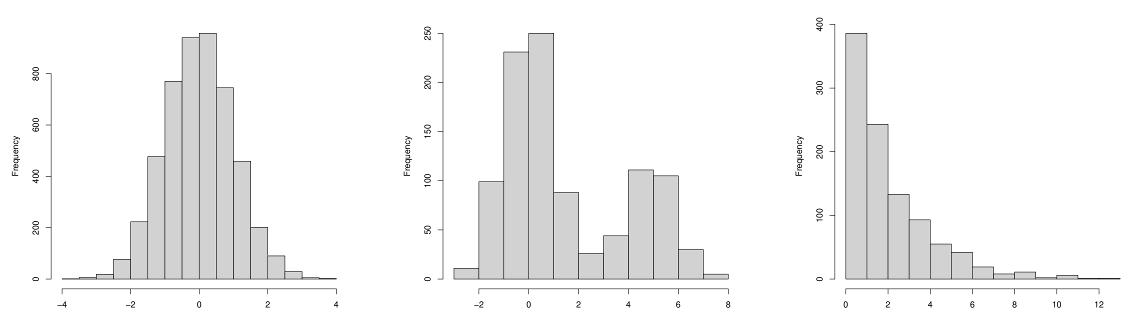
NONTECHNICAL INTRODUCTION

Imagine that you are a scientist trying to understand a complex disease. From your training to become a fancy scientist, you know that there is some sort of genetic heritability to this disease. If parents have the trait or illness, then their children are more likely to have it too. What the scientific community is not sure of, however, is what specific parts of the genome are associated with this disease. Because you are a scientist that also loves statistics, you decide to analyze some data to figure this out.

Think of the genome as a long sequence of data, where only a small number of locations are associated with our disease. There are specific spots that are “interesting” to look at, called single nucleotide polymorphisms (SNPs, pronounced “snips”). Imagine that your statistical test is like a metal detector passing over a sandy beach. For each of these SNP locations in the genome, you use this data analysis tool to tell if that there is an association with this genetic location and the disease or not.

One problem is that there are a large amount of these SNP locations to test with our statistical “detector.” Multi-marker methods are a strategy to combat this problem. Rather than testing each SNP location one at a time, we could test groups of locations instead. This makes sense because if a region is related to the disease, many of the individual SNPs within that area are likely to show significant associations as well. These location groups are sometimes defined by genes, which represent areas of the genome that prior research suggests have a similar function. There are a variety of different multi-marker association methods (metal detectors for groups of SNPs), but the kinds that I study in my research are called kernel association tests. Due to a series of mathematical results and statistical theory, these kernel-based tests are an intuitive and easy-to-use "metal detector." They are also fairly popular in the statistical genetics literature.

Another issue is that many kernel-based "metal detectors" do not work that well when the data looks "weird." Many tests assume that the variability of disease severity is distributed like a nice, bell curve (see Figure 1.1a). A lot is known about this shape, and many mathematical derivations become simpler with this assumption. Unfortunately, your data may look like Figure 1.1b or Figure 1.1c, both of which are clearly not a bell curve shape. If you use a "metal detector" built for bell-curve data but the disease responses are actually like Figure 1.1b or Figure 1.1c, then the detector starts to malfunction. In particular, it will scan over genes that are related to the disease but not detect a signal. In other words, as you crunch data across the genome with your metal detector, you would often be skipping over buried treasure.



(a) Nice, Bell-Curve Distribution **(b)** Bimodal Distribution **(c)** Skewed Distribution

Figure 1.1 Histograms of data generated from various distributions.

In my first project, I develop a new metal detector called RobKAT that solves this problem. It is similar to previous detectors, but this one does not assume a bell-curve distribution. If the data does look like Figure 1.1a, then it will still work; however, if the data happens to look like Figures 1.1b or 1.1c, RobKAT will not malfunction.

Up until now, we only considered one disease aspect at a time, but what if we wanted to detect many diseases? For example, suppose that we were interested in a disease like schizophrenia and wanted to detect for multiple severity measures like cognitive reasoning, memory, and presence of hallucinations all at the same time. Because these measures are related, there's literature suggesting that our analyses would work better by analyzing them together than separately. There are also kernel-based detectors for this type of analysis, but many of them suffer from similar problems as before; they assume that each response measurement is bell-curve shaped, so the tests malfunction when the data do not look this way. My second project, MrKAT (pronounced "Mister Cat"), solves this problem by extending the RobKAT methodology to also accept multiple disease measures.

For my third project, I turn my attention to a different type of genetic analysis that also has problems with "detector malfunctioning." Our previous goal was to find the specific parts of the genome that are associated with a disease. For example, studies have shown that the *15q24-25* region on Chromosome 15 is associated with lung cancer risk. Analyses have also shown that smoking tobacco may exacerbate this risk from Chromosome 15. These types of effects are called gene-environment interactions because an exposure (smoking) is affecting the association between a gene (*15q24-25*) and a trait (lung cancer risk). As you scan your detector across the genome, some regions may have such an interaction effect, and our detector would work better if it accounted for this possibility. There exist some kernel-based methods that do exactly this; however, many of them rely on assuming the bell-shaped distribution from Figure 1.1a. When this assumption is incorrect, these detectors malfunction. In particular, they tend to say "yes! this spot is related to the disease!" at locations that truly have no association going on.

My third project, "RoCKAT" (pronounced like "rocket"), develops a method to solve this problem. To do this, we first borrow ideas from RobKAT and MrKAT so that our new test works well when the distributions are messy. We also develop a new algorithm, because a theoretical assumption in our previous algorithms (the exchangability condition) does not always hold in this case of gene-environment interaction models. Through this project, we show that RoCKAT performs well both when the distribution is bell-shaped and when it's skewed.

While my dissertation projects focus on problems related to genetics, all of these methods can be applied to a myriad of analysis problems. Genetic association testing is one of the most popular application areas, but kernel methods have been applied to microbiome data, text analytics, image processing, and other areas that involve a large number of variables. The overall theme of my research has been simply to apply robust statistical ideas to kernel association tests, which is not exclusive to statistical genetics. This type of thinking can be used to improve many analytical tools is just the beginning of an exciting research area.

CHAPTER

2

TECHNICAL INTRODUCTION AND LITERATURE REVIEW

2.1 Introduction

A general goal of a linear regression model such as $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \gamma + \epsilon_i$ is to explore association between a response Y_i and variables $\boldsymbol{Z}_i \in \mathbb{R}^p$, while controlling for certain covariates $\boldsymbol{X}_i \in \mathbb{R}^q$. Many estimation and hypothesis testing problems in this setting are well-understood, but issues arise when the dimension of $\boldsymbol{Z}_i$ increases. For example, the average variance of prediction may increase. As a more general problem, accurately modeling a large number of variables may be difficult, especially when many interaction terms or a complex nonparametric structure is required.

Genetic data analysis often presents similar issues of modeling high-dimensional covariates. For example, researchers may regress an individual's phenotype Y_i on genetic markers $\boldsymbol{Z}_i$ in a gene or gene pathway, while controlling for clinical covariates $\boldsymbol{X}_i$ such as age and sex. Often, these genetic markers are sets of single nucleotide polymorphisms (SNPs) counting the number of minor alleles at each locus. In a marker set, there may be a large number of SNPs, and the relationship between these genetic markers is unknown and potentially complex. As a solution to this modeling problem, the single marker testing strategy involves performing p regressions, each with one Z_{ij}, then applying a multiple testing correction (Haines et al., 2005; Tusher et al., 2001). For large p, however, the multiple testing burden is high, and there have been issues with replication of results (Ioannidis et al., 2001; NCI-NHGRI Working Group on Replication in Association Studies, 2007). Testing only one loci at a time, single marker methods also lose the ability to test at the gene or gene pathway level. An alternative strategy is the burden-based test, which summarizes the $\boldsymbol{Z}_i$ vector into a summary value, usually via a strategic weighting scheme (Gauderman et al., 2007; Li and Leal, 2008; Madsen and Browning, 2009; Wang and Elston, 2007). However, many of these methods assume that the loci are additive with similar direction of effect, leading to power loss when the assumption does not hold (Basu and Pan, 2011).

Nonparametric methods are also capable of modeling a complex relationship among the $\boldsymbol{Z}_i$. However, many of these methods require specification of smoothness properties, which may be cumbersome when p is large. Alternatively, the kernel machine regression (KMR) framework models complex relationships between a large number of variables in an elegant and often intuitive manner through the use of a kernel function. Due to its connection with a linear mixed model, KMR methods easily lend themselves to hypothesis testing of the overall effect of $\boldsymbol{Z}_i$ on the response.

This paper aims to provide an overview of KMR hypothesis testing methods as well as common application areas. Section 2.2 discusses the kernel machine regression model and its theoretical foundations. The literature review then describes methods for univariate responses in Section 2.3, followed by a discussion of multivariate response methods in Section 2.4. Extensions of these tests to gene-environment interaction models is then discussed in Section 2.5. Popular applications of KMR to SNP genetic association analysis are discussed throughout Sections 2.2-2.5. For increased breadth, Section 2.6 explores additional applications to microbiome and copy number variance. This paper concludes by discussing open problems and directions for the future in Section 2.7.

2.2 Kernel Machine Regression

Many kernel machine regression methods utilize the partially linear model, $Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + h(\boldsymbol{Z}_i) + \epsilon_i$, where $h(\cdot)$ may be parametric or nonparametric. Although some authors choose to add an extra intercept term (Kong et al., 2016; Wu et al., 2011), we have included an intercept term among the $\boldsymbol{X}_i$ as in the works of Liu et al (Liu et al., 2008,0). From a genetics viewpoint, ϵ_i refers to an error term and $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ corresponds to the response, clinical covariates, and genetic markers for individual i. In this model, $h(\cdot)$ represents the effect of genetic markers in the gene pathway on the response after controlling for covariates. While one could explicitly model a parametric or nonparametric functional form of $h(\cdot)$ and use standard methods, explicitly specifying well-fitting $h(\cdot)$ may be difficult when p is moderate to large. Rather than constructing $h(\cdot)$ directly, we instead work with a function space $\mathscr{H}_K$ and assume that $h(\cdot)$ is an element. However, $\mathscr{H}_K$ is often high-dimensional or even infinite-dimensional, which may still be complicated or computationally burdensome.

KMR methodology instead utilizes a kernel function to provide a more computationally efficient solution. Kernels are symmetric, finitely positive semi-definite functions, meaning that the matrix K, with $(i,j)^{th}$ entry equal to $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$, is always positive semi-definite. Intuitively, the kernel is a real-valued function with our $\boldsymbol{Z}_i$'s as input that also works as an inner product of the images (ie: $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j) =< h(\boldsymbol{Z}_i), h(\boldsymbol{Z}_j) >$). The main benefit of kernels is that K contains all relevant information regarding inner products and distances among our data, the $\boldsymbol{Z}_i$'s. Algorithms and statistics that depend on $h(\boldsymbol{Z}_i)$ through this inner product information only require the lower-dimensional K rather than computations involving elements of the higher-dimensional function space. Thus, kernel methods are dimension reduction strategies. This process of leveraging kernels in high-dimensional problems is also prevalent in other machine learning methods such as support vector machines, PCA, and ridge regression (Hastie et al., 2009). For a detailed discussion of kernel functions in machine learning, many textbooks are available (Hastie et al., 2009; Shawe-Taylor and Cristianini, 2004; Theodoridis and Koutroumbas, 2009; Williams, 2003). Details and derivations of kernels' theoretical properties and results can also be found in various texts (Shawe-Taylor and Cristianini, 2004; Theodoridis and Koutroumbas, 2009; Wahba, 1990).

It can also be shown that a kernel function defines a reproducing kernel for a unique reproducing kernel Hilbert space (RKHS), in our case $\mathscr{H}_K$ (Shawe-Taylor and Cristianini, 2004; Wahba, 1990).

A RKHS is a Hilbert space defined by the kernel function $k(t,\cdot)$ with the reproducing property in which for $f \in \mathscr{H}_K$ we have $< f, k(t,\cdot) >= f(t)$. Thus, a RKHS would allow representation of $h(\boldsymbol{Z}_i)$ using only elements of a lower-dimensional kernel matrix, simplifying estimation in nonparametric modeling. Estimates of $h(\cdot)$ are often derived from minimizing the penalized loss criterion

$$\min_{h\in\mathscr{H}_K}\left[\sum_{i=1}^{N}\rho(Y_i-\boldsymbol{X}_i^T\boldsymbol{\beta}-h(\boldsymbol{Z}_i))+\lambda\,||h||^2_{\mathscr{H}_K}\right] \tag{2.1}$$

which may be an infinite-dimensional problem due to the nonparametric form of $h(\cdot)$ and dimensionality of $\mathscr{H}_K$. Since $\mathscr{H}_K$ is a RKHS, the Representer Theorem shows that the solution to this minimization criterion follows a linear combination $h(t) =< h, k(t,\cdot) >= \sum_{i=1}^{N}\alpha_i k(t,\boldsymbol{Z}_i)$, which is a finite-dimensional representation (Hastie et al., 2009; Kimeldorf and Wahba, 1971; Liu et al., 2007). In estimation procedures, we may substitute $h(t)=\sum_{i=1}^{N}\alpha_i k(t,\boldsymbol{Z}_i)$ and more easily derive estimates and statistics. Furthermore, these RKHS results link the functional form of $h(\cdot)$ to the user-specified kernel.

The connection between $h(\cdot)$ and $k(\cdot,\cdot)$ also leads to many kernel options. The linear kernel $k(\boldsymbol{Z}_i,\boldsymbol{Z}_j)=a\boldsymbol{Z}_i^T\boldsymbol{Z}_j+c$, for example, corresponds to a function space $\mathscr{H}_K$ of all linear functions. More generally, the kernel $k(\boldsymbol{Z}_i,\boldsymbol{Z}_j)=(a\boldsymbol{Z}_i^T\boldsymbol{Z}_j+c)^d$, corresponds to an $\mathscr{H}_K$ containing d-degree polynomial functions. Noting the kernel's connection to inner products, we see that the polynomial kernel functions modify a Euclidean inner product $\boldsymbol{Z}_i^T\boldsymbol{Z}_j$. The Gaussian kernel, $k(\boldsymbol{Z}_i,\boldsymbol{Z}_j)=exp\{-||\boldsymbol{Z}_i-\boldsymbol{Z}_j||^2/\xi\}$, is another popular choice and corresponds to an $\mathscr{H}_K$ containing radial basis functions, which have many theoretical applications (Buhmann, 2003). The parameter, ξ, can be used to affect the smoothness of h being modeled (Wu et al., 2010). Some methods also alter these popular kernels and incorporate weights based on application or data-specific quantities, such as minor allele frequencies or phylogenetic relationships (Kwee et al., 2008; Wu et al., 2011; Zhan et al., 2017a). Additionally, different kernels may be added, scaled, multiplied, and combined to create additional valid kernel functions tailored to the analysis of interest (Shawe-Taylor and Cristianini, 2004).

Kernel functions also can be interpreted as summarizing the similarity between two vectors. Results from Mercer's Theorem may be used to show that any kernel may be thought of as a covariance kernel for a Gaussian distribution (Shawe-Taylor and Cristianini, 2004). In the continuous case, the positive semi-definiteness of $k(\cdot,\cdot)$ implies that there exists a family of Gaussian random variables for which the kernel represents the covariance (Wahba, 1990). Therefore, the kernel function may be specified based on similarity measures of $\boldsymbol{Z}_i$ and $\boldsymbol{Z}_j$. A common kernel in statistical genetics literature, the Identity by State (IBS) kernel, is the proportion of alleles shared between two individuals' marker sets, $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = \frac{1}{2p}\sum_{k=1}^{p}[2\mathbb{I}\{Z_{ik} = Z_{jk}\} - \mathbb{I}\{|Z_{ik} - Z_{jk}| = 1\}]$.

As demonstrated with these examples, kernel specification provides intuitive modeling strategies for the analyst. Additionally, the KMR framework provides flexibility since the analyst may use any symmetric and well-behaved function as a kernel.

2.3 Hypothesis Testing with a Univariate Response

One common task in the kernel machine regression framework is to test for an overall effect of a high-dimensional variable on a univariate response, Y. In terms of the model, $Y_i = \boldsymbol{X}_i\boldsymbol{\beta} + h(\boldsymbol{Z}_i) + \epsilon_i$, this goal translates to testing the null hypothesis $H_0 : h(\cdot) = 0$. When $\boldsymbol{Z}_i$ represents a vector of an individual's genetic markers, $H_0 : h(\cdot) = 0$ signifies no overall effect of an individual's genetic marker set on his response. This type of hypothesis is useful when determining if a particular gene or pathway may be involved in a certain phenotypic response. This section will discuss hypothesis testing methodology designed for normal, non-normal, and discrete univariate responses.

2.3.1 Least Squares Kernel Machine

The most popular method of KMR hypothesis testing is the least squares kernel machine (LSKM), which has been described from multiple perspectives in the literature (Kwee et al., 2008; Liu et al., 2007; Wu et al., 2011). One hallmark of these LSKM methods is a tailoring to normally distributed, continuous responses.

Following the RKHS process described in Section 2.1 and by Liu et al. (2007) , LSKM methodology utilizes a user-specified kernel to define a function space $\mathscr{H}_K$ containing $h(\cdot)$. Then LSKM procedures change the minimization criterion in Equation 2.1 to a penalized least squares problem by inputting $\rho(t) = t^2$. Applying the Representer Theorem and substituting $h(t) = \sum_{j=1}^{N}\alpha_i k(t, \boldsymbol{Z}_j)$ into Equation 2.1, we may calculate estimates for α, $\boldsymbol{\beta}$, and $h(\cdot)$. In particular, $\widehat{\alpha} = \lambda^{-1}(I + \lambda^{-1}K)^{-1}(Y - X\widehat{\boldsymbol{\beta}})$, $\widehat{\boldsymbol{\beta}} = \{X^T(I + \lambda^{-1}K)^{-1}X\}^{-1}X^T(I + \lambda^{-1}K)^{-1}Y$, and $\widehat{h} = K\widehat{\alpha}$ (Liu et al., 2007).

Rather than implement this minimization process, LSKM methodology often leverages a connection with linear mixed models. For detailed derivations of the following discussion, see Liu et al. (2007). Creating this mixed model connection, suppose that $h \sim N(0, \tau K)$, $\epsilon \sim N(0, \sigma^2 I)$, and $Cov(h, \epsilon) = 0$. The notation (τ, σ^2) represents a set of parameters, and K signifies the Gram matrix whose $(i, j)^{th}$ entry is $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$. Within this mixed model framework, we are led to the joint normality of h and Y:

$$\begin{bmatrix} Y \\ h \end{bmatrix} \sim N\left(\begin{bmatrix} X\beta \\ 0 \end{bmatrix}, \begin{bmatrix} \tau K + \sigma^2 I & \tau K \\ \tau K & \tau K \end{bmatrix} \right) \tag{2.2}$$

We can then solve the subsequent mixed model equations to obtain the same estimates as from minimizing Equation 2.1. However, these estimates also correspond to familiar statistics. For example, $\widehat{\boldsymbol{\beta}}$ is a generalized least squares estimate, and $\widehat{h}$ is the best linear unbiased predictor (BLUP), where "best" is defined by minimizing $Var(\widehat{h} - h)$ while remaining linear and unbiased. For τ, σ^2, and any parameters defined in K, the LSKM procedure derives estimates through restricted maximum likelihood (REML). These mixed model estimates, $\widehat{\boldsymbol{\beta}}_{GLS}$ and $\widehat{h}_{BLUP}$, can be computed using mixed model software such as PROC MIXED or PROC GLMMIX in SAS and the `nlme` package in R. Since these software packages are fairly prevalent, estimation using the mixed model viewpoint may be easier to implement.

We also see simplifications in hypothesis testing from a mixed model framework. Since $H_0 : h(\cdot) = 0 \Leftrightarrow H_0 : \tau = 0$, mixed model methodology suggests testing $H_0 : h(\cdot) = 0$ with a variance component score statistic for τ. The resulting REML score test statistic for τ is

$$T_{LSKM} = \frac{1}{2\widehat{\sigma}^2}(Y - X\widehat{\boldsymbol{\beta}})^T K (Y - X\widehat{\boldsymbol{\beta}}) - tr\{(I - X(X^T X)^{-1} X)K\} \tag{2.3}$$

Due to normality, the null distribution of T_{LSKM} is a mixture of chi-square random variables and can be estimated by a Satterthwaite approximation (Liu et al., 2007).

A major benefit from this hypothesis testing method is that the nonparametric component vanishes under the null hypothesis $H_0 : h(\cdot) = 0$, leaving the fixed parametric linear model $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \epsilon_i$. Thus, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ in Equation 2.3 are the least square estimates from a fixed effects linear model, leaving T_{LSKM} and the corresponding p-value able to be implemented without nonparametric methods. With this simplification, LSKM is computationally efficient and well-suited to genome-wide studies requiring thousands of inference procedures. However, the assumption of a normally distributed response renders LSKM unsuited to clinical data sets with skewed, heavy tailed, bimodal, or outlier-containing responses (Kong et al., 2016; Martinez et al., 2020a). Additionally, proper data imputation must be performed on missing data prior to analysis with these methods.

2.3.1.1 SKAT

The most popular KMR method for hypothesis testing, the sequence kernel association test (SKAT), is a special case of the least squares kernel machine as it also uses a square loss function. Developed by Wu et al. (2011), SKAT is designed for detecting associations between a response and genetic marker set containing rare variants, which are SNPs with minor allele frequency (MAF) between 1% and 5%. A unique feature of SKAT is the use of a kernel matrix that includes a weighting scheme dependent on MAFs. This system of weights increases the impact of rare variant effects in hypothesis tests. Being a LSKM, SKAT also has the benefits of LSKM's simplicity and computational efficiency. However, SKAT also relies upon a normally distributed response and is similarly subject to power loss when normality does not hold (Martinez et al., 2020a).

An improvement upon SKAT is the optimal unified test, SKAT-O, developed by Lee et al. (2012). Burden-based methods assume that all loci have similar size and direction of effect, which leads to a powerful test when many of the loci in a genetic marker set are causal. SKAT, a non-burden-based method, does not carry this assumption, which makes it powerful when analyzing rare traits or using genetic marker sets with few causal variants. SKAT-O attempts to combine these two methods to create a powerful test for both rare and common causal variants. In particular, the SKAT-O test statistic for $H_0 : h(\cdot) = 0$ is a weighted sum of T_{SKAT} and T_{Burden} based on the correlation between gene effects. SKAT-O behaves like a burden test when marker correlation is strong, and the method behaves like SKAT when marker correlation is low.

2.3.2 Quantile Regression Kernel Machine

Like the least square methods in Section 2.3.1, the quantile regression kernel machine (QRKM) is based upon the RKHS structure, user-specified kernel function, and penalized minimization criterion in Equation 2.1. As an overview the derivations by Kong et al. (2016), the QRKM utilizes a check function for the loss in Equation 2.1. More specifically, the QRKM sets $\rho(t) = \eta t \ \mathbb{I}_{\{t \geq 0\}} + (\eta -$

$1)t\ \mathbb{I}_{\{t<0\}}$ where $\eta \in (0,1)$ is the desired conditional quantile for inference. For example, when testing association with the conditional median response, $\eta = 0.5$ and $\rho(t) = 0.5|t|$.

The QRKM idea is to then follow the score test procedure as in the LSKM methodology. Replacing the square loss with the non-differentiable check function, the QRKM derivation must instead use the check function's subgradient, $\psi(t) = \eta\mathbb{I}_{\{t>0\}} + (\eta-1)\mathbb{I}_{\{t<0\}} + (\eta-R)\mathbb{I}_{\{t=0\}}$, where R is the outcome of a $Bern(\eta)$ random variable. The resulting score-type statistic is

$$T_{QRKM} = \boldsymbol{w}^T K \boldsymbol{w} \tag{2.4}$$

Here, the i^{th} element of $\boldsymbol{w} \in \mathbb{R}^n$ is $\psi(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}})$, K is the user-specified kernel matrix, and $\widehat{\boldsymbol{\beta}}$ is the estimate under $H_0 : h(\cdot) = 0$. Since the model under H_0 is $Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + \epsilon_i$, we have that $\widehat{\boldsymbol{\beta}}$ is the quantile regression slope estimate that can be computed with any available quantile regression software, such as the `quantreg` R package. The null distribution of T_{QRKM} and resulting p-value can be derived using a permutation algorithm approach (for full algorithm details see Kong et al. (2016)).

Although permutation strategies are not the most computationally efficient, Zhan and Wu (2017) has recently developed a faster algorithm for computing the p-value based on approximating the null distribution of T_{QRKM}. Leveraging previously derived moments for statistics of the form $T = tr(X^T XY^T Y)$ by Kazi-Aoual et al. (1995) and applying Josse et al. (2008)'s fast permutation testing techniques designed for RV-type statistics of the form $tr(HAA^T HBB^T)$ with centering matrix H, we may center the kernels and match moments to the Pearson III distribution (Zhan and Wu, 2017). Inference using this moment matching technique can be 6,000 times faster than with standard permutation testing, while maintaining similar power and type I error (Zhan and Wu, 2017). Thus, even though QRKM implementation cannot use chi-square mixture inference as in LSKMs, there exists computationally efficient algorithms making QRKMs also viable for genome-wide analyses.

Unlike LSKM and SKAT, the QRKM model is not tailored to a normally distributed response. In the same way that a median is less sensitive to outliers than a mean, QRKM inference is less sensitive to responses with heavy tails, skew, and outliers. In these cases of non-normality, QRKM methods produce a more powerful test (Kong et al., 2016; Martinez et al., 2020a). Thus, the QRKM is a more robust method than the LSKM procedures described in the previous section, and the intended underlying distribution is a primary difference between these two methods.

2.3.3 Logistic and GLM Kernel Machines

Testing associations with categorical or binary responses is a common goal, especially in case-control studies. However, the methods described in Section 2.1 through Section 2.3.2 only apply to situations with a single continuous response. Extending the KMR hypothesis tests to a discrete response, Liu et al. (2008) developed the logistic kernel machine and GLM kernel machine.

The logistic kernel machine test proceeds similarly to LSKMs and QRKMs, but instead accommodates binary responses by using the modified model $logit(\mu_i) = \boldsymbol{X}_i^T \boldsymbol{\beta} + h(\boldsymbol{Z}_i)$ where $\mu_i = P(Y_i = 1|\boldsymbol{X}_i, \boldsymbol{Z}_i)$ and the covariate $\boldsymbol{X}_i$ includes an intercept. Again leveraging RKHS theory, logistic kernel machine testing first inputs a user-specified kernel function, sets up the penalized criterion in Equation 2.1, applies the Representer Theorem to substitute $h(t) = \sum_{i=1}^N \alpha_i k(t, \boldsymbol{Z}_i)$, and solves for parameters. To modify for binary responses, Liu et al. (2008) utilizes a loss function based on a log-likelihood in logistic regression.

The logistic kernel machine methodology exists with a similar connection to the logistic mixed model, $logit(\mu_i) = \boldsymbol{X}_i^T \boldsymbol{\beta} + h_i$ with $\boldsymbol{h} \sim N(0, \tau K)$ (Liu et al., 2008). Like LSKM, estimates from the mixed model match exactly the quantities derived from the penalized minimization criterion. The hypothesis test for the overall genetic effect, $H_0 : h(\cdot) = 0 \Leftrightarrow H_0 : \tau = 0$, is similarly accomplished through a variance component score test for τ, resulting in a familiar quadratic form test statistic

$$T_{GLM} = \frac{1}{2}(Y - \widehat{p})^T K (Y - \widehat{p}) \tag{2.5}$$

where $logit(\widehat{p}_i) = \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}$ is from the null parametric logistic model, and the usual logistic regression slope estimate, $\widehat{\boldsymbol{\beta}}$, is found under the null hypothesis (Wu et al., 2010; Zhang and Lin, 2003). Parameter estimation of $\widehat{\boldsymbol{\beta}}$ can be quickly computed using standard logistic regression methods such as PROC LOGISTIC in SAS or the R function `glm()`. The null distribution of T_{GLM} is a Chi-squared mixture distribution approximated using the Satterthwaite method and moment matching (Tzeng and Zhang, 2007; Wu et al., 2010; Zhang and Lin, 2003). Another approach views the score as a Gaussian process and derives a sharp upper bound on the p-value (Liu et al., 2008). Other strategies in the literature also have been proposed using Davies exact method, which is based on numerically inverting a characteristic function (Broadaway et al., 2016; Davies, 1980; Wu et al., 2011). Similar to QRKM, it would also be possible to employ a permutation-type strategy or fast approximation algorithm to obtain a p-value.

In the case of other discrete responses, Liu et al. (2008) also details a generalized kernel machine model, $g(\mu_i) = \boldsymbol{X}_i^T \boldsymbol{\beta} + h(\boldsymbol{Z}_i)$, that accommodates any response Y_i in the exponential family with monotone link function $g(\cdot)$. Estimation occurs similarly through either a penalized likelihood framework or the generalized linear mixed model, and hypothesis testing proceeds through a variance component score test. Tzeng and Zhang (2007) also developed a type of GLM kernel machine, however this method focuses on haplotype similarity and some have criticized the potential computational burden of obtaining haplotypes from large data sets (Wu et al., 2010).

2.4 Hypothesis Testing with a Multivariate Response

Although the methods discussed in Section 2.3 perform hypothesis tests between a high-dimensional variable and a univariate response, data sets with multivariate responses are common in scientific studies. There exist many clinical trials, for example, that record individuals' genetic information and multiple phenotypes to investigate potential associations (International Schizophrenia Consortium et al., 2009; Yolken et al., 2011). While it is possible to perform separate univariate tests on each response, this strategy fails to leverage correlations among the responses. Modeling response correlation structure is also relevant when genetic markers may be involved in a pleiotropic relationship, where a set of markers affect several phenotypes. From the perspective of a multiple testing burden, performing a single multivariate test may be preferred again over many univariate tests.

To address these issues, many multivariate KMR models and testing procedures exist in the literature. Like their univariate counterparts, multivariate KMR models vary in their distributional assumptions and intended applications. This section will discuss normality-based multivariate kernel machine methods (Section 2.4.1), robust methods (Section 2.4.2), and popular areas of application (Sections 2.4.3 - 2.4.4).

2.4.1 Normality-Based Methods

The multivariate kernel machine regression (MKMR) method is a multimarker association test for multiple, possibly correlated, responses. By leveraging the correlation structure among phenotypes, multivariate KMR methods have improved power over repeated univariate KMR tests, with even greater improvement for moderate to highly correlated phenotypes. While this section provides an overview, Maity et al. (2012) provides the development and derivation details of MKMR.

The MKMR model for the r^{th} recorded response from individual i is

$$Y_{ri} = \boldsymbol{X}_i^T \boldsymbol{\beta}_r + h_r(\boldsymbol{Z}_i) + \epsilon_{ri} \tag{2.6}$$

where $r \in \{1, 2, ..., R\}$, $(\epsilon_{1i}, ..., \epsilon_{Ri})^T \sim N(0, \Sigma)$, and $\Sigma = \{\sigma_{r_1 r_2}\}$ with $\sigma_{r_1 r_2} = corr(Y_{r_1 1}, Y_{r_2 1})$. The model can be simplified in matrix notation by $\mathbf{Y} = \mathbf{X}^T \boldsymbol{\beta} + \mathbf{h} + \epsilon$. This notation suggests a connection between MKMR and the LSKM model, since both KMR methods assume normally distributed errors. Like LSKM, the MKMR method begins familiarly with specifying a kernel function, with the new modification that one kernel for each response is required. Then a RKHS structure represents $h_r(\cdot)$ as a linear combination of kernel functions and employs a square loss. Like other kernel machine methods, the form of $h_k(\cdot)$ (parametric or nonparametric) is determined by the choice of kernel.

Similar to the connection between LSKM and a linear mixed model, the MKMR method relates to a multivariate linear mixed model. The estimating equations derived from the penalized log-likelihood and RKHS algebra match exactly the normal equations from a multivariate linear mixed model with the assumptions $\mathbf{h} \sim N(0, \mathbf{K}\Lambda)$, $\epsilon \sim N(0, \Sigma)$, $\mathbf{K} = diag(K_1, ..., K_R)$, and $\Lambda = diag(\tau_1 I_{J_1}, ...\tau_R I_{J_R})$. MKMR then tests the hypothesis, $H_0 : h_1(\cdot) = ... = h_R(\cdot) = 0$, through a score-type test. The resulting REML score statistic is

$$T_{MKMR} = (\mathbf{Y} - \mathbf{X}^T \widehat{\boldsymbol{\beta}})^T \mathbf{V}_0^{-1} \mathbf{K} \mathbf{V}_0^{-1} (\mathbf{Y} - \mathbf{X}^T \widehat{\boldsymbol{\beta}}) \tag{2.7}$$

with $\mathbf{V}_0 = (\Sigma + \mathbf{K}\Lambda)|_{H_0}$, and the null distribution follows a chi-square mixture that is approximated by an empirical algorithm (Maity et al., 2012). In the case of larger sample sizes, MKMR and its p-value computation algorithm may be computationally intensive and perform far slower than other multivariate KMR methods (Wu and Pankow, 2016).

The MKMR method also is restricted to testing a non-specific association among any of the responses. Recent literature provides modifications of the MKMR statistic in order to test a more specific association within a response subset, allowing investigation for which phenotypes drive the global association (Maity et al., 2018). Other variations on MKMR for rare-variant genetic markers exist and further detail will be discussed in Section 2.4.3 (Wu et al., 2011).

2.4.2 Robust Methods

While other multivariate methods may depend upon a normal distribution, the dual kernel association test (DKAT) presents a more robust approach to multivariate KMR hypothesis testing. DKAT analyses are also computationally efficient even for high dimensional responses, $y \in \mathbb{R}^{200}$, and large sample sizes, $n = 1000$. DKAT performance and methodology are fully detailed in a 2017 paper by Zhan et al. (2017b).

As an overview, DKAT is a generalization of MKMR whose derivations accommodate non-normal responses through the model $g(E[Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i]) = \boldsymbol{X}_i^T \boldsymbol{\beta} + h(\boldsymbol{Z}_i)$ where $g(\cdot)$ is the link function commonly found in the GLM framework. DKAT then tests the hypothesis, $H_0 : h(\cdot) = 0$, with a score type statistic as in previously discussed methods. In this way, the DKAT may be viewed as a generalization of the MKMR test. For continuous uncorrelated responses, the MKMR variance component score test utilizes $T_{MKMR} = \frac{1}{2\widehat{\sigma}^2}(\mathbf{Y} - \widehat{\mathbf{Y}})^T K_G (\mathbf{Y} - \widehat{\mathbf{Y}})$ where K_G is the overall genetic marker kernel and $\widehat{\sigma}$ is the scale estimate from the null model. DKAT's generalization comes from the observation $T_{MKMR} = \frac{1}{2\widehat{\sigma}^2} tr[K_G(\mathbf{Y} - \widehat{\mathbf{Y}})(\mathbf{Y} - \widehat{\mathbf{Y}})^T] = \frac{1}{2\widehat{\sigma}^2} tr[K_G K_Y]$ where K_Y is a kernel matrix describing the response similarity structure. By incorporating two kernel matrices, the DKAT statistic can compare the response similarity to the genetic marker similarity structure. Like the genotype kernel, K_Y can also be selected by the user. Through customizing this phenotype kernel, DKAT may be tailored to binary, continuous, or mixed responses.

The centered version of DKAT's score-type statistic with centering matrix H is

$$T_{DKAT} = \frac{tr(HK_G HK_Y)}{\sqrt{tr(HK_G HK_G) tr(HK_Y HK_Y)}} \tag{2.8}$$

By centering the statistic and noting that both K_G and K_Y are symmetric positive semi-definite, we see that T_{DKAT} is a RV-type statistic of the form $tr(HAA^T HBB^T)$. This allows DKAT procedures to be implemented with Zhan and Wu (2017)'s fast permutation algorithm described in Section 2.3.2. Although DKAT accommodates multiple response types, adequate hypothesis testing power has only been demonstrated for continuous responses. The literature has also only explored the effectiveness of the phenotype kernel $K_Y = \mathbf{Y} - \widehat{\mathbf{Y}}$, where $\widehat{\mathbf{Y}}$ are values fitted from a least squares regression model.

2.4.3 Methods for Rare Variant Covariates

A popular variation on MKMR is the multivariate sequence kernel association test (MSKAT). Derived by Wu and Pankow (2016), MSKAT also depends on normally distributed responses, a multivariate linear mixed model, and a variance component score test. Like its univariate counterpart, SKAT (Wu et al., 2011), MSKAT incorporates weights into the kernel matrix based on minor allele frequency. This weighting scheme tailors the method to association testing with rare genetic variants.

Another special case designed for rare variants, Gene Association with Multiple Traits (GAMuT) is a version of DKAT and MKMR that allows either continuous or discrete responses (Broadaway et al., 2016). Using the DKAT notation, the test statistic is $T_{GAMuT} = \frac{1}{N} tr(HK_Y HK_G)$ with a suggested genotype matrix $K_G = ZWZ^T$ that upweight rare variants. The methodology of GAMuT is in kernel distance covariance (KDC). The KDC method tests for independence of two groups of multivariate variables, and literature has shown a mathematical equivalence between KDC and the kernel machine regression framework (Hua and Ghosh, 2015). Upon comparison, T_{GAMuT} is the DKAT statistic with different scaling and becomes a version of MKMR when $K_Y = (\mathbf{Y} - \widehat{\mathbf{Y}})^T (\mathbf{Y} - \widehat{\mathbf{Y}})$. However, GAMuT computes p-values based on asymptotic results, while DKAT employs a fast permutation testing algorithm. This implementation difference renders DKAT more powerful than GAMuT, especially in samples sizes less than 500 individuals (Zhan et al., 2017b).

2.4.4 Methods for Pleiotropic Relationships

Sun et al. (2016) developed a variation on MKMR tailored to pleiotropic effects, the multivariate rare-variant association test (MURAT). Pleiotropy describes when one genetic variant affects outwardly unrelated phenotypes, suggesting correlated effects of variants on an individual's responses. MURAT then incorporates a specified correlation coefficient and MAF weights into the kernel matrix, K. Like MKMR, MURAT assumes multivariate normality of response, performs a variance component score test, and obtains p-values through a chi-square mixture distribution. The original MKMR discussed in Section 2.4.1 is the special case when the added correlation coefficient is set to zero (Maity et al., 2012).

2.5 Hypothesis Testing with Gene-Environment Interaction

The methods described in Sections 2.3-2.4 have been tests to detect associations between genetic marker sets and phenotypes, while controlling for clinical covariates; however, some researchers may need to consider environmental exposures that may affect these genetic associations. For example, SNPs in the *15q24-25* region are thought to be associated with lung cancer risk, but analyses have suggested that this risk is affected by smoking status (Lin et al., 2013). As another example, other studies have suggested that the *MTR* gene's effect on homocysteine methylation differs between age groups (Marceau et al., 2015). These types of effects are called gene-environment interactions.

Although models vary somewhat from method to method, a common conceptualization of gene-environment interactions involves

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}_X + h_E(\boldsymbol{E}_i) + h_G(\boldsymbol{G}_i) + h_{GE}(\boldsymbol{G}_i, \boldsymbol{E}_i) + \epsilon_i \tag{2.9}$$

where $(Y_i, \boldsymbol{X}_i, \epsilon_i)$ represent the response, clinical covariates, and random error for individual i and $h_E(\cdot)$, $h_G(\cdot)$, and $h_{GE}(\cdot)$ are nonparametric functions for the environmental, genetic, and gene-environment interaction effects, respectively. When there are only a small number of environmental variables, a common model for $h_E(\boldsymbol{E}_i)$ is a linear form, $\boldsymbol{E}_i^T \boldsymbol{\beta}_E$.

In kernel-based approaches to this problem, $h_E(\cdot)$ and $h_G(\cdot)$ may be modeled with any common kernel described in Section 2.2; therefore, gene-environment KMR tests are multi-kernel methods. The interaction similarity structure is often modeled with a combination of these environmental and genotype kernels. A common choice is produced by element-wise multiplication of genotype (K_G) and environmental kernels (K_E) in the definition $K_{GE} = K_E \odot K_G$ (Marceau et al., 2015; Wu et al., 2013; Zhao et al., 2015a,1). If there is only one environmental exposure and K_E is linear, this kernel can also be represented by $K_{GE} = diag(E_i) K_G diag(E_i)$. When there are many environmental effects, other kernels have been developed that take care not to repeat terms (Wang et al., 2015b).

There are two main approaches for gene-environment interaction analyses, and each addresses a different type of research question. For one type, the goal is to investigate known genes and environmental factors to test whether they affect each other. These research questions would address the presence and characterization of gene-environment interactions in order to identify at-risk populations, as modifying environmental exposures is often more feasible than altering genes (Hutter et al., 2013). In the framework of Equation 2.9, these tests would focus on $H_0 : h_{GE}(\cdot,\cdot) = 0$. Unlike variance component score tests for marginal genetic effects, the nonparametric function $h_G(G_i)$ is part of the null model and must be estimated, which poses a major challenge.

The gene-environment set association test (GESAT) is a methodological approach to this problem (Lin et al., 2013). The genetic marker effect is assumed to be fixed and the genotype kernel set to be linear, resulting in the modified model: $g(\mu) = X\boldsymbol{\alpha}_1 + E\alpha_2 + G\boldsymbol{\alpha}_3 + S\boldsymbol{\beta}$ where $g(\cdot)$ is the GLM link function, $S_i = E_i\boldsymbol{G}_i$, and $S\boldsymbol{\beta} \sim (0, \tau^2 K_{GE})$ where $K_{GE} = diag(E)K_G diag(E) = diag(E)GG^T diag(E)$. The GESAT then proposes; to estimate $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1\ \alpha_2\ \boldsymbol{\alpha}_3]^T$ and nuisance parameters using ridge regression. Simulations have shown that when linearity assumptions are incorrect, type I error may be slightly inflated (Zhao et al., 2015a). Researchers have also generalized GESAT through similarity regression methods (Zhao et al., 2015a). Using a model that treats both genetic and interaction effects as random, $g(\boldsymbol{\mu}) = X\boldsymbol{\gamma} + \boldsymbol{h}_G + \boldsymbol{h}_{GE}$, this similarity regression approach does not fix a linear form for G. The parameter, $\boldsymbol{\gamma}$, and the nuisance parameters from the variance of $\boldsymbol{h}_G$.

The second approach to gene-environment interactions is to focus on overall genetic marker effects with the goal of discovering novel associations (Hutter et al., 2013). For example, a research study may aim to discover genes that are associated with lung cancer risk through a genome-wide study. In this analysis, researchers may also wish to account for the possibility that some genetic effects differ between smokers and non-smokers. When these types of interactions exist, joint hypothesis tests for both interaction and genetic main effects are typically more powerful than tests for main effects alone. Additionally, if a gene region does not have an interaction effect, testing for overall genetic marker effects may yield similar power as tests for only main effects (Lindström et al., 2009). For these reasons, joint tests have been recommended when researchers are interested in the genetic associations more than whether a gene-environment interaction exists or not (Kraft et al., 2007). In a model such as in Equation 2.9, this type of test would correspond to $H_0 : h_G(\cdot) = h_{GE}(\cdot,\cdot) = 0$.

A main challenge for these types of interaction analyses is developing a way to adequately combine kernels for genetic markers and gene-environment interactions. There are many ways to combine kernels so that the resulting kernel is valid (Shawe-Taylor and Cristianini, 2004); however, the weighted average of K_G and K_{GE} provides an overall genetic kernel that works well (Wu et al., 2013; Zhao et al., 2019). The composite kernel likelihood ratio test (CKLRT) is a kernel-based association test that models the overall genetic effect nonparametrically through this type of weighted average kernel (Zhao et al., 2019). Although the weights could be selected and fixed *a priori*, the CKLRT methodology maximizes the statistic over a range of potential weights. After assuming that the errors are normally distributed, the CKLRT approximates the null distribution of a likelihood ratio test statistic by a Monte Carlo sampling algorithm.

In the CKLRT's methodology, derivations for both the likelihood and null distribution are dependent upon normality assumptions, which may not be appropriate in all cases. As discussed in Section 2.3.2, kernel association tests that assume normality lead to decreased power when the underlying errors are skewed, heavy-tailed, or have undefined moments (Martinez et al., 2020a,2).

2.6 Other Applications in KMR Hypothesis Testing

2.6.1 Microbiome Data

Microbiome research is another field where KMR testing strategies are an important tool, as data may be high dimensional and requiring complex models. Researchers often sequence 16S rDNA from microbial communities, and the sequences are clustered into pseudo-taxa called operational taxonomic units (OTUs) (Schloss and Handelsman, 2005; Zhao et al., 2015b). A common goal is testing the association between these abundances of OTUs and a response such as disease or environmental exposure.

The Microbiome Regression-Based Kernel Association Test (MiRKAT) developed by Zhao et al. (2015b) implements a hypothesis test through the LSKM or logistic kernel machine regression models. From this perspective, $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ represent the response, covariates, and OTU abundances for subject i. MiRKAT also incorporates a standard phylogenetic distance metric of the OTUs, D, into the kernel matrix: $K = -\frac{1}{2}(I - \frac{11^T}{n})D^2(I - \frac{11^T}{n})$. Because microbiome studies often have small sample sizes, the kernel machine regression p-value is computed through permutation testing.

Zhan and Wu (2017) recently created MMiRKAT, a multivariate generalization of MiRKAT. From a general perspective, MMiRKAT is similar to MKMR with traditional microbiome similarity measures incorporated into the kernel matrix. Unlike its univariate counterpart, however, MMiRKAT assumes a continuous phenotype.

2.6.2 Copy Number Variation Data

Many of the KMR testing methods in this review were originally designed for SNP genetic marker data, which can be relatively simple to model. SNPs only involve one numeric quantity, the number of minor alleles at a single locus. Copy number variation (CNV) is another source of genetic variation linked to many disease outcomes such as schizophrenia, autism, and bipolar disorder (Girirajan et al., 2013; St Clair, 2009). CNV refers to a region of the genome that is either repeated or deleted a number of times. The length of a CNV region may be on the scale of entire gene regions or as small as a few base pairs. A certain CNV may also be a deletion, duplication, or normal with varying dosages (Tzeng et al., 2015). These additional features require modification of the traditional variance component-based testing methods for SNP data.

The CNV Collapsing Random Effects Test (CCRET) developed by Tzeng et al. (2015) is a mixed model method designed for detecting associations between a binary or continuous response and rare CNV features. This mixed model approach allows CCRET to be viewed as a generalization of univariate KMR testing tailored to CNV data. The overall model is $g(\mu_i) = \boldsymbol{X}_i^T \boldsymbol{\beta}_0 + h_{DS}(Z_i^{DS}) + h_{GI}(Z_i^{GI}) + h_{Len}(Z_i^{Len})$, a semiparametric model containing functions for the three CNVR features dosage (DS), gene intersection (GI), and length (Len). Researchers can then perform an association test on one of the features of interest. Then the model simplifies to a linear mixed effects model, $g(\mu_i) = \boldsymbol{X}_i^T \boldsymbol{\beta}_0 + \boldsymbol{\beta}_{GI} Z_i^{GI} + \boldsymbol{\beta}_{Len} Z_i^{Len} + h_i(Z_i^{DS})$ with $\mathbf{h}^{DS} \sim N(0, \tau_{DS} K_{DS})$, and a variance component score test can be carried out for $H_0 : \tau_{DS} = 0$.

The CNV Kernel Association Test (CKAT) is a logistic kernel machine method designed to detect associations within a CNVR, using the model $logit[P(Y_i = 1|\boldsymbol{X}_i, R_i)] = \boldsymbol{X}_i^T \boldsymbol{\beta} + h(R_i)$ where R_i is subject i's CNVs in the CNVR (Zhan et al., 2016). CKAT incorporates a user-specified kernel, connects to a logistic mixed model, and performs a variance component score test. Zhan et al. (2016)'s proposed CNVR kernel incorporates a Gaussian kernel for the continuous CNV length multiplied with the binary kernel from the deletion/duplication data. For the binary kernel to work, however, the value when elements are unequal must be greater than zero. Unlike CCRET, all CNV features are incorporated into the random effect. Thus, in scenarios where CNV dosage effect is small, association testing through CKAT may be more powerful (Zhan et al., 2016).

2.7 Discussion

In this review, we have discussed hypothesis testing methods using kernel machines for a variety of response data types. Section 2.3.1 describes tests for normally distributed responses, Section 2.3.2 reviews methods for non-normal data, and Section 2.3.3 covers a generalization for discrete responses. For many of these diverse tests, literature also exists for multivariate extensions as described in Section 2.4. The KMR framework also permits extensions to specialized covariates, such as rare variants (Section 2.4.3) and instances of pleiotropy (Section 2.4.4).

The flexibility of KMR hypothesis testing lies in a common process that is tailored to different situations. In almost all discussed methods, the researcher selects an appropriate kernel matrix and inputs $h(t) = \sum_{i=1}^{N} \alpha_i k(t, \boldsymbol{Z}_i)$ into Equation 2.1 using the Representer Theorem. As this minimization criterion is often equivalent to solving a linear mixed model estimation problem, the analyst then performs a variance component score test, or some variation on such a test. Finally, the p-value is calculated via permutations or asymptotic results.

While this process is well studied, there are still limitations in KMR hypothesis regarding distributional robustness. Many popular KMR testing methods, namely LSKM and MKMR, yield tests that are not robust to heavy tails, multiple modes, skew, and outliers. In the univariate case, QRKM provides some robustness by using a check loss function, but the method could be more flexible. QRKM inference is to a single class of loss function, while robust statistics literature provides a plethora of other loss functions (Huber, 1996; Maronna et al., 2006).

In Chapter 3, we develop a robust kernel association test (RobKAT) that addresses these issues. RobKAT is a variance component score test like LSKM, but the null model is fit using a general loss function that allows robust regression estimates. We implement a computationally efficient algorithm to approximate the null distribution, and we demonstrate through simulation study that RobKAT performs well in the case of non-Gaussian error distributions.

Little is known regarding mixed responses in multivariate kernel machine literature. DKAT does offer a general framework of phenotype kernels to accommodate both continuous, discrete, and mixed responses (Zhan et al., 2017b). However, the proposed method of forming K_Y by multiplying a binary and continuous kernel together was only demonstrated for a two-dimensional data matrix, one binary and one continuous (Zhan et al., 2016). Powerful KMR testing methods for multiple binary responses have recently been proposed, though the case of mixed responses is not addressed (Davenport et al., 2017). Thus, additional options for DKAT phenotype kernels or novel KMR methods to model mixed responses may be further explored in the future.

In Chapter 4, we use a variance component score testing approach to develop a multivariate robust KAT (MrKAT) that performs well for continuous, binary, and mixed responses. By using a general loss function for each individual response, MrKAT becomes robust to non-normal errors and high leverage points when residuals are fit using robust GLM regression. Due to its similarity with RobKAT, we implement a similar fast approximation algorithm when conducting inference.

In Chapter 5, we apply our robust strategies to gene-environment interaction models by developing a robust composite kernel association test (RoCKAT) for overall genetic marker effects. To make the test robust to non-normally distributed errors, RoCKAT fits its null model with robust regression estimates and constructs a variance component score test. To adequately model both the genetic main and interaction effects, our test utilizes a weighted average of kernels and maximizes over a range of weights. Additionally, Chapter 5 develops an algorithm to approximate RoCKAT's null distribution through permutations within exposure groups. This novel scheme is necessary to preserve the exchangability condition in the case of a gene-environment interaction model.

Because KMR's flexibility lends itself to many applications, methodological breakthroughs in the KMR space may solve a plethora of problems. The simple modeling of high-dimensional relationships often is difficult to achieve through other parametric approaches. Thus, understanding and progressing KMR methodology may prove impactful both now and in the future.

CHAPTER

3

ROBUST KERNEL ASSOCIATION TESTING

Previously published article: **Martinez, K., Maity, A., Yolken, R., Sullivan, P., and Tzeng, J.Y. (2020) Robust Kernel Association Testing (RobKAT).** ***Genetic Epidemiology*, 44(3), 272-282.**

3.1 Introduction

Detecting associations between a set of genetic markers, such as SNPs, and a continuous response is a common problem. These tasks lend themselves to multi-marker approaches, such as kernel methods, that analyze variants at the gene level and may gain power due to their lower multiple testing burden. Kernel methods are also appealing over burden-based multi-marker approaches, which treat the overall genetic marker effect as fixed and test for the mean genetic effect with various weighting schemes (Gauderman et al., 2007; Li and Leal, 2008; Madsen and Browning, 2009). Many of these burden-based methods are optimal in the case of additive loci with similar size and direction of effect, but kernel methods typically avoid these loci effect restrictions by viewing the overall genetic effect as random or the output of a nonparametric function.

Least squares kernel machines (LSKM), developed by Liu et al. (2007) and Kwee et al. (2008), are popular kernel methods that utilize a linear mixed model with normally distributed errors. Mixed model methods such as Restricted Maximum Likelihood (REML) may be applied for simultaneous estimation of parameters, and LSKM estimators can be shown to be the best linear unbiased predictors. Score tests can be used for testing the overall genetic marker effect, and due to distributional assumptions this score statistic follows a Chi-Squared mixture under the null hypothesis (Liu et al., 2007). A type of LSKM that has been developed is the sequence kernel association test (SKAT), which incorporates weights into the kernel according to genetic marker minor allele frequencies (Wu et al., 2011). However, the assumption of normally distributed errors limits inference to the conditional mean response. In the case of a skewed or heavy-tailed response, the null distribution of the test statistic does not follow a Chi-Squared mixture, resulting in moderate to severe loss of power.

A variation on LSKMs, the quantile regression kernel machines (QRKM) utilizes quantile regression methods rather than least squares for parameter estimation. When testing the genetic marker effect, Kong et al. (2016) proposes a QRKM score-type statistic based on the subgradient of the check function rather than the traditional score test. Although QRKM is more robust to extreme observations, the method is restricted only to inference on the conditional quantile. While these quantiles offer more choice than in LSKMs, additional flexibility may be desired.

This lack of robust kernel association testing options poses a problem for clinical data sets with non-normal responses, and particularly for the antibody response data from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) study. The CATIE trial aimed to compare the effectiveness of atypical antipsychotics in unrelated patients suffering from schizophrenia and measured patient antibody levels on three neurotropic herpesviruses. As discussed with additional detail in Section 3.4, these antibody levels are notably right-skewed. Even after a log-transformation, the data display a bimodal distribution. Violating the assumption of normality, tests such as LSKM and SKAT may not be appropriate. Although QRKM is valid with such distributions and robust to outliers, QRKM restricts inference to the conditional quantile and limits the loss function used in the analysis. Thus, we propose a robust and general method to test for multi-marker association with this non-normally distributed response.

Our robust kernel association test accommodates any valid loss function and is robust to outliers, heavy tails, and skewed distributions. We achieve this robust framework by generalizing previous kernel methods and eliminating distributional assumptions on the response. While LSKM and SKAT specify a square loss and QRKM specifies a 'check' loss, we follow similar score function derivations with a general loss function subject to a few minor restrictions. The user-specified, general loss function may be taken from the robust statistics literature, rendering RobKAT methodology robust to conditionally non-normal responses. As another score-type statistic, we describe in Section 3.2 that the RobKAT statistic simplifies to LSKM, SKAT, and QRKM in special cases; however, RobKAT's general loss function also provides alternatives to LSKM's conditional mean inference and QRKM's conditional quantile inference. The null distribution of our test statistic and associated p-values can be calculated through a permutation test, as originally proposed for QRKM. Since the permutation testing procedure is computationally inefficient, we instead apply a fast permutation testing method described by Zhan and Wu (2017) that estimates the null distribution of the test statistic through moment matching.

The rest of the article is organized as follows. Section 3.2 details model notation, derivation of our RobKAT statistic, and p-value computation. Section 3.3 then demonstrates the validity, type I error control, and power of our test through simulation study comparing to normal-based methods. We show that our proposed testing procedure is robust to non-normal distributions, even providing a powerful and type I error-controlled test when errors follow a Cauchy distribution. Although comparable to SKAT when the response is normal, RobKAT shows greater power in all non-normal cases that we explored. In Section 3.4, we apply our proposed testing procedure to detect the association between genes in chromosome six, including the MHC region, and antibody levels using CATIE clinical trial data. We find four significantly associated gene sets using the skewed antibody response, three of which are not detected by SKAT. In Section 3.5, we conclude that RobKAT's computational efficiency and robust nature make it a useful procedure even on the genome-wide scale.

3.2 Methodology

3.2.1 Model Specification and Testing Problem

Suppose we observe data $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, n$, where Y_i is a scalar response of interest, $\boldsymbol{Z}_i$ denotes a $p \times 1$ vector of genetic covariates, and $\boldsymbol{X}_i$ denotes a $q \times 1$ vector of other confounders observed for the ith subject. We consider the model

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + h(\boldsymbol{Z}_i) + \epsilon_i,$$

where $\boldsymbol{\beta}$ is an unknown coefficient vector, $h(\cdot)$ is an unknown function quantifying the effect of the genetic covariates, and ϵ_i, $i = 1, \ldots, n$ are independent mean zero random errors. Testing for the overall genetic marker effect corresponds to a hypothesis test for $H_0 : h(\cdot) = 0$.

One issue in kernel machine regression is that the potentially nonparametric function $h(\cdot)$ may be difficult to model. We instead assume that $h(\cdot)$ is an element of a reproducing kernel Hilbert space (RKHS), $\mathscr{H}_K$, which can be uniquely defined by a positive semi-definite kernel function $k(\cdot, \cdot)$ (Shawe-Taylor and Cristianini, 2004; Wahba, 1990). Due to certain mathematical properties of RKHS, this structure allows us to represent $h(\boldsymbol{Z}_i)$ with images of the kernel function. Further, we may summarize this information in the positive semi-definite matrix $\mathbb{K}$, whose (i, j) element is defined by $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$. To obtain estimates of $h(\cdot)$, researchers often minimize the penalized loss criterion

$$\min_{\boldsymbol{\beta}, h \in \mathscr{H}_K} \left[\sum_{i=1}^{N} \rho\{Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta} - h(\boldsymbol{Z}_i)\} + \lambda \, ||h||^2_{\mathscr{H}_K} \right] \tag{3.1}$$

We note that the nonparametric functional form of $h(\cdot)$ may be infinite-dimensional. Regarding the form of $\rho(\cdot)$, the least squares kernel machine methods specify $\rho(t) = t^2/2$ in order to implement squared-loss.

A key observation is that estimates of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{h}}$ from Equation 3.1 match best linear unbiased estimators obtained from the mixed effects model, $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + h_i + \epsilon_i$ where $\boldsymbol{h} \sim N(0, \tau\mathbb{K})$ and $\epsilon_i \sim N(0, \sigma^2)$ (Liu et al., 2007). Further, estimates for the remaining parameters, τ and σ^2, can be computed using restricted maximum likelihood estimates. With the mixed effects model in mind, LSKM performs a hypothesis test for $H_0 : \tau = 0$ using a variance component score test. Leveraging the model's normality assumptions, inference on the test statistic comes from a Chi-Squared mixture and a Satterthwaite approximation (Liu et al., 2007).

QRKM also uses the penalized minimization criterion but instead specifies a check function for the loss: $\rho(t) = \eta t \ \mathbb{I}_{\{t\geq 0\}} + (\eta - 1)t \ \mathbb{I}_{\{t<0\}}$ where $\eta \in (0,1)$ is the conditional quantile of interest (Kong et al., 2016). When researchers are interested in association with the conditional median response, for example, $\eta = 0.5$ and $\rho(t) = 0.5|t|$. Estimates for $\boldsymbol{\beta}$ are the values that minimize $\sum_{i=1}^{N} \rho(Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta})$, Equation 1 under the null hypothesis, and can be computed using quantile regression packages such as `quantreg` in R. As with LSKM, the QRKM test statistic follows the variance component score test procedure. Since the check loss function is not differentiable, however, the derivation uses the subgradient, $\psi(t) = \eta \mathbb{I}_{\{t>0\}} + (\eta - 1)\mathbb{I}_{\{t<0\}} + (\eta - B)\mathbb{I}_{\{t=0\}}$, where B is the outcome of a $Bern(\eta)$ random variable. Without using an assumption of normality, the null distribution of the QRKM test statistic and resulting p-value are computed using a permutation algorithm (Kong et al., 2016).

In RobKAT, we derive the variance component score statistic but with a general ρ, subject only to a small number of conditions. In this way, LSKM and QRKM will be special cases of our robust and generalized method. Let ρ be a differentiable function $\rho : \mathbb{R} \to \mathbb{R}^+$ such that

1. $\rho(0) = 0$
2. $\rho(x)$ is a convex function of x
3. $\psi(x) = \frac{d}{dx}\rho(x)$ exists and is non-decreasing

The first assumption is taken from standard robust statistics procedures described in Maronna et al. (2006), and the last two are modifications to exclude redescending ψ-functions. Violations of assumptions two and three can be accommodated; however performance of the resulting test may be suboptimal. The exclusion of these redescending ψ-functions and such performance impairment are discussed further in Section 3.5. Note that we do not make any assumptions on the distribution of ϵ_i.

3.2.2 Test Statistic

To motivate our method, we first consider the simpler case in which $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \boldsymbol{\gamma}$ so that the underlying model is $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \boldsymbol{\gamma} + \epsilon_i$. Testing $H_0 : h(\cdot) = 0$ in this case is equivalent to testing

$H_0 : \gamma = 0$. A score-type test follows from minimizing the criterion

$$L(\boldsymbol{\beta}, \gamma) = \sum_{i=1}^{n} \rho\left(\frac{\epsilon_i}{s}\right) = \sum_{i=1}^{n} \rho\left(\frac{Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta} - \boldsymbol{Z}_i^T \gamma}{s}\right)$$

where s is the scale parameter for the distribution of ϵ_i. Differentiating and setting equal to zero, the corresponding score function then yields the M-estimation equations

$$S(\boldsymbol{\beta}, \gamma) = \sum_{i=1}^{n} \boldsymbol{Z}_i \ \psi\left(\frac{\epsilon_i}{s}\right) = \sum_{i=1}^{n} \boldsymbol{Z}_i \ \psi\left(\frac{Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta} - \boldsymbol{Z}_i^T \gamma}{s}\right) = 0$$

Let $\widehat{\boldsymbol{\beta}}$ be an estimate of $\boldsymbol{\beta}$ under the null hypothesis $H_0 : h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \gamma = 0$, and let $\widehat{\epsilon_i} = Y_i + \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}$. Under H_0, fitting $\widehat{\boldsymbol{\beta}}$ and accordingly $S(\widehat{\boldsymbol{\beta}}, \gamma)|_{\gamma=0}$ reduces to the simple task of fitting the linear regression model $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \epsilon_i$, which does not require knowledge of the form of $h(\boldsymbol{Z}_i)$ and can be done easily using standard robust regression methods. When obtaining an M-estimate of s, minimizing $\sum_{i=1}^{n} \rho(\frac{\epsilon_i}{s})$ can ensure that regression estimates are invariant. However, we desire an estimate of s that is also robust. One solution is to use Huber's Proposal Two robust estimate of scale, which solves

$$\frac{1}{n-p} \sum_{i=1}^{n} \psi^2\left(\frac{Y_i - \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}}{s}\right) = E_{\Phi}(\psi^2(\boldsymbol{\epsilon}))$$

where Φ denotes an expectation under the assumption of a standard normal error distribution Huber (1977); Schrader and Hettmansperger (1980). Huber's Proposal Two estimate is preferred over other robust estimates of scale as it has desirable convergence properties under our assumption of convex ρ (Schrader and Hettmansperger, 1980).

We then derive the test statistic $T = S(\widehat{\boldsymbol{\beta}}, \gamma)^T S(\widehat{\boldsymbol{\beta}}, \gamma) = \boldsymbol{w}^T \mathbb{K} \boldsymbol{w}$. The last expression uses the simplifications $\boldsymbol{w}$ and $\mathbb{K}$ where the i^{th} element of $\boldsymbol{w}$ is defined by $w_i = \psi(\frac{\widehat{\epsilon_i}}{s})$ and the $(i,j)^{th}$ element of $\mathbb{K}$ is defined by $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = \boldsymbol{Z}_i^T \boldsymbol{Z}_j$.

This $\mathbb{K}$ can be generalized to any kernel function $k(.,.)$. The effect of $k(.,.)$ is linked with the form of $h(\boldsymbol{Z}_i)$. In the previous derivation, setting $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i \gamma$, a model with only main genetic-marker effects, corresponds to the linear kernel $\mathbb{K}$ with entries defined by the function $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = \boldsymbol{Z}_i^T \boldsymbol{Z}_j$. If instead for $\boldsymbol{Z}_i \in \mathbb{R}^p$ we chose in our derivation $h(\boldsymbol{Z}_i) = \sum_{j<k} Z_{ij} Z_{ik} \gamma_{jk} + \sum_{j=1}^{p} Z_{ij}^2 \gamma_{2j}$, a model with quadratic main effects and all two-way interactions, then we would obtain a quadratic kernel whose entries are defined by $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = (\boldsymbol{Z}_i^T \boldsymbol{Z}_j)^2$. The kernel function $k(.,.)$ generates a function

space that contains $h(\boldsymbol{Z}_i)$. This theoretical framework thus represents $h(\boldsymbol{Z}_i)$ by a linear combination $h(\boldsymbol{Z}_i) = \sum_j \alpha_j \; k(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$ for some weights α_j, which gives a clear connection between choice of $k(.,.)$ and form of $h(\boldsymbol{Z}_i)$. Since $\mathbb{K}$ is symmetric positive semi-definite, $k(.,.)$ is often thought of as a similarity metric between two individuals, which may be more intuitive to specify than the form of $h(\boldsymbol{Z}_i)$ itself. Further, any symmetric positive semi-definite matrix describing similarity may be used for $\mathbb{K}$. When $\boldsymbol{Z}_i$ represents genetic markers, for example, the IBS kernel is defined as $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = \frac{1}{2p}\sum_{k=1}^{p}[2\mathbb{I}\{Z_{ik} = Z_{jk}\} + \mathbb{I}\{|Z_{ik} - Z_{jk}| = 1\}]$ or the proportion of alleles shared between individuals i and j.

By instead defining the form of $h(\boldsymbol{Z}_i)$ through $k(.,.)$, we generalize our test statistic to any valid kernel function. Thus, we produce a test statistic that is intuitively specified through a kernel function, $k(.,.)$, and loss function, ρ. Note that any ρ function following the general assumptions listed above is easily incorporated into this test statistic. For example, the case in which $\rho(t) = \frac{1}{2}t^2$ leads to the usual LSKM test statistic and $\rho(t) = 0.5|t|$ leads to the QRKM statistic (Kong et al., 2016; Liu et al., 2007). Another common choice of ρ is from the family of Huber functions defined by

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & |x| \leq \kappa \\ \kappa(|x| - \frac{\kappa}{2}) & |x| > \kappa \end{cases}$$

where the value of κ may be chosen for a desired asymptotic variance (Huber, 1977; Maronna et al., 2006). For $-\kappa < x < \kappa$, the function is quadratic as in least squares, and for $|x| > \kappa$ the function is linear as in quantile regression. Thus, this family may be seen as an intermediate between least squares and quantile regression.

To calculate our general kernel machine test statistic, we first fit the model under the null hypothesis, $H_0 : h(\cdot) = 0$, by regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$ and minimizing $\sum_{i=1}^{n}\rho(\epsilon_i)$ with the iterated re-weighted least squares algorithm. Then we use the null model to compute $\widehat{s}$, Huber's Proposal Two robust estimate of scale. We then compute the score-type statistic described above, $T = \boldsymbol{w}^t\mathbb{K}\boldsymbol{w}$, where the i^{th} element of $\boldsymbol{w}$ is defined by $w_i = \psi(\widehat{\epsilon}_i/\widehat{s})$.

3.2.3 Null Distribution of the Test Statistic

It is possible to obtain the null distribution for a statistic of the form $\boldsymbol{w}^T\mathbb{K}\boldsymbol{w}$ using permutations (Kong et al., 2016). However, there is a substantial computational burden associated with permutation methods when sample sizes are moderate to large. Slightly modifying our test statistic by centering $\mathbb{K}$, we can apply the method of Zhan and Wu (2017) that avoids much of the computational burden by instead approximating the null permutation distribution of previous methods. Following this fast approximation algorithm, we instead consider $T = \boldsymbol{w}^T\mathbb{P}\mathbb{K}\mathbb{P}\boldsymbol{w}$, where $\mathbb{P} = \mathbb{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the centering matrix. We then see that $T = \boldsymbol{w}^T\mathbb{P}\mathbb{K}\mathbb{P}\boldsymbol{w} = tr(\boldsymbol{w}^T\mathbb{P}\mathbb{K}\mathbb{P}\boldsymbol{w}) = tr(\mathbb{P}\mathbb{K}\mathbb{P}\boldsymbol{w}\boldsymbol{w}^T) = tr\{\mathbb{P}\mathbb{K}^{\frac{1}{2}}(\mathbb{K}^{\frac{1}{2}})^T\mathbb{P}\boldsymbol{w}\boldsymbol{w}^T\}$, and there exists a matrix $\mathbb{K}^{\frac{1}{2}}$ such that $\mathbb{K} = \mathbb{K}^{\frac{1}{2}}(\mathbb{K}^{\frac{1}{2}})^T$ because $\mathbb{K}$ is symmetric and positive semi-definite.

The form $tr\{\mathbb{P}\mathbb{K}^{\frac{1}{2}}(\mathbb{K}^{\frac{1}{2}})^T\mathbb{P}\boldsymbol{w}\boldsymbol{w}^T\}$ can then be leveraged with previously established distribution approximation methods. Primarily, Zhan and Wu (2017) utilize the fast techniques developed for RV-type statistics of the form $tr(\mathbb{P}\mathbb{A}\mathbb{A}^T\mathbb{P}\mathbb{B}\mathbb{B}^T)$, which approximate the null permutation distribution by moment matching to a Pearson type III distribution (Josse et al., 2008). For the moment matching in this technique, the method uses the analytical expressions of the first three moments of $T = tr(\mathbb{X}^T\mathbb{X}\mathbb{Y}^T\mathbb{Y})$ derived by Kazi-Aoual et al. (1995). Although it is possible to approximate the empirical null distribution with the normal, log-normal, or Edgeworth distributions, the Pearson type III approximation is both accurate and more efficient than these competitors (Josse et al., 2008). As described by Zhan and Wu (2017), we use the centered test statistic, analytical moments, and approximated Pearson type III density to obtain a p-value for our test statistic. In the case of $\rho(t) = 0.5|t|$, this computationally efficient algorithm has been shown to be about 6,000 times faster than the standard permutation test with comparable power and type I error (Zhan and Wu, 2017).

3.3 Simulation Study

We perform a simulation study to investigate the properties of our proposed method with the robust Huber ψ-function when the data have various error distributions, the genetic effect takes various forms, and the marker set is either small or large. We then compare the performance of our proposed method against that of SKAT, which is not robust and assumes a normally distributed error. We also include the special case of our robust method where $\rho(t) = 0.5|t|$, which corresponds to the QRKM method with $\tau = 0.5$. We refer to the QRKM $\rho(t) = 0.5|t|$ as least absolute deviation, LAD.

3.3.1 Simulation Study Design

We generate data from the model

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + c * h(\boldsymbol{Z}_i) + \epsilon_i$$

where the scalar, c, is multiplied with the nonparametric function value to control genetic marker effect sizes in our simulations. We generate elements of $\boldsymbol{X}_i$ from $N(0, \mathbb{I}_2)$ and set $\boldsymbol{\beta} = \mathbf{1}_2$, and perform simulation studies with both the linear $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \mathbf{1}_9$ and the nonlinear $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{p} \eta_k \boldsymbol{Z}_{ik} + 2\sum_{k=2}^{p} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$, where $\eta_1 = 0.4$, $\eta_2 = \eta_3 = ... = \eta_9 = 0.7$ and $\boldsymbol{Z}_i$ is a randomly sampled nine-SNP genotype of the *SLC17A1* gene in the CATIE antibody study. We compare testing results from i.i.d. error terms, ϵ_i, generated from t_3, χ^2_1, $N(0,1)$, and $Cauchy(0,1)$. Additionally, we compare the bimodal setting through a mixture of normal distributions, where $\epsilon_i = BW_0 + (1-B)W_{10} - 10(1-\theta)$ where $W_X \sim N(X, 1)$, $B \sim Bern(\theta)$, and $\theta \in \{0.9, 0.7\}$. For all testing procedures, we use an IBS kernel and two sample sizes ($n = 100, 300$). We use the Huber function defined by $k = 1.345$ for 95% efficiency of least squares when the errors are normally distributed.

To examine Type I error, we consider significance levels $\alpha = 0.01$ and $\alpha = 0.05$. We set $c = 0$ for the null model under $H_0 : h(\boldsymbol{Z}_i) = 0$ and use $N = 10,000$ Monte Carlo replications. As $c = 0$, the type I errors are identical for both linear and nonlinear $h(\boldsymbol{Z}_i)$ functions. To investigate the power of our testing procedure compared to SKAT, we test increasing values of c and use $N = 1,000$ Monte Carlo replications for each value.

We then perform a second simulation study to explore high-dimensional and low-causal gene regions. Rather than sampling $\boldsymbol{Z}_i$ from the nine-SNP *SLC17A1* gene with nine causal markers, this setting instead samples from the 787-SNP Major Histocompatibility Complex (MHC) gene region, and only two of these SNPs are used when generating the response. The two causal markers are common with MAFs of 0.44 and 0.18, respectively.

3.3.2 Simulation Results

The type I error from the low-dimensional simulation setting ($p = 9$) are displayed in Table 3.1, and our testing method produces reasonable type I errors for each of the error distributions. Although SKAT produces similar type I errors for t_3, χ_1^2, and $N(0,1)$, the type I errors with Cauchy distributed ϵ_i are notably deflated. For the simulation study exploring high-dimensional SNP-sets ($p = 787$) shown in Table 3.2, our robust methodology using Huber and LAD also controls type I error for each of the error distributions. The type I errors when using SKAT are conservative for all distributions in Table 2, although SKAT is known to be conservative when sample sizes are small Wu et al. (2011). Appendix Tables A.2 and A.6 show that SKAT becomes less conservative with a larger sample size ($n = 300$).

Table 3.1 Simulation study for type I error of RobKAT (Huber loss), QRKM (LAD loss), and SKAT. The simulation includes IBS kernel, sample size $n = 100$, and 9 simulated SNPs per sample. Estimates show proportion of $10,000$ iterations with p-value below α cutoff.

Error	α	Loss Function (ρ)		
		Huber	LAD	SKAT
t_3	0.01	0.0100	0.0083	0.0081
	0.05	0.0456	0.0458	0.0466
χ_1^2	0.01	0.0105	0.0113	0.0080
	0.05	0.0532	0.0507	0.0475
$N(0,1)$	0.01	0.0104	0.0089	0.0105
	0.05	0.0473	0.0492	0.0455
$Cauchy$	0.01	0.0120	0.0109	0.0032
	0.05	0.0488	0.0509	0.0319
10% Mix	0.01	0.0115	0.0094	0.0088
	0.05	0.0484	0.0476	0.0494
30% Mix	0.01	0.0108	0.0103	0.0096
	0.05	0.0484	0.0487	0.0489

The power curves for the nine-SNP setting and nominal level 0.05 are shown in Figures 3.1-3.2. For standard normal errors, RobKAT with a Huber ψ-function displays similar power as SKAT and a moderate power gain over LAD. Visualized in the Figure 3.1 power curves for standard normal errors, our robust method shows a maximum power gain of 49.6% over the LAD ψ-function when data was generated with a linear $h(\boldsymbol{Z}_i)$ and a 42.6% maximum power gain for data generated with a

Table 3.2 Simulation study for type I error of RobKAT (Huber loss), QRKM (LAD loss), and SKAT. The simulation includes IBS kernel, sample size $n = 100$, and 787 simulated SNPs per sample. Estimates show proportion of 10,000 iterations with p-value below α cutoff.

Error	α	Loss Function (ρ)		
		Huber	LAD	SKAT
t_3	0.01	0.0096	0.0089	0.0028
	0.05	0.0461	0.0484	0.0269
χ_1^2	0.01	0.0117	0.0107	0.0041
	0.05	0.0508	0.0454	0.0268
$N(0,1)$	0.01	0.0103	0.0105	0.0037
	0.05	0.0491	0.0522	0.0270
$Cauchy$	0.01	0.0085	0.0097	0.0035
	0.05	0.0471	0.0470	0.0226
10% Mix	0.01	0.0100	0.0102	0.0033
	0.05	0.0473	0.0474	0.0265
30% Mix	0.01	0.0083	0.0117	0.0031
	0.05	0.0461	0.0504	0.0264

nonlinear $h(\boldsymbol{Z}_i)$. For the χ_1^2 skewed errors, our robust method has similar power to LAD and has power that is up to double that of SKAT in both the linear and nonlinear settings. In the case of heavy-tailed t_3-distributed errors, our robust method performs similarly to the other methods but, as visualized in Figure 3.1, we see a moderate maximum power gain of 24.6% over LAD and 51% over SKAT when data was generated with a linear $h(\boldsymbol{Z}_i)$ (20.5% and 42.4% for nonlinear $h(\boldsymbol{Z}_i)$). For Cauchy-distributed errors with very heavy tails, however, SKAT loses a substantial amount of power, while the Huber ψ-function and LAD ψ-function methods perform reasonably well. For the power curves shown in Figure 3.1, the power using Huber loss is observed to be at most 8 times that of SKAT when data was generated with a linear $h(\boldsymbol{Z}_i)$ and 7 times for nonlinear $h(\boldsymbol{Z}_i)$.

For the 30% normal mixture distribution in Figure 3.2, the Huber ψ-function performs similarly to SKAT, while the LAD ψ-function has a notably greater power than the other two methods. For the 10% normal mixture distribution, the Huber method is more powerful than SKAT and similar to LAD. The association tests under the 30% bimodal mixture overall are less powerful than the association test under the 10% bimodal mixture.

These trends are similar between the linear and nonlinear form of $h(\boldsymbol{Z}_i)$. However, the nonlinear modeling of the genetic markers required smaller effect size for maximal power. The previously described trends for power are also consistent when the sample size increases to $n = 300$ (See

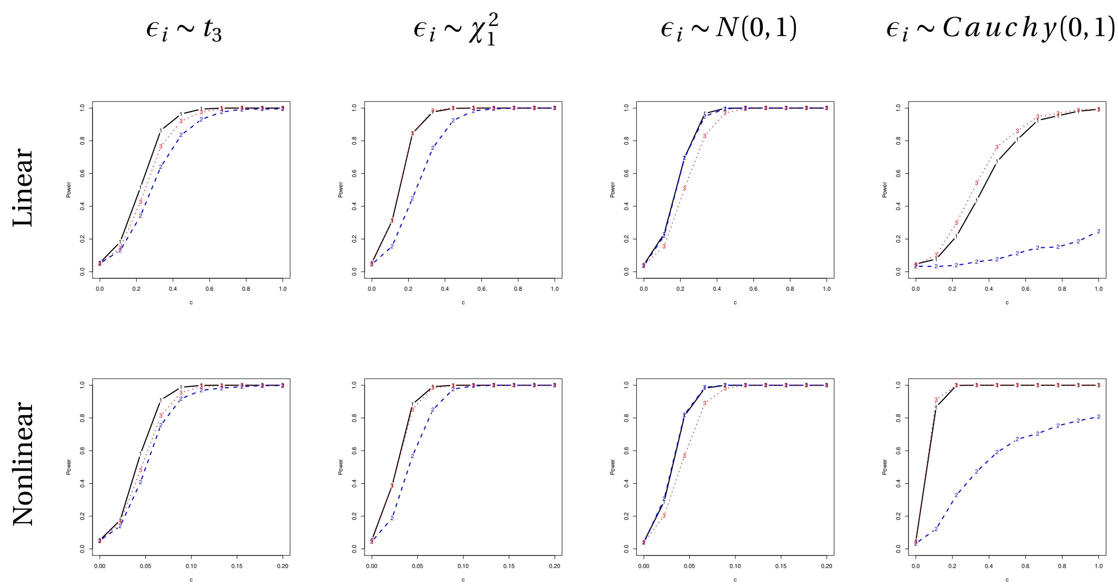
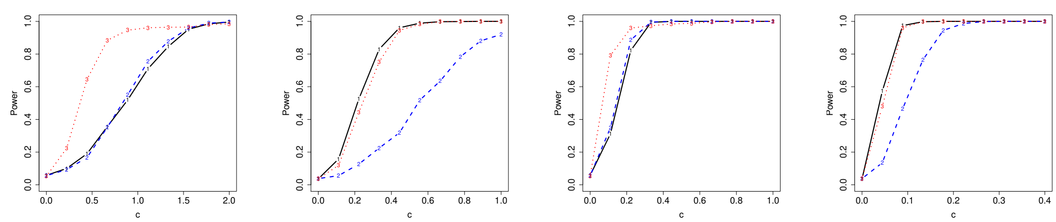


Figure 3.1 Power simulations of RobKAT (black solid line), SKAT (blue dashed line), and LAD (red dotted line). The simulation uses IBS kernel, sample size $n = 100$, $p = 9$ SNPs, and 1,000 iterations at each setting of c. The top row corresponds to the linear case $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \boldsymbol{1}_9$, and the bottom row corresponds to the nonlinear case $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{p} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{p} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$.



(a) 30% mixture; n=100 $h(\boldsymbol{Z}_i)$ linear

(b) 10% mixture; n=100 $h(\boldsymbol{Z}_i)$ linear

(c) 30% mixture; n=100 $h(\boldsymbol{Z}_i)$ nonlinear

(d) 10% mixture; n=100 $h(\boldsymbol{Z}_i)$ nonlinear

Figure 3.2 Power simulations with bimodal errors using RobKAT (black solid line), SKAT (blue dashed line), and LAD (red dotted line). The simulation uses IBS kernel, sample size $n = 100$, $p = 9$ SNPs, and 1,000 iterations at each setting of c. Figures (a) and (b) show the linear case $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \boldsymbol{1}_9$, and figures (c) and (d) display nonlinear case $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{p} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{p} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$.

Appendices A.2 and A.4).

As summarized in Figures 3.3-3.4, we see that the Huber loss function is superior in terms of power when the SNP-set is large compared to the sample size. Although some values of c exist in the Cauchy and 30% bimodal settings where the LAD loss is more powerful, the Huber loss function results in more power for most of the c values. For all other error distributions considered in this simulation study, Huber loss yielded the most powerful test. These trends for power are also consistent when the sample size increases to $n = 300$ (See Appendix A.7).

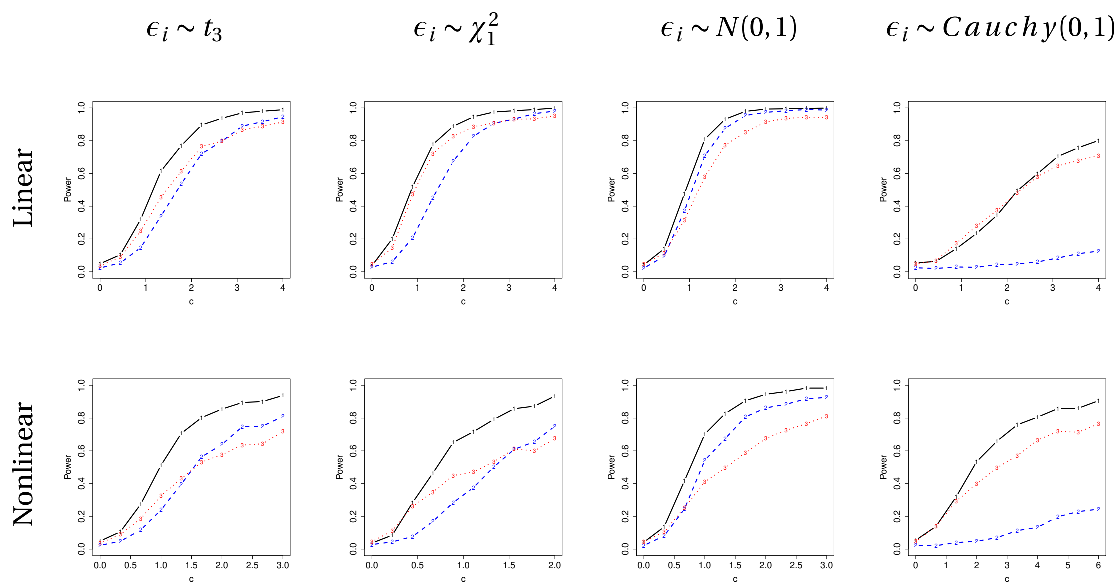
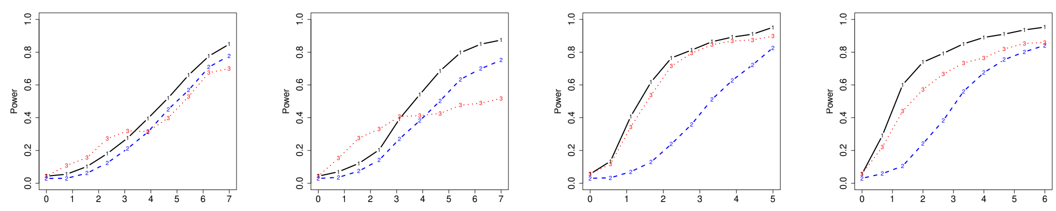


Figure 3.3 Power simulations of RobKAT (black solid line), SKAT (blue dashed line), and LAD (red dotted line). The simulation uses IBS kernel, sample size $n = 100$, 1,000 iterations at each setting of c, 787 simulated SNPs per sample, and 2 causal SNPs. The top row corresponds to the linear case $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_{i1} + \boldsymbol{Z}_{i2}$, and the bottom row corresponds to the nonlinear case $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{2} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{2} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$.



(a) 30% mixture; n=100 $h(\boldsymbol{Z}_i)$ linear

(b) 30% mixture; n=100 $h(\boldsymbol{Z}_i)$ nonlinear

(c) 10% mixture; n=100 $h(\boldsymbol{Z}_i)$ linear

(d) 10% mixture; n=100 $h(\boldsymbol{Z}_i)$ nonlinear

Figure 3.4 Power simulations of RobKAT (black solid line), SKAT (blue dashed line), and LAD (red dotted line). The simulation uses IBS kernel, sample size $n = 100$, 1,000 iterations at each setting of c, 787 simulated SNPs per sample, and two causal SNPs. Figures (a) and (b) show the linear case $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_{i1} + \boldsymbol{Z}_{i2}$, and figures (c) and (d) display nonlinear case $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{2} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{2} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$.

3.4 Application to CATIE Herpesvirus Data

Several genes in chromosome six, including the Major Histocompatibility Complex (MHC), region have been found to have an established connection to schizophrenia, and previous GWAS analyses have identified multiple SNPs in this region associated with schizophrenia (Aberg et al., 2013; International Schizophrenia Consortium et al., 2009; Shi et al., 2009). As described in Lieberman et al. (2005), the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) study aimed to compare the effect of various atypical antipsychotics on unrelated patients' schizophrenia symptoms. DNA samples were donated from 51% of the 1,460 patients in the CATIE study, and genotyping was performed for 492,000 SNPs (Sullivan et al., 2008).

Genetic variation in chromosome six is also thought to contribute to immune function and response to infectious diseases, such as neurotropic herpesviruses. This class of infectious agent has been linked to cognitive impairments, and in the case of Herpes Simplex Virus Type 1 have shown significant association with aspects of the decreased cognition found in schizophrenia disorders (Yolken et al., 2011). At baseline in the CATIE trial, participants provided a blood sample from which investigators measured Ig class antibodies to three neurotropic herpesviruses: Herpes Simplex Virus type one (HSV1), Herpes Simplex Virus type two (HSV2), and Cytomegalovirus (CMV) (Yolken et al., 2011).

We aim to apply our robust kernel machine methodology to determine genes near the MHC region with association to levels of three herpesvirus antibodies: HSV1, HSV2, and CMV. We investigate 12 genes on chromosome 6p22.1 including MHC region as was done previously in Maity et al. (2012) and Davenport et al. (2017). For each gene, we perform our proposed kernel association test with the Huber ψ-function and three different kernels (linear, quadratic, IBS), controlling for age and sex. We use a Bonferroni multiple testing correction to an $\alpha = 0.05$ significance level ($\alpha^* = 0.05/(12*3*3) = 0.00046$) as we are performing tests for 12 genes, 3 kernels, and 3 methods. For comparison, we also apply SKAT and LAD to analyze each of the 12 candidate genes with the same three kernel functions. The results for our analyses are summarized in Table 3.3.

For the HSV1 and CMV antibody responses, no significant associations were detected for any gene with any loss function. For the HSV2 antibody response and each kernel, both SKAT and RobKAT detected significant association with the MHC region. However, our robust kernel association test was able to detect significant associations to the *POM121L2* and *SLC17A1* genes, while SKAT and LAD did not detect any significant associations. LAD detected a significant association with the *HIST1H2BJ* gene, while the other ψ-functions did not.

It is sensible that both SKAT and the proposed test detected the MHC region, which has the most established association with both schizophrenia and immune function. However, the data for each antibody response are skewed (see Figure A.5 in Appendix A.5 for data distributions). As our previous simulations suggest improved power in non-normally distributed data sets, perhaps our model detected a signal in the *SLC17A1* and *POM121L2* genes whereas SKAT failed. The association between *POM121L2* and schizophrenia has been previously noted by Aberg et al. (2013). The power described in the simulation study represents a probability, so it is not unreasonable that LAD detected *HIST1H2BJ* gene when our robust test did not.

Table 3.3 Data analysis of CATIE data using RobKAT (huber loss), QRKM (LAD loss), and SKAT. The significance threshold was determined via a Bonferroni correction based on 12 gene analyses, 3 kernels, and 3 methods ($\alpha = 0.05/(12*3*3) = 0.00046$). Significant results are shown in bold.

			HSV1			HSV2			CMV		
gene	SNPs	ψ	linear	quadratic	IBS	linear	quadratic	IBS	linear	quadratic	IBS
BTN2A1	4	SKAT	0.07679	0.15524	0.10567	0.00176	0.00335	0.00494	0.00150	0.00759	0.00236
		huber	0.00803	0.02631	0.01436	0.00541	0.00639	0.02336	0.00155	0.00863	0.00251
		LAD	0.03475	0.10657	0.04764	0.00106	0.00150	0.00514	0.01024	0.03866	0.01116
BTN2A2	5	SKAT	0.05244	0.08935	0.05702	0.05251	0.06792	0.07835	0.00324	0.00763	0.00286
		huber	0.01515	0.03424	0.01131	0.01443	0.02229	0.01920	0.00517	0.01266	0.00425
		LAD	0.11892	0.21728	0.08933	0.03188	0.07616	0.03467	0.01244	0.01766	0.01149
BTN3A2	5	SKAT	0.03494	0.02282	0.01762	0.07271	0.06011	0.08863	0.04910	0.03126	0.07955
		huber	0.01431	0.00673	0.01121	0.15886	0.19427	0.19263	0.07666	0.04850	0.11918
		LAD	0.00384	0.00212	0.00339	0.84444	0.73582	0.90418	0.31518	0.23253	0.44299
HIST1H2AG	2	SKAT	0.92213	0.93267	0.70116	0.19853	0.23608	0.25445	0.43686	0.66758	0.21429
		huber	0.72157	0.81800	0.85490	0.06201	0.12390	0.03250	0.38483	0.55682	0.25441
		LAD	0.48233	0.38472	0.80218	0.00241	0.00306	0.00407	0.56707	0.73559	0.45521
HIST1H2BJ	2	SKAT	0.79122	0.58380	0.70037	0.01310	0.01899	0.01649	0.18609	0.30461	0.11541
		huber	1.00000	1.00000	1.00000	0.00121	0.00236	0.00129	0.16770	0.25322	0.13759
		LAD	1.00000	1.00000	1.00000	**0.00028**	**0.00018**	0.00098	0.27022	0.21404	0.26863
MHC	787	SKAT	0.04048	0.08628	0.04483	**0.00010**	**0.00016**	**0.00013**	0.10080	0.14007	0.10371
		huber	0.00877	0.01932	0.00660	**2.2314e-06**	**1.7107e-06**	**2.4198e-06**	0.05149	0.06349	0.05627
		LAD	0.03487	0.02926	0.02175	0.00419	0.00159	0.00180	0.07406	0.07532	0.11273
NOTCH4	24	SKAT	0.04606	0.04072	0.06212	0.26256	0.40066	0.28563	0.60714	0.93015	0.64687
		huber	0.03791	0.03982	0.05217	0.06431	0.21994	0.08140	0.43858	0.79330	0.48190
		LAD	0.04826	0.05320	0.07832	0.16485	0.44452	0.25259	0.28569	0.53319	0.33174
POM121L2	4	SKAT	0.26827	0.36745	0.20218	0.00073	0.00146	0.00076	0.10933	0.16350	0.07037
		huber	0.21433	0.29563	0.15850	**0.00023**	**0.00042**	**0.00021**	0.09777	0.13930	0.07341
		LAD	0.03044	0.05003	0.02184	0.00725	0.01584	0.00337	0.01474	0.01903	0.00989
PRSS16	1	SKAT	0.41067	0.41067	0.41067	0.65578	0.65578	0.65578	0.14173	0.14173	0.14173
		huber	0.24503	0.24503	0.24226	0.32980	0.32980	0.33131	0.07214	0.07214	0.07305
		LAD	0.09568	0.10038	0.10510	0.26890	0.26890	0.26107	0.03620	0.03620	0.03802
SLC17A1	9	SKAT	0.19398	0.17851	0.32453	0.00122	0.00396	0.00250	0.00529	0.01382	0.00961
		huber	0.08473	0.09419	0.14602	**0.00020**	0.00070	0.00062	0.00492	0.01555	0.00913
		LAD	0.24536	0.33055	0.39316	0.00400	0.00265	0.00420	0.03292	0.07575	0.04848
SLC17A3	13	SKAT	0.57656	0.40083	0.66907	0.02843	0.07313	0.05747	0.09016	0.05664	0.06619
		huber	0.48986	0.39620	0.43752	0.01814	0.05846	0.03350	0.10804	0.06930	0.10833
		LAD	0.97778	0.78567	0.66117	0.05555	0.17217	0.07699	0.16769	0.14177	0.13404
ZNF184	8	SKAT	0.58984	0.64917	0.61829	0.29280	0.19342	0.33406	0.32298	0.32206	0.44491
		huber	0.46517	0.51445	0.63247	0.42303	0.39284	0.61290	0.30613	0.30321	0.39351
		LAD	0.41699	0.46274	0.54990	0.58686	0.69201	0.61074	0.10163	0.10491	0.15728

3.5 Discussion

We have developed the methodology of a general and robust kernel association test for the model $Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + h(\boldsymbol{Z}_i) + \epsilon_i$ with the score-type test statistic $T = \boldsymbol{w}^t\mathbb{K}\boldsymbol{w}$. Our kernel association test is general, and many previously established kernel association tests such as SKAT and QRKM are special cases when specific loss functions are chosen. RobKAT is flexible in that it places few assumptions on both the choice of kernel and the choice of loss function. The test also does not place any assumptions, including existence of moments, on the error distribution. We then implemented RobKAT with fast permutation testing techniques, which also provides the null distribution of our test statistic without assuming a response distribution. This permutation testing approach is key, as the LSKM Chi-squared mixture null distribution with our statistic decreases power and fails to control for type I error in small and moderate sample sizes (See Appendix A.6 for simulation evidence).

We ran a simulation study that included an IBS kernel, three different loss functions (Huber, LAD, SKAT), and various response distributions ($t_3, \chi_1^2, N(0,1), Cauchy(0,1)$) to investigate the performance of our robust kernel association test. In all distributional settings with the nine-SNP gene region, the type I error was controlled and the robust tests showed similar or greater power than SKAT. In the 787-SNP simulation setting where the genetic marker set is large compared to the sample size, we see that the Huber loss function has more power overall than both LAD and SKAT.

As no single loss function is universally superior across response distributions, there is a demonstrated benefit of allowing a variety of loss functions, which is a feature of RobKAT. However, the robust test with a Huber loss function is the most powerful when analyzing large SNP-sets and is among the most powerful in all other simulation settings explored. For these reasons, the authors recommend implementing RobKAT with a Huber loss function in data analyses where the error distribution is unknown.

We also apply our methodology to the CATIE data set to test for associations between levels of neurotropic herpesviruses and genes located in the MHC region of chromosome six. Although the antibody level response was non-normal, our robust kernel association test was able to detect significant associations between HSV2 and the MHC region and the *HIST1H2BJ*, *POM121L2*, and *SLC17A1* genes. Among these genes, SKAT only detected association between HSV2 levels and the MHC region, which is unsurprising given the observed trend from simulation that SKAT is less powerful than our robust test when the response is non-normal.

Simulation studies for linear and quadratic kernels were also performed and produced similar power and type I error trends among the various loss functions. Appendix Tables A.1-A.2 show the type I error estimates for these additional kernels for sample sizes 100 and 300, respectively. For both the linear and nonlinear forms of $h(\boldsymbol{Z}_i)$, the tests utilizing different kernels had similar power with the quadratic kernel producing the greatest power. Appendix Figures A.1-A.2 and Figures A.3-A.4 show the power estimates for the linear and quadratic kernels with sample sizes 100 and 300, respectively.

Simulation studies for two additional ψ-functions, Hampel and Bisquare, were also performed. The Hampel family of functions is defined by

$$\psi(x) = \begin{cases} x & |x| \le a \\ a \text{ sign}(x) & a < |x| \le b \\ a \text{ sign}(x)\frac{r-|x|}{r-b} & b < |x| \le r \\ 0 & r < |x| \end{cases}$$

and setting $a = 1.353, b = 3.157, r = 7.216$ in our simulations for 95% efficiency (Hampel, 1986). Tukey's Bisquare family of functions is defined by $\psi(x) = x\left(1-(\frac{x}{k})^2\right)^2 I_{\{|x|\le k\}}$ and setting $k = 4.685$ for 95% efficiency in our simulations (Maronna et al., 2006). These two ψ-functions are called redescending functions, meaning that there exists a rejection point $c < \infty$ such that $\psi(x) = 0$ for all $|x| \ge c$ (Huber, 1981).

Although these redescending functions violate our assumptions on ψ, redescending functions may be attractive in data sets with extreme outliers in the distribution's tails, as the outliers would likely receive no weight in model fitting. Some researchers such as Huber (1981) warn against the use of these redescending functions, arguing that the risks from sensitivity to incorrect ψ scale parameter specification do not outweigh the slight improvement in asymptotic variance. For these reasons, simulation results that include the Hampel and Bisquare ψ-functions are only included in the Appendix. Additionally, a CATIE antibody data analysis including these redescending loss functions is included in Appendix Tables A.3-A.4

Appendix Tables A.1-A.2 include the type I error estimates for these additional Hampel and Bisquare ψ-functions for sample sizes 100 and 300, respectively. For the Cauchy and 30% mixture distributions, the redescending ψ-functions had slightly more power than with the other ψ-functions, likely due to the heavy tails and outliers in these distributions. For all other distributions, the power and type I error control were similar to that of the Huber ψ-function. With increasing effect sizes in our simulation study, we observed an eventual decreasing power trend for the redescending ψ-functions. With large effect size, we suspect that the plato in ρ-function inherent from the redescending ψ fails to detect strong signals.

Our algorithm is also computationally efficient through the utilization of fast kernel machine testing methods developed by Zhan and Wu (2017). Thus, not only is the RobKAT test general and robust, but our robust kernel association test is also scalable to moderate and large amounts of SNPs. For example, the average time to perform one RobKAT test with sample size 300, IBS kernel, and linear $h(\boldsymbol{Z}_i)$ is 0.17 seconds (2.9 GHz Intel Core i5 processor). Additionally, we have produced a user-friendly R package containing the RobKAT software. Thus, our robust kernel association test methodology is a robust generalization with fewer distributional and inferential restrictions than previously established methods, while remaining computationally efficient and user friendly.

Data Source

Data used in the preparation of this article were obtained from the limited access datasets distributed from the NIH-supported “Clinical Antipsychotic Trials of Intervention Effectiveness in Alzheimer’s Disease" (CATIE-AD). This is a multisite, clinical trial of persons with Alzheimer’s Disease comparing the effectiveness of randomly assigned medication treatment. The study was supported by NIMH Contract #N01MH90001 to the University of North Carolina at Chapel Hill. The ClinicalTrials.gov identifier is NCT00015548. This work was partially supported by National Institutes of Health grant P01CA142538 (to JYT) and National Institutes of Health training grant T32GM081057, Biostatistics Training in the Omics Era (to KM).

RobKAT software is available in the publicly available repository https://github.com/karamartinez00/RobKAT/tree/master/Downloads/RobKAT-master-3

CHAPTER

4

MULTIVARIATE ROBUST KERNEL ASSOCIATION TESTING WITH CONTINUOUS, BINARY, AND MIXED RESPONSES

4.1 Introduction

Testing for associations between genetic marker sets and a response is an important task in statistical genetics literature, and kernel machine regression (KMR) methods offer a useful set of solutions. Rather than performing separate hypothesis tests on individual markers, these KMR methods collectively test the association of the marker set on the response.

In the univariate continuous response case, many kernel association tests such as SKAT have been explored and popularized (Kwee et al., 2008; Liu et al., 2007; Wu et al., 2011). Extensions to discrete responses (Liu et al., 2008; Tzeng and Zhang, 2007) and domain-specific applications such as microbiome data (Zhan et al., 2017a; Zhao et al., 2015b) and copy number variants (CNVs) (Tzeng et al., 2015; Zhang, 2016) have also been explored. In recent literature, researchers have developed generalizations that replace square loss functions, which are tailored to normally distributed errors, with functions that may be more appropriate for non-normal distributions (Kong et al., 2016; Martinez et al., 2020a). In particular, we created in Chapter 3 a robust kernel association test, RobKAT, that removes normality assumptions for continuous data by employing a general loss function, thus accommodating responses with outliers, skew, and heavy-tails (Martinez et al., 2020a).

Although RobKAT provides distributionally robust KATs for univariate responses, data analyses with multivariate responses represent another application area for KAT research. Jointly analyzing multiple responses allows researchers to incorporate phenotype correlations and model pleiotropic relationships, leading to potentially more powerful tests (Zhu and Zhang, 2009). Maity et al. (2012) set a framework for a multivariate kernel machine regression (MKMR) tailored to continuous, marginally normal, multivariate responses. The MKMR methodology represents a multivariate generalization of univariate, least-squares-based KATs (Liu et al., 2007). Domain-specific univariate KATs, such as MiRKAT and SKAT, have multivariate extensions that are similar to MKMR in that they are tailored to normal conditional responses through a square loss function (Wu and Pankow, 2016; Zhan et al., 2017a). Davenport et al. (2017) developed a multivariate kernel association test for multivariate binary responses. For count and mixed data, other multivariate KATs have also been created (Zhan et al., 2017b).

Many of these multivariate kernel association testing methods, however, incorporate ordinary least squares parameter estimates and residuals based on a Gaussian linear model. This approach may be problematic when multivariate responses are marginally non-normal or the independent variables contain influential observations. As discussed in further detail in Section 4.4, the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) presents one such problematic data set. The CATIE clinical trial collected information on three neurotropic herpesvirus antibody levels as well as genotyped SNPs from the major histocompatibility complex region on Chromosome 6. As shown in Appendix Figure A.5, all three marginal herpesvirus antibody levels are notably skewed right and the log-transformed antibody levels are visibly bimodal. With such a deviation from multivariate normality, KATs that incorporate ordinary least squares estimates and their resulting residuals may not be ideal.

The dual kernel association test (DKAT), developed by Zhan et al (2017), is one such multivariate KAT whose derivation is not based on distributional assumptions. Rather, the DKAT statistic utilizes kernels to describe the response and genetic marker structures then compares the similarity between these two kernel structures (Zhan et al., 2017b). This comparison is quantified using a standardized Frobenius inner product, a quantification of similarity between two matrices (Shawe-Taylor and Cristianini, 2004). Since it uses an algorithm that approximates the exact distribution through moment matching, DKAT is also seen as more computationally efficient than MKMR (Wu and Pankow, 2016; Zhan et al., 2017b). With this distribution-free derivation, DKAT can accommodate any type of multivariate response if a proper response kernel is constructed. Although DKAT does provide an avenue for analyzing continuous, binary, count, and mixed multivariate responses, guidance to appropriate response kernels may not be straightforward. In the case of response kernels that incorporate covariate information, a common approach involves residuals based on ordinary least squares estimates that are tailored to Gaussian responses. In this aspect, the area of appropriate kernels for multivariate responses containing outliers, skew, or resulting from high leverage points remains largely unexplored.

In this paper, we aim to create a multivariate robust kernel association test (MrKAT) that can accommodate multiple response types and guide the user to appropriate robust phenotype kernels based on covariate information. To accomplish this goal, we propose a sum of univariate statistics based on the robust method, RobKAT, to serve as a multivariate extension. As the RobKAT loss functions are general, we also derive and implement loss functions for continuous, binary, and count response types that are robust to non-normal errors in the continuous case and robust to influential observations in the discrete case.

The rest of this article is organized as follows. In Section 4.2, we justify the summation method for MrKAT and derive appropriate robust loss functions in the continuous, binary, and mixed response cases. In Section 4.3, we then investigate through simulation study the type I error rate and testing power of MrKAT compared to DKAT. Our studies indicate that both methods adequately control type I error when the response is normal, non-normal, and contains outliers. Simulation results also suggest that MrKAT has similar power to DKAT when errors are normal but shows moderate to substantial power gains when errors are non-normal or influential observations exist in the covariate information. Through a case study in Section 4.4, MrKAT and DKAT are used to detect significant gene regions on Chromosome 6 that are associated with neurotropic herpesvirus antibody levels, which have notable skew in the continuous case. In our analysis, MrKAT was able to detect five additional gene regions when compared to testing results from DKAT. In Section 4.5, we conclude that MrKAT's software implementation is computationally efficient and appropriate for use on the genome-wide scale.

4.2 Methods

4.2.1 Univariate Continuous Response

In this section we develop a robust and general multivariate KMR hypothesis testing statistic that accommodates continuous, discrete, and mixed responses. For clarity, we first develop our methodology with a univariate, continuous response. In this case, a common model is $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + h(\boldsymbol{Z}_i) + \epsilon_i$ where $h(\cdot)$ is a potentially nonparametric function belonging to the function space, $\mathscr{H}_K$. The notation $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ represents an individual's response, covariates, and genetic markers. Testing for an overall genetic marker effect in this model translates to rejecting or failing to reject the hypothesis $H_0 : h(\cdot) = 0$.

Rather than specifying the complex, nonparametric form of $h(\cdot)$, the researcher works with the function space $\mathscr{H}_K$ generated by a kernel function, $k(\cdot,\cdot)$. In our methodology, $k(\cdot,\cdot)$ may be any function that is symmetric, continuous, and well-behaved according to Mercer's Theorem (Theodoridis and Koutroumbas, 2009). Through Mercer's Theorem, the kernel function defines an inner product of the functions' images in $\mathscr{H}_K$ (ie: $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = h(\boldsymbol{Z}_i) \cdot h(\boldsymbol{Z}_j)$) and may be thought of as a similarity measure of two individuals' genetic marker sets (Theodoridis and Koutroumbas, 2009). This connection between $h(\cdot)$ and $k(\cdot,\cdot)$ suggests the choice of kernel based on the believed form of $h(\cdot)$. For example, the linear kernel $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = a\boldsymbol{Z}_i^T \boldsymbol{Z}_j + c$ corresponds to an $\mathscr{H}_K$ of linear functions and defines $h(\cdot)$'s functional form as linear. The quadratic kernel $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = (1 + \boldsymbol{Z}_i^T \boldsymbol{Z}_j)^2$ implies an $\mathscr{H}_K$ of quadratic functions. Another popular choice is the identity by state (IBS) kernel, $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = \frac{1}{2p} \sum_{k=1}^{p} [2\mathbb{I}\{Z_{ik} = Z_{jk}\} - \mathbb{I}\{|Z_{ik} - Z_{jk}| = 1\}]$, which views the kernel as a similarity measure and sets the kernel output to the proportion of alleles shared between two individuals' sets of genetic markers. The matrix $\mathbb{K}$ with $(i,j)^{th}$ entries defined by $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$ is then symmetric positive semi-definite.

From the view of estimating $h(\cdot)$, the next step is to solve the penalized minimization criterion

$$\min_{h \in \mathcal{H}_K} \left[\sum_{i=1}^{N} \rho(Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta} - h(\boldsymbol{Z}_i)) + \lambda \, ||h||^2_{\mathcal{H}_K} \right] \tag{4.1}$$

and utilize the Representer Theorem to input $h(t) = \sum_{i=1}^{N} \alpha_i k(t, \boldsymbol{Z}_i)$ into Equation 4.1 where α_i are weights (Kimeldorf and Wahba, 1971). The more complicated minimization task then simplifies to solving a linear mixed model estimation problem as if $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{h}_i + \epsilon_i$ with $\boldsymbol{h} \sim N(0, \tau \mathbb{K})$ and, often, $\epsilon \sim N(0, \boldsymbol{\Sigma})$. Thus, KMR hypothesis testing of $H_0 : h(\cdot) = 0$ is often carried out through a variance component score test or some variation on such a test (Kong et al., 2016; Liu et al., 2008,0; Wu et al., 2011).

Such a score test only requires parameter estimates under the null hypothesis, which simplifies KMR statistic computations. Since the nonparametric component vanishes under the null assumption, only parameter estimates from a linear model are required when computing the test statistic. In some cases, the normality assumptions placed on ϵ_i lead to $\rho(t) = \frac{1}{2}t^2$ in Equation 4.1. While this simplifies the null distribution of test statistics, many such KMR methods may lose power when distributional assumptions are violated (Martinez et al., 2020a).

The robust kernel association test (RobKAT) in Chapter 3 is a robust generalization of these univariate KMR hypothesis tests for continuous responses, and this extension is facilitated through the loss function $\rho(\cdot)$ in Equation 4.1. RobKAT does not directly employ distributional assumptions, but the method instead uses a small number of loss function requirements from robust statistics literature and to exclude redescending loss functions (Martinez et al., 2020a)

1. $\rho : \mathbb{R} \to \mathbb{R}^+$ is a differentiable, positive-valued function
2. $\rho(0) = 0$
3. $\rho(x)$ is a convex function of x
4. $\psi(x) = \frac{d}{dx}\rho(x)$ exists and is non-decreasing

The RobKAT test statistic is $T = \boldsymbol{w}^T \mathbb{K} \boldsymbol{w}$ where $\boldsymbol{w}_i = \psi(\frac{\widehat{\epsilon}_i}{\widehat{s}})$, $\widehat{s}$ represents the Huber Proposal 2 robust estimate of scale, $\widehat{\epsilon}_i = Y_i - \widehat{Y}_i$ is the residual from the null model, and $\mathbb{K}$ is a user-specified kernel matrix. When $\rho(t) = \frac{1}{2}t^2$, we get the least square kernel machine statistic of Liu et al. (2007) that is sensitive to outliers but appropriate for normally distributed responses (Martinez et al., 2020a). This square loss is also found in SKAT methodology (Wu et al., 2011). When the loss is set to the least absolute deviation (LAD), $\rho(t) = \frac{1}{2}|t|$, we get the quantile regression kernel machine statistic of Kong et al that is less sensitive to outliers but limited to inference on the conditional quantile (Kong et al., 2016). As a robust intermediate between the these two loss functions, the Huber loss function is $\rho(t) = \frac{1}{2}t^2\ \mathbb{I}\{|t| \leq b\} + b(|t| - \frac{b}{2})\ \mathbb{I}\{|t| > b\}$ where b is a scalar chosen for desired asymptotic properties. The Huber function behaves like a square loss near zero and like LAD for more extreme values.

4.2.2 Multivariate Continuous Response

We now extend RobKAT to our proposed multivariate statistic. When each individual has $r \geq 1$ responses, we use the model

$$Y_{ij} = \boldsymbol{X}_i^T \boldsymbol{\beta}_j + h_j(\boldsymbol{Z}_i) + \epsilon_{ij} \tag{4.2}$$

for $i = 1, 2, ..., n$ and $j = 1, 2, ..., r$. We will use the notation $\Psi_j = [\psi(\epsilon_{1j}/s_j), ..., \psi(\epsilon_{nj}/s_j)]^T$ in our derivation where $\psi(\cdot)$ signifies the Huber loss function, which will then weight the fitted residuals robustly for response j. The scale parameters and error terms are represented in this model by s_j and ϵ_{ij}. As in Section 4.2.1, we will estimate these terms with robust regression residuals for ϵ_{ij} and Huber's Proposal 2 robust estimate of scale for s_j. Substituting these estimates into Ψ_j to create $\widehat{\Psi}_j$, the RobKAT test statistic for the j^{th} individual response is then $T_j = \widehat{\Psi}_j^T \mathbb{K} \widehat{\Psi}_j$. We propose a cumulative sum of these r RobKAT statistics for our robust multivariate test:

$$\begin{aligned}
\sum_{j=1}^{r} T_j &= \sum_{j=1}^{r} tr\left(\widehat{\Psi}_j^T \mathbb{K} \widehat{\Psi}_j\right) \\
&= tr\left(\mathbb{K} \sum_{j=1}^{r} \widehat{\Psi}_j \widehat{\Psi}_j^T\right) \\
&= tr(\mathbb{K} \Psi \Psi^T)
\end{aligned}$$

where the j^{th} column of Ψ is defined by $\widehat{\Psi}_j$.

Computing a p-value from our test statistic then requires either permutation procedures or an approximation method. In recent literature, permutation-based inference has been successful, though computationally burdensome in cases of large sample sizes (Kong et al., 2016; Maity et al., 2012). Our methodology and software leverages the form of our test statistic and implements the null-distribution approximation algorithm developed by Zhan and Wu (2017). The algorithm utilizes literature on fast permutation-approximation techniques for RV-type statistics of the form $tr(\mathbb{P}\boldsymbol{A}\boldsymbol{A}^T\mathbb{P}\boldsymbol{B}\boldsymbol{B}^T)$ where $\mathbb{P}$ is the centering matrix $\mathbb{P} = \mathbb{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ (Josse et al., 2008). Noting that the positive semi-definite matrix $\mathbb{K}$ may be written as $\mathbb{K} = \mathbb{K}^{\frac{1}{2}}(\mathbb{K}^{\frac{1}{2}})^T$, our statistic after centering $\mathbb{K}$ is $T = tr(\mathbb{P}\mathbb{K}^{\frac{1}{2}}(\mathbb{K}^{\frac{1}{2}})^T\mathbb{P}\Psi\Psi^T)$, which follows the RV-statistic form. Using the first three analytical moments of $tr(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{Y}^T\boldsymbol{Y})$ derived by Kazi-Aoual et al. (1995), the algorithm then approximates the null distribution by matching analytical moments to a Pearson III distribution (Zhan and Wu, 2017). This fast approximation approach is included in other methods such as DKAT, and has been shown to be about 6,000 times faster than permutation test procedures when applied to a univariate response (Zhan and Wu, 2017).

In addition to sharing the same null distribution approximation methods, there is a connection between the DKAT and MrKAT test statistics. The DKAT statistic is derived through a standardized Frobenius inner product of phenotype and genotype kernels, which measures the similarity of the two structures (Shawe-Taylor and Cristianini, 2004; Zhan et al., 2017b). The resulting DKAT statistic derived is

$$T_{DKAT} = \frac{tr(\mathbb{P}\boldsymbol{K}_G\mathbb{P}\boldsymbol{K}_Y)}{\sqrt{tr(\mathbb{P}\boldsymbol{K}_G\mathbb{P}\boldsymbol{K}_G)tr(\mathbb{P}\boldsymbol{K}_Y\mathbb{P}\boldsymbol{K}_Y)}} \tag{4.3}$$

where $\boldsymbol{K}_Y$ and $\boldsymbol{K}_G$ are kernel matrices describing similarity among the phenotypes and genetic markers, respectively (Zhan et al., 2017b). In the case of normally distributed data, a typical phenotype kernel is $\boldsymbol{K}_Y = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^T$ where $\widehat{\boldsymbol{Y}}$ is an estimate from a multivariate linear model. In this case, DKAT follows the form of MrKAT where $\Psi_j = \hat{\boldsymbol{\epsilon}}_j$. The ψ used here is derived from a square loss $\rho(t) = \frac{1}{2}t^2$ as is appropriate when $\boldsymbol{Y}$ is multivariate normal and contains no outliers. This connection to DKAT suggests that, intuitively, the MrKAT method is also comparing the genotype and phenotype similarity structures. Based on this connection to DKAT via square loss, MrKAT may be viewed as proposing a specialized phenotype kernel to down-weight the influence of outliers in a hypothesis test. However, in the case of MrKAT, this phenotype kernel is derived from univariate, robust score statistics.

4.2.3 Multivariate Binary Responses

We now extend our MrKAT methodology to accommodate multivariate binary responses. KMR hypothesis testing for a discrete univariate response involves a modified model $g(E[\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{Z}_i]) = \boldsymbol{X}_i^T \boldsymbol{\beta} + h(\boldsymbol{Z}_i)$ where $g(\cdot)$ is the link function from generalized linear model methodology. For binary responses, g is the logit link function. To fit into the continuous methodology already developed, we must derive an appropriate, robust $\psi(\cdot)$ function for a binary response. In Section 4.2.2 we looked to the Huber loss function from robust regression methods, so we look to robust logistic regression for this binary case.

The traditional logistic regression log-likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[Y_i log(p_i(\boldsymbol{\beta})) + (1 - Y_i) log(1 - p_i(\boldsymbol{\beta})) \right]$$

with $p_i(\boldsymbol{\beta}, \gamma)$ defined by the logistic distribution $p_i(\boldsymbol{\beta}) = \frac{exp\{\boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \gamma\}}{1 + exp\{\boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \gamma\}}$ (Agresti, 2013; Maronna et al., 2006). Creating a more robust logistic regression procedure, Carroll and Pederson (1993) modify this log-likelihood by weighting observations according to their leverage:

$$l(\boldsymbol{\beta}, \gamma) = \sum_{i=1}^{n} w_i \left[Y_i log(p_i(\boldsymbol{\beta})) + (1 - Y_i) log(1 - p_i(\boldsymbol{\beta})) \right]$$

As utilized by Carroll and Pederson (1993), we use the weight, $w_i = W(h_n(\boldsymbol{X}_i))$, where $W(u)$ is the trisquared redescending function from $W(u) = (1 - \frac{u^2}{c^2})^3 I(|u| \leq c)$ so that observations with large leverage are downweighted. This $W(u)$ weight includes a prespecified tuning constant, $c > 0$, and the value, $c = 8$, is the default implementation in our software as suggested in an example by Carroll and Pederson (1993). The quantity $h_n(\boldsymbol{X}_i) = (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_n)^T \widehat{\boldsymbol{\Sigma}}_n^{-1} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_n)$ is the leverage for an observation, $\boldsymbol{x} =< x_1, x_2, ..., x_n >$ with estimates of location and scale $\widehat{\boldsymbol{\mu}}_n$ and $\widehat{\boldsymbol{\Sigma}}_n$ (Maronna et al., 2006). We then use this weighted log-likelihood to derive the appropriate loss function, $\psi_j(\cdot)$, for the binary case:

$$\begin{aligned}
\frac{dl}{d\gamma} &= \sum_{i=1}^{n} w_i (Y_i - p_i(\boldsymbol{\beta}, \gamma)) \boldsymbol{Z}_i \\
&= \sum_{i=1}^{n} \psi(\epsilon_i) \boldsymbol{Z}_i
\end{aligned}$$

Thus, using robust logistic regression guides us to set $\psi(\epsilon_{ij}) = w_i \epsilon_{ij}$. With this thinking, the i^{th} element of $\widehat{\Psi}_j$ is $w_i \widehat{\epsilon}_{ij}$, where w_i down-weights observations with high leverage and $\widehat{\epsilon}_{ij}$ are the robust logistic regression residuals. We then propose to substitute the robust binary $\widehat{\Psi}_j$ into our $T_j = \widehat{\Psi}_j^T \mathbb{K} \widehat{\Psi}_j$ statistic. This T_j then represents a robust kernel association test statistic for a univariate binary response.

4.2.4 Multivariate Mixed Responses

Using our statistics for continuous and binary responses, we now develop the KAT statistic that accommodates mixed responses. This extension to multivariate mixed responses comes from observing that each Ψ_j defines a column of Ψ. We then use the robust binary Ψ_j in the columns corresponding to a binary response, and we use the robust continuous Ψ_j in columns corresponding to continuous responses. We then proceed with the mixed-phenotype Ψ to obtain a new version of $T = tr(\mathbb{K}\Psi\Psi^T)$, and we utilize the same fast null-distribution approximation algorithm from Zhan et al (2017) that was implemented in RobKAT and DKAT.

4.3 Simulation Study

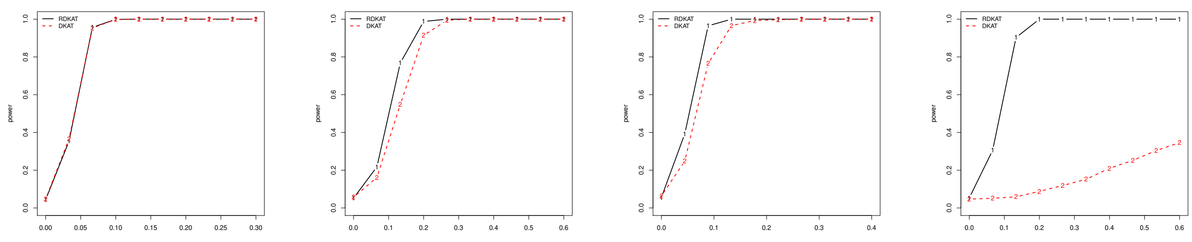
We perform simulation studies to compare the testing performance of MrKAT and existing methods such as DKAT. In the first simulation, we evaluate the power and type I error when the multivariate responses, $\boldsymbol{Y}_i \in \mathbb{R}^4$, are all continuous but have a variety of error distributions. In the second simulation, we explore the effect of adding a high-leverage observation when some or all marginal responses in $\boldsymbol{Y}_i \in \mathbb{R}^4$ are binary.

4.3.1 Non-normal Responses

In the first simulation, we investigate the effects of continuous multivariate responses generated with errors that are not normally distributed. We generate $i = 1, 2, ..., n = 100$ observations with continuous responses $j = 1, 2, 3, 4$ from the model

$$Y_{ij} = \boldsymbol{X}_i^T \boldsymbol{\beta} + c * h_j(\boldsymbol{Z}_i) + \epsilon_{ij}$$

where $\boldsymbol{X}_i \sim N(0, \mathbb{I}_4)$ and $\boldsymbol{\beta} = -5 * \mathbf{1}_4$. We randomly sample 9-SNP, $\boldsymbol{Z}_i$ observations from the *SLC17A1* gene of the CATIE clinical trial data discussed in Section 4.4, and we borrow from Liu et al (2008) a linear model for the genetic marker effect $h_1(\boldsymbol{Z}_i) = h_2(\boldsymbol{Z}_i) = h_3(\boldsymbol{Z}_i) = h_4(\boldsymbol{Z}_i) = 2Z_{i1} + 3Z_{i2} + Z_{i3} + 2Z_{i4} + Z_{i5}$. We implemented the MrKAT statistic by specifying $\Psi_j(\cdot)$ as residuals from the Huber loss function with $b = 1.345$ (for 95% efficiency to the normal). Both MrKAT and DKAT modeled genotype similarity, $\mathbb{K}$ or $\boldsymbol{K}_G$, with the IBS kernel described in Section 4.2.



(a) $N(0,1)$ Errors **(b)** χ_1^2 Errors **(c)** t_3 Errors **(d)** Cauchy Errors

Figure 4.1 Power simulation results for four continuous, independent, multivariate responses that are symmetric $N(0,1)$, skewed (χ_1^2), heavy-tailed (t_3), and substantially heavy-tailed (Cauchy). MrKAT is shown in black, and DKAT is shown in red. The simulation implements IBS kernel, n=100, $h_1(\boldsymbol{Z}_i) = h_2(\boldsymbol{Z}_i) = h_3(\boldsymbol{Z}_i) = h_4(\boldsymbol{Z}_i) = 2Z_{i1} + 3Z_{i2} + Z_{i3} + 2Z_{i4} + Z_{i5}$, and $\boldsymbol{\beta} = -5\mathbf{1}_4$.

In this simulation study, we choose to alter the distribution of ϵ_{ij} to create marginal responses that are symmetrically distributed ($\epsilon_{ij} \underset{iid}{\sim} N(0,1)$), skewed ($\epsilon_{ij} \underset{iid}{\sim} \chi_1^2$), heavy-tailed ($\epsilon_{ij} \underset{iid}{\sim} t_3$), and substantially heavy-tailed ($\epsilon_{ij} \underset{iid}{\sim} Cauchy(0,1)$). Although the error terms are independent, the elements of Y_{ij} are not independent for $j = 1, ..., r$ as they contain the same values of $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$, and $h_j(\boldsymbol{Z}_i)$. For columns of Ψ corresponding to continuous responses, the elements of Ψ_j are computed using robust linear regression residuals with Huber loss ($b = 1.345$), as described in Section 4.2.2.

To evaluate type I error, we set $c = 0$ to create a null model and compared Monte Carlo estimates from $N = 10,000$ iterations at the $\alpha = 0.01$, $\alpha = 0.05$, and $\alpha = 0.10$ levels of significance. As shown in Table 4.1, type I error was adequately controlled for both methods across all error settings.

To compare testing power of MrKAT and DKAT, we modeled increasing values of c and computed the proportion of rejected values at the $\alpha = 0.05$ significance level. We generated $N = 1,000$ iterations for each of ten equally-spaced values of c between 0 and 0.8. The resulting power curves are shown in

Table 4.1 Simulation study for type I error of MrKAT and DKAT the non-normal setting. The simulation includes four continuous responses, various error distributions, IBS genetic kernel, and a sample size of 100. Estimates show the proportion of 10,000 iterations with p-value below the α cutoff.

Error	α	MrKAT	DKAT
	0.01	0.0108	0.0101
$N(0,1)$	0.05	0.0479	0.0490
	0.10	0.1023	0.1035
	0.01	0.0096	0.0107
χ_1^2	0.05	0.0494	0.0491
	0.10	0.0972	0.0999
	0.01	0.0096	0.0088
t_3	0.05	0.0472	0.0503
	0.10	0.1005	0.0988
	0.01	0.0110	0.0106
Cauchy	0.05	0.0480	0.0490
	0.10	0.0995	0.0975

Figure 4.1. When the distribution is symmetric, we see in Figure 4.1a that MrKAT and DKAT produce nearly identical power. For the skewed and moderately heavy-tailed marginal responses in Figure 4.1b-4.1c, the robust MrKAT has a moderate power advantage over DKAT. When the errors are Cauchy, there is a substantial power decrease with DKAT, while MrKAT produces a powerful hypothesis test. These results suggest that MrKAT is more robust to continuous, non-normal multivariate responses methods based on square loss functions, such as this instance of DKAT.

4.3.2 High Leverage Observations

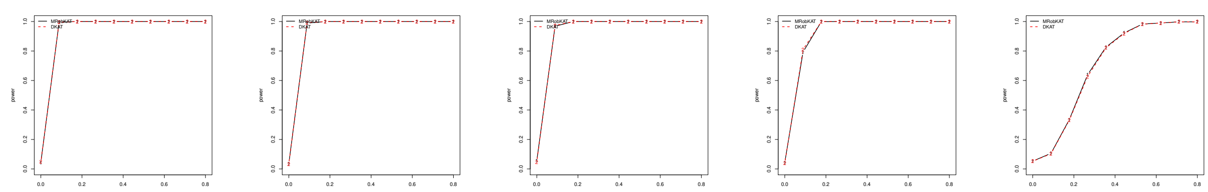
In our second simulation, we evaluate the power and type I error when the multivariate responses, $\boldsymbol{Y}_i \in \mathbb{R}^4$, contain any number of binary elements. We generate $n = 100$ observations under two settings: a standard setting where all data points follow the given model, and a robust setting where one data point is influential and deviates from the underlying model's pattern.

For the standard setting, we generate continuous responses using the same process described in Section 4.3.1. For the observations Y_{ij} corresponding to binary responses, we again simulate $\boldsymbol{X}_i$, $\boldsymbol{\beta}$, c, and $h_j(\boldsymbol{Z}_i)$ as detailed in Section 4.3.1; however, response data are generated with the logistic model $logit(P(Y_{ij}=1|\boldsymbol{X}_i, \boldsymbol{Z_i})) = \boldsymbol{X}_i^T\boldsymbol{\beta} + c * h_j(\boldsymbol{Z}_i)$. For columns of Ψ corresponding to continuous responses, the elements of Ψ_j are computed using robust linear regression residuals with Huber loss ($b = 1.345$), as described in Section 4.2.2. For columns of Ψ corresponding to binary responses, the elements of Ψ_j are computed by $w_i(Y_{ij} - \widehat{p}_{ij})$, where w_i is the Mallow's leverage-dependent weight value and $\widehat{p}_i$ is the fitted value using Mallow's estimates, as described in Section 4.2.3 (Carroll and Pederson, 1993). We implement DKAT with $\boldsymbol{K}_Y = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^T$ so that Ψ_j contains the residuals from a logistic regression estimates for j corresponding to binary responses and residuals from Gaussian linear models for j corresponding to continuous responses. As with the first simulation, we note that all elements of Y_{ij} are not independent as $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$, and $h_j(\boldsymbol{Z}_i)$ are shared among $j = 1, 2, 3, 4$ for either the standard or robust settings.

For the robust setting, we generate the data as described in the previous paragraph, but we also alter the first observation to be an outlier that has high influence. To create this influential observation, we generate a large value of $X_1 \underset{iid}{\sim} N(3 * \mathbf{1}_4, \mathbb{I}_4)$ and a large response value ($Y_{1j} \sim N(10, 1)$ for continuous responses and $Y_{1j} = 1$ for binary responses). The trend of this point disagrees with the modeled trend of $\boldsymbol{\beta} = -5 * \mathbf{1}_4$ and may affect slope estimates when non-robust methods are used. In a practical sense, this point may mirror transcribing the incorrect response sign or a single patient who does not fit the general trend. As computed in Appendix Table B.2, the leverage associated with this covariate matrix is over 7 times what would be expected if the observations had equal leverage ($q/n = 4/100$). Measuring the effect on least squares regression of this observation, Appendix Table B.2 details how the first observation has Cook's Distance far higher than the non-outlier setting; however, we also see that the outlier has more influence in the continuous response setting, likely due to the capped magnitude of a binary response.

Table 4.2 Simulation study for the type I error rate of MrKAT and DKAT in the Standard (no high leverage points) and Robust (1 high leverage point) settings. The simulation includes IBS genetic kernel and sample size of 100. Estimates show the proportion of 10,000 iterations with p-value below the α cutoff.

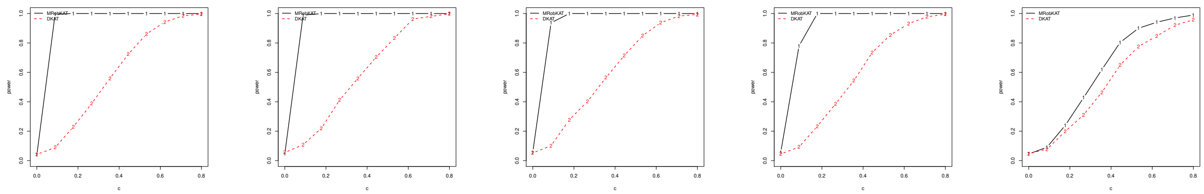
		MrKAT		DKAT	
Error	α	Standard	Robust	Standard	Robust
0 Binary	0.01	0.0108	0.0102	0.0101	0.0094
	0.05	0.0479	0.0527	0.0490	0.0465
	0.10	0.1023	0.1026	0.1035	0.0959
1 Binary	0.01	0.0082	0.0083	0.0089	0.0114
	0.05	0.0448	0.0481	0.0442	0.0503
	0.10	0.0940	0.0965	0.0892	0.0948
2 Binary	0.01	0.0096	0.0106	0.0101	0.0105
	0.05	0.0470	0.0513	0.0489	0.0489
	0.10	0.0961	0.0971	0.0945	0.0995
3 Binary	0.01	0.0096	0.0097	0.0097	0.0104
	0.05	0.0469	0.0502	0.0482	0.0507
	0.10	0.0941	0.0975	0.0987	0.0979
4 Binary	0.01	0.0101	0.0116	0.0093	0.0099
	0.05	0.0522	0.0518	0.0483	0.0491
	0.10	0.1044	0.0997	0.1014	0.0989



(a) 0 Binary Resp. **(b)** 1 Binary Resp. **(c)** 2 Binary Resp. **(d)** 3 Binary Resp. **(e)** 4 (all) Binary Resp.

Figure 4.2 Power simulation results from the standard setting (no high leverage points). MrKAT is shown in black, and DKAT is shown in red. This simulation included 4 Responses, and IBS kernel, n=100, $h_1(\boldsymbol{Z}_i) = h_2(\boldsymbol{Z}_i) = h_3(\boldsymbol{Z}_i) = h_4(\boldsymbol{Z}_i) = 2Z_{i1} + 3Z_{i2} + Z_{i3} + 2Z_{i4} + Z_{i5}$, and $\boldsymbol{\beta} = -5\mathbf{1}_4$

As shown in Table 4.2, type I error is adequately controlled for both methods across both the standard and robust settings. Summarized in Figure 4.2, there is almost no power difference between MrKAT and DKAT in the standard setting, though this result is to be expected. For observations that do not contain outliers or high leverage points, both testing methods use a similar, quadratic loss function when computing their test statistics. Thus, both methods handle continuous, mixed, and binary data equally well when there are neither high leverage points nor response outliers.



(a) 0 Binary Resp. **(b)** 1 Binary Resp. **(c)** 2 Binary Resp. **(d)** 3 Binary Resp. **(e)** 4 (all) Binary Resp.

Figure 4.3 Power simulation results from the robust setting (1 high leverage points). MrKAT is shown in black, and DKAT is shown in red. This simulation included 4 Responses, and IBS genotype kernel, sample size $n = 100$, $h_1(\boldsymbol{Z}_i) = h_2(\boldsymbol{Z}_i) = h_3(\boldsymbol{Z}_i) = h_4(\boldsymbol{Z}_i) = 2Z_{i1} + 3Z_{i2} + Z_{i3} + 2Z_{i4} + Z_{i5}$, and $\boldsymbol{\beta} = -5\mathbf{1}_4$. For the influential observation, we set $Y_{100} \sim N(10, 1)$ or $Y_{100} = 1$, $X_{100} \sim N(3, 1)$.

The results from the robust setting in Figure 4.3, however, suggest that DKAT may be much more sensitive to high leverage points than MrKAT. Even when 1% of the observations deviate from the overall trend, there is potential for a substantial decrease in DKAT testing power. These results indicate that, in addition to providing a powerful test for binary and mixed data, MrKAT is far more robust to high leverage points than the method based on ordinary least squares residuals. We also see that this advantage over DKAT decreases as the number of binary responses increases, as the added data point has less influence in the logistic model. As shown in Appendix Figures B.1-B.2 these trends also hold when the genetic markers are modeled with a nonlinear function.

4.4 Application to CATIE Herpesvirus Data

Various studies and clinical trials have linked the Major Histocompatibility Complex (MHC) region on chromosome 6 to symptoms of schizophrenia (Aberg et al., 2013; International Schizophrenia Consortium et al., 2009; Shi et al., 2009; Stefansson et al., 2009). However, the highly polymorphic MHC region is best known for its involvement with immune response to infectious agents (Janeway et al., 2001). Neurotropic herpesviruses are a type of virus thought to affect the central nervous system. These herpesvirus infections have been linked to cognitive impairments, and in some cases this decreased cognition has been similar to those of patients suffering from schizophrenia (Yolken et al., 2011).

The Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) provides both herpesvirus antibody and MHC genetic information. The primary goal of the CATIE study was to compare the efficacy of atypical antipsychotics in schizophrenia patients, and 51% of the 1,460 enrolled subjects also submitted DNA samples (Lieberman et al., 2005; Sullivan et al., 2008). From these genetic samples, 429,000 SNPs were available for analysis (Sullivan et al., 2008). CATIE patient blood samples were also obtained at baseline. Focusing on viruses capable of infecting the human central nervous system, researchers analyzed exposure to three viruses: Herpes Simplex Virus type 1 (HSV-1), Herpes Simplex Virus type 2 (HSV-2), and Cytomegalovirus (CMV). Exposure to these three herpesviruses were then quantified by measuring IgG class antibodies in patient blood samples, providing continuous measurements (Yolken et al., 2011). A binary representation of the IgG class antibody levels was also provided, indicating whether denoted values were greater than a pre-determined cutoff or not. As shown in Figure A.5, the continuous HSV-1, HSV-2, and CMV antibody level values are notably skewed. After a log-transformation, the distributions are visibly bimodal, which again fails typical normality assumptions.

Due to the connection between neurotropic herpesviruses and schizophrenia, our goal is to investigate the association of genes within the MHC region and IgG levels for the three neurotropic herpesviruses measured. We analyze each of 11 genes in the MHC region as well as the overall MHC region, which contains a total of 787 SNPs.

For the 3 continuous antibody responses, we employed both DKAT and MrKAT in multivariate analyses. For the DKAT method, we utilized the phenotype kernel, $\boldsymbol{K}_Y = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^T$, where each column of $\widehat{\boldsymbol{Y}}$ was modeling with a Gaussian linear model. The MrKAT method was implemented as described in Section 4.2.2 with a Huber loss function. For both association testing methods, we repeated analysis with linear, quadratic, and IBS genotype kernels. To control for multiple comparisons, we employed a Bonferroni correction of $0.05/(2*3*12) = 0.00069$ for tests on 2 testing methods, 3 genotype kernels, 12 marker sets.

For the continuous antibody responses, our MrKAT analysis detected five significant gene regions (*BTN2A1*, *BTN2A2*, MHC, *POM121L2*, *SLC17A1*). Associations to some of these gene regions (MHC, *POM121L2*, *SLC17A1*) align with previous analyses in Chapter 3, and links between *POM121L2* and schizophrenia have been suggested in the literature (Aberg et al., 2013). The DKAT method did not detect any significant associations, even when testing the overall MHC region with strongest links to immune response. This difference in gene detection may be due to the DKAT phenotype kernel utilizing a loss function derived from normal responses. MrKAT, however, employs a robust

Table 4.3 Applying MrKAT and DKAT to the continuous form of HSV-1, HSV-2, and CMV antibody levels. The bold values highlight significant p-values after a Bonferroni correction $(0.05/(2*3*12) = 0.00069)$ for the 2 methods, 3 kernels, and 12 genes.

Gene	Test	Linear	Quadratic	IBS
BTN2A1	DKAT	0.00164	0.00987	0.00368
	MrKAT	**0.00003**	**0.00036**	**0.00019**
BTN2A2	DKAT	0.00423	0.01211	0.00459
	MrKAT	**0.00028**	0.00137	**0.00020**
BTN3A2	DKAT	0.01089	0.00731	0.00703
	MrKAT	0.00959	0.00474	0.01281
HIST1H2AG	DKAT	0.65120	0.80305	0.49072
	MrKAT	0.22443	0.42351	0.12470
HIST1H2BJ	DKAT	0.19307	0.22607	0.17298
	MrKAT	0.00849	0.01795	0.00889
MHC	DKAT	0.00141	0.00462	0.00143
	MrKAT	**5.78e-07**	**1.21e-06**	**5.24e-07**
NOTCH4	DKAT	0.07155	0.08818	0.09974
	MrKAT	0.03893	0.12917	0.05763
POM121L2	DKAT	0.02974	0.06077	0.01821
	MrKAT	0.00072	0.00190	**0.00043**
PRSS16	DKAT	0.33324	0.33324	0.33324
	MrKAT	0.13580	0.13580	0.13693
SLC17A1	DKAT	0.00755	0.01607	0.01990
	MrKAT	**0.00003**	**0.00023**	**0.00018**
SLC17A3	DKAT	0.15663	0.12839	0.19999
	MrKAT	0.03558	0.05209	0.04951
ZNF184	DKAT	0.54356	0.49649	0.61742
	MrKAT	0.51123	0.51134	0.72252

loss function that down-weights extreme values to provide more powerful tests in the situation of skewed marginal distributions.

For demonstrative purposes, we then repeated our analysis with the three binary representations of the antibody levels and include these results in Appendix Table B.1. To accommodate the binary responses in DKAT, we instead computed each column of $\widehat{\boldsymbol{Y}}$ using logistic regression. For MrKAT, we implemented the weighted logistic regression loss function as described in Section 4.2.3. All other settings were identical to the continuous analysis. For the binary representation of the antibody response, both methods detect the similar significant genes (*BTN2A1*, *BTN2A2*, MHC, *SLC17A1*), which match the results from our MrKAT analysis on continuous responses. As we see in Appendix Table B.3, there is little evidence of extreme leverage points or highly influential observations among the clinical trial data, so the similar performances of the two methods may be expected. Many p-values from these binary analyses are much smaller than those from the corresponding continuous response analyses. Since the response is binary, nuance regarding association strength may be lost.

4.5 Discussion

Through our multivariate extension of RobKAT, we have developed a multivariate kernel association test that can accommodate combinations of binary and continuous responses. In the absence of influential observations and non-normal conditional responses, type I error control and testing power are similar to previous methods such as DKAT. However, MrKAT is also robust to high leverage points and non-normal error distributions through the incorporation of robust regression and robust logistic regression methods when modeling the phenotype. We demonstrate through simulation study adequate type I error control and improved testing power under a variety of robust settings (Section 4.3). In particular, the case of responses containing Cauchy errors and high-leverage covariates show substantial power gains over DKAT when residual-dependent phenotype kernels are used.

As explored through our simulation study in Section 4.4, these power gains are useful in analyses where the marginal responses are non-normal, such as in the CATIE antibody level data. Shown in Table 4.3, the power of residual-dependent DKAT seemed insufficient to detect any significant gene regions on Chromosome 6, while MrKAT detected five significant gene regions in the continuous case.

Although the CATIE analysis in Section 4.4 only included three responses, it should be noted that MrKAT may accommodate larger dimensions without losing substantial power. Simulation evidence for up to 200 marginal responses are presented in Appendix Figure B.4 and Appendix Table B.4. The performance of MrKAT is also reasonably robust to sample sizes as small as 25 (Appendix Figure B.5 and Appendix Table B.5). Simulation results for large SNP-sets ($p = 787$) show similar trends in testing power (Appendix Figure B.6).

The MrKAT methodology and software may also be extended to general responses from the exponential family, by adapting the log-likelihood in the MrKAT derivations. Following the derivation in Section 4.2.3, an appropriate robust loss function ψ for exponential family traits may be derived similarly through the log-likelihood

$$l(Y_i;\theta_i)=\sum_{i=1}^{n}\Big[\frac{Y_i\theta_i-b(\theta_i)}{a(\phi)}+c(Y_i,\phi)\Big]$$

where ϕ is the dispersion parameter and θ_i is the natural parameter. As in the binary case, to make this ψ robust to high leverage points and outliers, the residuals would be estimated using a robust regression method and a weight would be added. When following the score derivation in Section 4.2.3 and using the GLM canonical link function $g(\mu_i)=\frac{\theta_i}{a(\phi)}=\boldsymbol{\beta}^T\boldsymbol{X}_i+\gamma^T\boldsymbol{Z}_i$, we then see:

$$\begin{aligned}\frac{dl}{d\gamma}&=\sum_{i=1}^{n}w_i\Big[Y_i\,\frac{d}{d\gamma}\Big(\frac{\theta_i}{a(\phi)}\Big)-\frac{\frac{d}{d\gamma}b(\theta_i)}{a(\phi)}\Big]\\&=\sum_{i=1}^{n}w_i\Big[Y_i-\frac{db}{d\theta_i}\Big]\boldsymbol{Z}_i\\&=\sum_{i=1}^{n}w_i\big[Y_i-E[Y_i]\big]\boldsymbol{Z}_i\end{aligned}$$

The result, $E[Y_i]=\mu_i=\frac{db}{d\theta_i}$, follows from a property of the exponential family (Agresti, 2013). In Section 4.2.3, we used residuals from robust logistic regression by employing the weights $w_i=W(h_n(\boldsymbol{X}_i))$ to produce Mallows estimates. These weight observations based on leverage only, since $Y_i\in\{0,1\}$ and leverage is the issue of interest in logistic regression. While one could also use w_i from Section 4.2.3 in the GLM setting, we suggest an alternative robust weighting scheme that down-weights according to atypical leverage and residual values. Employing these weights in the derivation, we are led to substituting the weighted residual as Ψ_j in MrKAT if the j^{th} response comes from the exponential family.

As implemented in our software and other software packages, an alternative robust weighting for a general GLM would come from the conditionally unbiased bounded influence (CUBIF) method (Künsch et al., 1989; Maronna et al., 2006). The CUBIF implementation utilizes the function

$$\psi(Y_i,\boldsymbol{X}_i,\boldsymbol{\beta},b,\boldsymbol{B})=W(Y_i,\boldsymbol{X}_i,\boldsymbol{\beta},b,\boldsymbol{B})\ \ r(Y_i,\boldsymbol{X}_i,\boldsymbol{\beta},b,\boldsymbol{B})$$

which is intuitively thought of as a weighted residual. The residual portion, $r(Y_i,\boldsymbol{X}_i,\boldsymbol{\beta},b,\boldsymbol{B})=Y_i-g(\boldsymbol{\beta}^T\boldsymbol{X}_i)-a(\boldsymbol{\beta}^T\boldsymbol{X}_i,\frac{b}{h(\boldsymbol{X}_i,\boldsymbol{B})})$, contains a bias correction term $a(\boldsymbol{\beta}^T\boldsymbol{X}_i,\frac{b}{h(\boldsymbol{X}_i,\boldsymbol{B})})$ to accommodate CUBIF's conditional Fisher-consistency requirement that $E[\Psi(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta},b,\boldsymbol{B})\boldsymbol{X}|\boldsymbol{X}_i] = \mathbf{0}$. The function, $h(\boldsymbol{X}_i,\boldsymbol{B})$, computes an observation's leverage, and the scalar, b, is the cutoff value in the Huber loss function. A common software default based on desired asymptotic efficiency is $b = 2$. The matrix, $\boldsymbol{B}$, is an estimated dispersion matrix whose derivation and properties are detailed by Künsch et al. (1989). The weighting portion of ψ is given by

$$\begin{aligned} W(Y_i,\boldsymbol{X}_i,\boldsymbol{\beta},b,\boldsymbol{B}) &= W_b\big(r(Y_i,\boldsymbol{X}_i,\boldsymbol{\beta},b,\boldsymbol{B}),h(\boldsymbol{X}_i,\boldsymbol{B})\big) \\ &= min\left\{1,\frac{b}{|r(Y_i,\boldsymbol{X}_i,\boldsymbol{\beta},b,\boldsymbol{B})h(\boldsymbol{X}_i,\boldsymbol{B})|}\right\} \end{aligned}$$

is the Huber weighting function for the cutoff, b. If an observation's corrected residual or leverage is too large, then the observation's contribution towards ψ will be decreased.

The MrKAT statistic structure being a sum of marginal test statistics leads us to further observations regarding which marginal response drives an overall association. We note that the individual summands are computed separately and may be stored, and then the proportion T_j/T could suggest the contribution of individual response statistics to the overall MrKAT statistic. Upon exploration through simulation study (Appendix Tables B.6-B.7), we note that the ratios T_j/T corresponding to null marginal responses tend to be small compared to ratios from non-null marginal responses in some settings. Thus, when a significant global association is detected using MrKAT, viewing the decomposition of T_j/T may guide researchers to candidate marginal responses for further analysis. In the case of no underlying global association and non-significant p-values for T, we tend to observe $T_j \approx \frac{1}{r}$ for all $j = 1,2,...,r$.

We must note, however, that this observation is not yet supported by any probabilistic results. Although we approximate the null distribution of T, deriving the null distribution of $< T_1, T_2, ..., T_r >$ and of T_j/T is a non-trivial problem. The marginal statistics are likely correlated since the marginal responses may be correlated and the T_j are derived using the same covariates. Further, the variation present in each T_j may vary, which Table B.7 suggests to be potentially problematic. In the future, a more rigorous exploration and theoretical support of these empirical observations could be explored.

We have also included a user-friendly implementation MrKAT in the R package, RobKAT. Like DKAT, our MrKAT software incorporates the computationally efficient algorithms of Zhan and Wu

(2017) to approximate the null distribution of our statistic. Using our implementation, the average time to compute the MrKAT statistic with sample size 300, 200 SNPs, IBS kernel, and 3 continuous responses is 0.1769 seconds (2.9 GHz Processor). Further effects on computation time are explored in Appendix Figure B.3. Thus, genome-wide analysis using MrKAT is a computationally practical option. This multivariate kernel association test methodology not only handles continuous and discrete responses, but also MrKAT provides computationally efficient methodology to robustly incorporate non-normal continuous responses and high-leverage covariates.

CHAPTER

5

ROBUST COMPOSITE KERNEL ASSOCIATION TESTING

5.1 Introduction

Testing for associations between phenotypes and genetic markers, such as single nucleotide polymorphisms (SNPs), is a common task, but researchers may also want to consider the possibility that certain environmental exposure groups affect these genetic associations. For example, SNPs in the *15q24-25* region are thought to be associated with lung cancer risk, but analyses have suggested that this genetic risk differs between smokers and non-smokers (Lin et al., 2013). Another example motivating our methodology comes from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) study. In this clinical trial, cognitive functioning scores and genotype information were collected from patients suffering from schizophrenia. As detailed further in Section 5.4, genetic effects on cognitive scores may differ between patients who were exposed to neurotropic herpesviruses and those unexposed. Our main focus in this article is to develop a genetic association test that accounts for these types of gene-environment interaction effects involving exposure groups.

In one type of gene-environment interaction study, the goal is to test known genes and environmental factors that affect each other. These hypothesis tests only focus on characterizing the gene-environment interaction and would guide clinicians to target interventions for the most at-risk populations, as modifying environmental exposures is often more feasible than altering genes (Hutter et al., 2013). There are also many instances of these types of gene-environment interaction tests in the literature (Chatterjee et al., 2006; Wang et al., 2015a).

Alternatively, we are interested in developing methods to identify novel genetic markers by incorporating gene-environment interaction effects, providing new scientific discoveries (Hutter et al., 2013). For example, in the CATIE study, we aim to discover the gene regions that have associations with cognitive functioning in schizophrenia patients; however, we also wish to account for possible differences between exposure groups to improve modeling. If a gene-environment interaction effect is present, studies have shown that tests for an overall genetic marker effect that includes a gene-environment interaction typically would be more powerful than tests for main effects alone (Kraft et al., 2007; Lindström et al., 2009). For these reasons, joint tests are preferred when researchers are interested in discovery of significant genes rather than characterizing a specific gene-environment interaction (Kraft et al., 2007).

An approach for these joint hypothesis tests would be to consider one, individual genetic marker at a time. For example, Kraft et al. (2007) has applied a likelihood ratio test that jointly considers one SNP, one environmental effect, and a gene-environment interaction; however, such methods may be biased when multiple SNPs are associated with a trait (Lin et al., 2013). This problem is in addition to the large multiple-testing burden from testing individual markers. We instead consider multi-marker approaches, kernel-based methods in particular, that analyze genetic marker sets at the gene level, which result in lower multiple testing burden and greater flexibility in modeling. These kernel association tests typically employ semiparametric models that describe complex genetic effects with kernel functions. In the case of tests for marginal genetic effects, a common strategy is a variance component score test (Kwee et al., 2008; Liu et al., 2007; Wu et al., 2011).

While there are many examples of kernel based tests for gene-environment interactions (Lin et al., 2013; Marceau et al., 2015; Wang et al., 2015b; Zhao et al., 2015b), there are fewer kernel association tests for joint gene and interaction problems, in part due to the challenge of combining models for genetic marker and gene-environment interaction similarity structures.

One such method that has been successful, however, is the composite kernel likelihood ratio test (CKLRT) (Zhao et al., 2019). The CKLRT is a kernel-based association test that models the overall genetic effect nonparametrically through a weighted average of genetic main effect and gene-environment interaction kernels. Leveraging the assumption of normally distributed errors, Zhao et al. (2019) use a likelihood ratio test statistic and develop a Monte Carlo-based algorithm to approximate its null distribution. While the weights to construct the overall genetic kernel could be selected by the analyst, the approximation algorithm also maximizes the likelihood over a range of potential weights.

A key assumption used to build the likelihood and null distribution in the CKLRT methodology is that the errors and response are normally distributed, but researchers may not wish to make such distributional assumptions. In tests of marginal genetic effects, for example, using normality-dependent kernel association tests leads to decreased power when the underlying errors are skewed, heavy-tailed, or have undefined moments (Martinez et al., 2020a). If the distributional assumptions that construct the likelihood are violated, the accuracy of the derived null distribution may also be compromised.

In this article, we propose a robust composite kernel association test (RoCKAT) that performs a joint hypothesis test that does not assume any distributional forms for the errors. To accomplish this, we propose to derive variance component score test statistics that fit the model with a general, robust loss function such as Huber loss. RoCKAT also models the joint gene and interaction similarity structure in a similar fashion to the CKLRT. Unlike the CKLRT, we propose a permutation-based algorithm that approximates the null distribution by permuting within exposure groups. By doing so, we create a method that does not rely on known error distributions and constructs a permutation test without violating the exchangability condition.

This article is arranged as follows. Section 5.2 details the model notation and motivation of our RoCKAT statistic as well as the algorithm to approximate the null distribution. In Section 5.3, we then conduct two simulation studies to compare the performance of RoCKAT to that of the CKLRT. As detailed in Section 5.3.2, our simulations suggest that the CKLRT has inflated type I error when the error distribution is skewed, while RoCKAT controls the type I error. When the errors follow a Gaussian distribution, the CKLRT is more powerful overall; however, the power of RoCKAT is similar to that of the CKLRT when a genetic main effect is present. Section 5.4 then applies our RoCKAT methodology to the CATIE study to analyze the relationship between cognitive scores and the major histocompatibility complex (MHC) on Chromosome 6, while accounting for possible interactions with neurotropic herpesvirsus exposures. In this case study, we find that RoCKAT detects a significant association between reasoning, a skewed cognitive measure, and the MHC region, which literature shows has links to schizophrenia. The only significant association detected with the CKLRT is the *SLC17A3* gene. Finally, Section 5.5 concludes that RoCKAT would be preferred in gene discovery studies where the assumption of normally distributed errors may not hold for all genetic marker sets considered.

5.2 Methods

5.2.1 Model and Notation

We begin with the semiparametric model

$$Y_i = \beta_0 + \boldsymbol{X}_i^T \boldsymbol{\beta}_X + \boldsymbol{E}_i^T \boldsymbol{\beta}_E + h_G(\boldsymbol{G}_i) + h_{GE}(\boldsymbol{G}_i, \boldsymbol{E}_i) + \epsilon_i \tag{5.1}$$

where $(Y_i, \boldsymbol{X}_i, \boldsymbol{G}_i, \boldsymbol{E}_i)$ represents an individual's quantitative response, clinical covariates, genetic marker set, and environmental exposure, respectively. For this methodology, $\boldsymbol{E}_i$ is a vector containing the $l-1$ dummy variables for a nominal categorical variable with l levels. If the exposure is binary as in the CATIE data, then $\boldsymbol{E}_i$ becomes a scalar in $\{0,1\}$. We use β_0 as an intercept term, and ϵ_i as an independent and identically distributed error term among individuals. The potentially semiparametric functions, $h_G(\cdot)$ and $h_{GE}(\cdot,\cdot)$, model the genetic marker and gene-environment interaction effects. As we are interested in discovering genetic regions while leveraging information about environmental interactions, we use the null hypothesis $H_0 : h_G(\cdot) = h_{GE}(\cdot,\cdot) = 0$.

5.2.2 Modeling Functions with Kernels

For the model given in Equation 5.1, we would first seek to model a functional form for the non-parametric $h_G(\boldsymbol{G}_i) + h_{GE}(\boldsymbol{G}_i, \boldsymbol{E}_i)$. While we could choose a parametric form, such as $\boldsymbol{G}_i^T \boldsymbol{\gamma}_0 + \sum_{k=1}^{l-1} E_{ik} \sum_{j=1}^{p} G_{ij} \gamma_{kj}$, this may introduce a large number of parameters, especially if the number of genetic markers, p, is large. Also problematic, the genetic markers may have a far more complex functional form than the linear model specified.

Kernel methods provide a more intuitive way of modeling nonparametric forms such as those in Equation 5.1. Suppose that for a nonparametric function, $h(\cdot)$, we assume that $h \in \mathscr{H}_K$, where $\mathscr{H}_K$ is a reproducing kernel Hilbert space (RKHS) defined by a kernel function $k(\cdot,\cdot)$. For data $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$, the similarity information $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$ can be summarized as the $(i,j)^{th}$ entry in a kernel matrix, K. This kernel function may be any real-valued, positive semi-definite function where the output represents the inner product of the images ($k(\boldsymbol{Z}_i, \boldsymbol{Z}_j) =< h(\boldsymbol{Z}_i), h(\boldsymbol{Z}_j) >$). In this sense, the choice of kernel function defines the inner product information and, therefore, the similarity structure of the data. The benefit is that the data structure is often more intuitive to specify than the nonparametric functional form.

Existing literature provides guidance on choice of kernels for genetic markers, based either on functional form or similarity measures. If, for example, an analyst wished to model the genetic markers with a linear form, $h(\boldsymbol{G}_i) = \boldsymbol{G}_i^T \boldsymbol{\gamma}$ for some coefficients $\boldsymbol{\gamma}$, then a linear kernel, $k(\boldsymbol{G}_i, \boldsymbol{G}_j) = \boldsymbol{G}_i^T \boldsymbol{G}_j$, could be used. If the quadratic functional form, $h(\boldsymbol{G}_i) = \sum_{j<k} G_{ij} Z_{ik} \gamma_{jk} + \sum_{j=1}^{p} G_{ij}^2 \gamma_{2j}$, seemed more appropriate then a quadratic kernel, $k(\boldsymbol{G}_i, \boldsymbol{G}_j) = (\boldsymbol{G}_i^T \boldsymbol{G}_j)^2$, could be utilized. Alternatively, researchers may view the kernel as a similarity measure and use intuition to guide kernel choice. For example, when $\boldsymbol{G}_i$ represents genetic markers, the identity by state (IBS) kernel describes the proportion of alleles shared between individuals i and j, computed as $k(\boldsymbol{G}_i, \boldsymbol{G}_j) = \frac{1}{2p} \sum_{k=1}^{p} [2\mathbb{I}\{G_{ik} = G_{jk}\} + \mathbb{I}\{|G_{ik} - G_{jk}| = 1\}]$.

To specify a gene-environment interaction kernel that adequately captures the similarity structure, existing literature also provides some options to combine individual K_E and K_G. One such kernel is produced by element-wise multiplication, $K_{GE} = K_E \odot K_G$, of the two individual kernels (Marceau et al., 2015; Wu et al., 2013; Zhao et al., 2015a,1). If the environmental exposure is scalar and a K_E is a linear kernel, then K_{GE} can also be represented by $K_{GE} = diag(E) K_G diag(E)$.

In addition to providing intuitive modeling, the RKHS structure may simplify tests involving nonparametric functions. We couple the RKHS assumption and some regularity conditions with $h(\cdot)$ being the solution to a penalized loss criterion, then we apply the Representer Theorem to view $h(\cdot)$ as a finite linear combination of kernel function images (Kimeldorf and Wahba, 1971; Wahba, 1990). When $\rho(t) = \frac{1}{2}(t^2)$ is a square loss function, the minimizing solution to our penalized loss criterion has been shown to be the same as the best linear unbiased predictor from a linear mixed model

$$\boldsymbol{Y} = \boldsymbol{\beta}_0 + \boldsymbol{X}\boldsymbol{\beta}_X + \boldsymbol{E}\boldsymbol{\beta}_E + \boldsymbol{h} + \boldsymbol{\epsilon}$$

with $\boldsymbol{h} \sim N(0, \tau K)$ and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$ (Liu et al., 2007). With this mixed model representation, our test for the effect of $\boldsymbol{G}_i$, $H_0 : h(\cdot) = 0$, is translated to $H_0 : \tau = 0$. This greatly simplifies our testing process, since only $[\beta_0\ \boldsymbol{\beta}_X^T\ \boldsymbol{\beta}_E^T]^T$ requires estimation under the null model. With this simplification, kernel machine-based variance component score tests that estimate parameters with least squares are popular in the literature (Kwee et al., 2008; Liu et al., 2007; Wu et al., 2010; Zhao et al., 2015b).

5.2.3 RoCKAT Statistic

Our goal is to build a test statistic for the hypothesis $H_0 : h_G(\cdot) = h_{GE}(\cdot, \cdot) = 0$ in Equation 5.1 while making few assumptions on the error distribution. As described in Section 5.2.2, there exist many methods to model the similarity structures for genetic markers, environmental exposures, and gene-environment interactions separately. In our method, we aim to model these data structures jointly.

As a way to combine the genotype kernel and the gene-environment interaction kernel, we use $K_{G+GE} = \lambda K_G + (1-\lambda) K_{GE}$ with $0 \leq \lambda \leq 1$. Here, K_{G+GE} is a weighted average of the genotype kernel (K_G) and the gene-environment interaction kernel (K_{GE}), where $\lambda = 1$ models no interaction effect and $\lambda = 0$ models no genetic marker main effect. We also use the fact that weighted combinations of kernel matrices are valid kernels (Shawe-Taylor and Cristianini, 2004). The properties of this composite kernel have been explored by Wu et al. (2013) and incorporated into the Composite Kernel Likelihood Ratio Test (CKLRT) (Zhao et al., 2019).

While other methods such as Ionita-Laza et al. (2013) strategically select λ values beforehand, simulation studies have shown these methods to be less powerful when both genetic main effects and interactions are present (Zhao et al., 2019). Therefore, we follow the CKLRT methodology and maximize λ over the range of possible values.

The next challenge for our robust composite kernel association test is that many of the existing methods assume an error distribution, typically Gaussian. In the CKLRT, this assumption is necessary when constructing the likelihood. Many variance component score test methods also are designed for normally-distributed errors, as parameter values are estimated using a square loss function, $\rho(t) = \frac{1}{2}t^2$ (Kwee et al., 2008; Liu et al., 2007; Wu et al., 2010; Zhao et al., 2015b). Another reason for this dependence on normality is that the resulting null distribution often is a Chi-squared mixture, which can be computed using Davies method or a Satterthwaite approximation (Liu et al., 2007).

While many of these kernel association tests use the square loss function, $\rho(t) = \frac{1}{2}(t^2)$, recent literature has relaxed this requirement by incorporating robust parameter estimates (Kong et al., 2016; Martinez et al., 2020a,2). The quantile regression kernel machine (QRKM), for example, relaxes distributional assumptions by replacing the square loss function, which is designed for Gaussian errors, with the check loss function $\rho(t) = \eta t\ \mathbb{I}_{\{t\geq 0\}} + (\eta - 1)t\ \mathbb{I}_{\{t<0\}}$ where $\eta \in (0,1)$ is the conditional quantile of interest (Kong et al., 2016). While least squares kernel machines fit the null model with a least squares regression, QRKM fits the null model with the check loss to produce quantile regression estimates and employs a permutation approach to approximate the statistic's null distribution.

In the Robust Kernel Association Test (RobKAT), these variance component score statistics are generalized further to accommodate any loss function, subject only to a few conditions (Martinez et al., 2020a). Rather than pre-specifying a loss function or error distribution, RobKAT allows $\rho : \mathbb{R} \to \mathbb{R}^+$ to be any differentiable function such that

1. $\rho(0) = 0$
2. $\rho(x)$ is a convex function of x
3. $\psi(x) = \frac{d}{dx}\rho(x)$ exists and is non-decreasing

These requirements are based on common assumptions in the robust statistics literature (Maronna et al., 2006). As discussed in Chapter 3 Section 3.5, there is evidence that the redescending loss functions excluded by Assumption 3 are valid in certain circumstances. Allowing the relaxation of Assumption 3, we note that the choice of square loss and check loss functions make least squares-based kernel machine and QRKM methods special cases of RobKAT. Like QRKM, RobKAT uses a permutation approach; however, the exact distribution is approximated with a Pearson III distribution via the moment-matching algorithm of Zhan and Wu (2017).

Combining the composite kernel from the CKLRT with the robust loss functions of RobKAT, our

RoCKAT test statistic is

$$T = \max_{0\leq\lambda\leq 1} \boldsymbol{w}^T[\lambda K_G + (1-\lambda)K_{GE}]\boldsymbol{w}$$

Following RobKAT notation, $w_i = \psi(\frac{\widehat{\epsilon}_i}{\widehat{s}})$ where $\widehat{\epsilon}_i = Y_i - \widehat{\beta}_0 - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_X + \boldsymbol{E}_i^T\widehat{\boldsymbol{\beta}}_E$ are the robust residuals, $\widehat{s}$ is the Huber's Proposal 2 robust estimate of scale, and $\psi(\cdot)$ is the derivative of the selected loss function $\rho(\cdot)$.

5.2.4 Null Distribution and Inference

To compute the observed test statistic, $T_{obs} = \max_{0\leq\lambda\leq 1} \boldsymbol{w}^T[\lambda K_G + (1-\lambda)K_{GE}]\boldsymbol{w}$, the analyst inputs the desired genotype and gene-environment interaction kernels. For the subsequent maximization step, we make the simplifying observation that our statistic is maximized at either $\lambda = 1$ or $\lambda = 0$:

$$\begin{aligned}
\max_{0\leq\lambda\leq 1} \ \boldsymbol{w}^T\big[\lambda K_G + (1-\lambda)K_{GE}\big]\boldsymbol{w} &= \max_{0\leq\lambda\leq 1} \ \boldsymbol{w}^T\big[\lambda(K_G - K_{GE}) + K_{GE}\big]\boldsymbol{w} \\
&= \max_{0\leq\lambda\leq 1} \lambda\Big\{\boldsymbol{w}^T\big[K_G - K_{GE}\big]\boldsymbol{w}\Big\} + \Big\{\boldsymbol{w}^T\big[K_{GE}\big]\boldsymbol{w}\Big\} \\
&= \max_{0\leq\lambda\leq 1} \lambda\Big\{\boldsymbol{w}^T K_G \boldsymbol{w} - \boldsymbol{w}^T K_{GE}\boldsymbol{w}\Big\} + \Big\{\boldsymbol{w}^T K_{GE}\boldsymbol{w}\Big\} \\
&= \max\Big\{\boldsymbol{w}^T K_G \boldsymbol{w},\ \boldsymbol{w}^T K_{GE}\boldsymbol{w}\Big\}
\end{aligned}$$

While computing the statistic is relatively simple, finding the null distribution is a more complex task. Similar kernel-based methods for gene-environment interaction leverage asymptotic results, assumed error distributions, or likelihoods when computing the null distribution (Liu et al., 2007; Zhao et al., 2019). Since we are using a robust, general loss function and not specifying the error distribution's functional form, we cannot mirror these approaches. As an alternative, we employ a permutation approach.

Permutation-based methods for gene-environment interaction studies are a difficult task (Bůžková et al., 2011). In RobKAT under the null hypothesis, permuting the rows and columns of the genotype kernel, K_G, does not affect the residuals under the null model. Since all permutations are equally likely under the null hypothesis, exchangability is not violated in the permutation approximation algorithm Zhan and Wu (2017).

In our model, unfortunately, both $K_{G+GE} := \lambda K_G + (1-\lambda)K_{GE}$ and the null model contain the environmental effect. Permutations of K_{G+GE} or K_{GE} without accounting for the environmental effect risk violates the exchangability assumption, leading to inflated type I errors (Anderson and Robinson, 2001; Zhan and Wu, 2017). To approximate the null distribution of T, we develop a permutation algorithm that accounts for levels of the environmental exposure.

We propose to permute the rows of $\boldsymbol{G}$ within the levels coded by $\boldsymbol{E}$, and thus we also rearrange the rows and columns of K_G and K_{GE}. We denote such a permutation of K_G and K_{GE} with K_G^* and K_{GE}^*. For a binary environmental effect, for example, one would permute separately the observations with $E_i = 0$ and $E_i = 1$. For a categorical environmental effect with l levels, we permute within each combination of the $l-1$ dummy variables. Within each level coded by $\boldsymbol{E}$ under the null hypothesis, the response values are equally likely regardless of K_{G+GE} and the exchangability condition is not violated.

We then calculate the statistic

$$T^* = \max_{0\le\lambda\le1} \boldsymbol{w}^T[\lambda K_G^* + (1-\lambda)K_{GE}^*]\boldsymbol{w} = \max_{\lambda\in\{0,1\}} \boldsymbol{w}^T[\lambda K_G^* + (1-\lambda)K_{GE}^*]\boldsymbol{w}$$

for each permutation. Since we are permuting the $\boldsymbol{G}_i$ values, we also gain the computational advantage that the null model and residuals only need to be computed once.

Next, our algorithm repeats this permutation step many times to generate $\{T_j^* : j = 1, 2, ..., N_p\}$. that approximate the null distribution of T. Although each individual T_j^* incorporates only the genotype kernel ($\lambda = 1$) or gene-environment interaction kernel ($\lambda = 0$), the collection of T^* may contain a mixture of the two types of statistics. Finally, to compute the p-value, our algorithm computes the proportion of T_j^* that are greater or equal to the observed test statistic, T_{obs}.

5.3 Simulation Study

To explore and compare the performance of RoCKAT and the CKLRT, we simulate data from the model

$$Y_i = \beta_0 + \boldsymbol{X}_i^T \boldsymbol{\beta}_X + h_E(E_i) + c_G * h_G(\boldsymbol{G}_i) + c_{GE} * h_{GE}(\boldsymbol{G}_i, E_i) + \epsilon_i$$

for $i = 1, 2, ..., 100$. The term, $\beta_0 = 1$, serves as an intercept, and $\boldsymbol{X}_i \underset{i.i.d}{\sim} N(0, \mathbb{I}_2)$ represent two covariates with slope parameters $\boldsymbol{\beta}_X = \mathbf{1}_2$. We generated $\boldsymbol{G}_i$ by randomly sampling from the $p = 787$-SNP, major histocompatibility complex (MHC) region on Chromosome 6, whose biological relevance is described with further detail in Section 5.4. In the generation of Y_i, only the first 20 SNPs were causal and made any contribution. We used the simple, linear functions $h_G(\boldsymbol{G}_i) = \sum_{j=1}^p G_{ij}$ to model the underlying genetic marker and gene-environment interaction effects. To modulate the genetic and interaction effect sizes, we use the constants, c_G and c_{GE}, with different settings throughout the simulation study.

5.3.1 Simulation Study Design

We then perform two simulation studies that explore the effect of normally distributed errors ($\epsilon_i \underset{i.i.d.}{\sim} N(0,1)$) and skewed errors ($\epsilon_i \underset{i.i.d.}{\sim} \chi_1^2$). In the first study, we generate a scalar, binary exposure variable such that $P(E_i = 1) = 0.5$. We then model this binary exposure with $h_E(E_i) = E_i \beta_E$. The gene-environment interaction effect was modeled linearly with $h_{GE}(\boldsymbol{G}_i, E_i) = \sum_{i=1}^p G_{ij} E_i$. The similarity structure for $\boldsymbol{E} = [E_1, ..., E_n]^T$ was modeled with a linear kernel, $K_E = \boldsymbol{E}\boldsymbol{E}^T$. In the second study, $E_i \in \{1, 2, 3\}$ is a nominal categorical variable with $P(E_i = 1) = P(E_i = 2) = P(E_i = 3)$. The three levels of E_i in this simulation are represented by two indicator variables ($E_i^{(2)} = I\{E_i = 2\}$ and $E_i^{(3)} = I\{E_i = 3\}$) with slope parameters $\beta_{E^{(2)}} = \beta_{E^{(3)}} = 1$ and $h_E(E_i) = E^{(2)}\beta_{E^{(2)}} + E^{(3)}\beta_{E^{(3)}}$. In this setting, the gene-environment interaction is modeled by $h_{GE}(\boldsymbol{G}_i, E_i) = \sum_{i=1}^p G_{ij}(E^{(2)} + 2E^{(3)})$. The linear environmental kernel is constructed on the model matrix, $\boldsymbol{E}_{(2,3)} = [\boldsymbol{E}^{(2)} | \boldsymbol{E}^{(3)}]$, so that $K_E = \boldsymbol{E}_{(2,3)} \boldsymbol{E}_{(2,3)}^T$

After generating these data, we compare the testing performance of RoCKAT and the CKLRT. For both methods, we modeled the genotype similarity structure (K_G) with an IBS kernel and the gene-environment interaction structure with $K_{GE} = K_G \odot K_E$. For the RoCKAT implementation, we use $N_p = 500$ permutations. When implementing the CKLRT, we used 10,000 iterations to generate the empirical null distribution of the likelihood ratio test, and we searched 200 values of λ between 0 and 1.

To evaluate the type I error when testing $H_0 : h_G(\cdot) = h_{GE}(\cdot, \cdot) = 0$, we set $c_G = c_{GE} = 0$. The type I error rate was then estimated by the proportion of $N = 1,000$ iterations below a 0.05 significance level. To compare the power of RoCKAT and CKLRT under the two error distributions, we chose settings with various combinations of genetic and interaction effects. First, we generated data with a genetic effect but no interaction effect by setting $c_G \in \{0.1, 0.3, 0.5\}$ and $c_{GE} = 0$. Second, we investigated the power when only an interaction effect was present by setting $c_G = 0$ and $c_{GE} \in \{0.1, 0.3, 0.5\}$. Third, we investigated power when both genetic and interaction effects were similarly weighted $c_G = c_{GE} \in \{0.1, 0.3, 0.5\}$. Finally, we compared the power when both genetic and interaction effects were present but differentially weighted. In this setting, we implemented the settings $(c_G = 0.2, c_{GE} = 0.3)$ and $(c_G = 0.3, c_{GE} = 0.2)$

5.3.2 Simulation Results

As shown in Table 5.1, both RoCKAT and the CKLRT adequately control the type I error rate when the errors are normally distributed in both the binary and multi-level E_i settings. When the errors are skewed ($\epsilon_i \underset{i.i.d.}{\sim} \chi_1^2$), however, only RoCKAT controls the type I error rate. The simulation results for the CKLRT show substantially inflated type I error rates for both the binary and multi-level E_i settings. This inflation is likely due to its dependence on Gaussian distributions when constructing likelihoods and approximating the empirical null distribution.

Table 5.1 Type I error simulation showing the proportion of tests with p-value less than $\alpha = 0.05$. The simulation uses IBS genetic marker kernel, sample size $n = 100$, and $1,000$ iterations at each setting of β. RoCKAT used 500 permutations per test, CKLRT used $N = 5000$ generated variables for the empirical null distribution of the LRT. Data generated using the MHC gene ($p = 787$ SNPs). Bold numbers represent values that fall 1.96 standard errors or more above the $\alpha = 0.05$ target.

E Data Type	Error	RoCKAT	CKLRT
Binary	$N(0,1)$	0.044	0.053
	χ_1^2	0.043	**0.181**
3-Level	$N(0,1)$	0.048	0.062
	χ_1^2	0.043	**0.188**

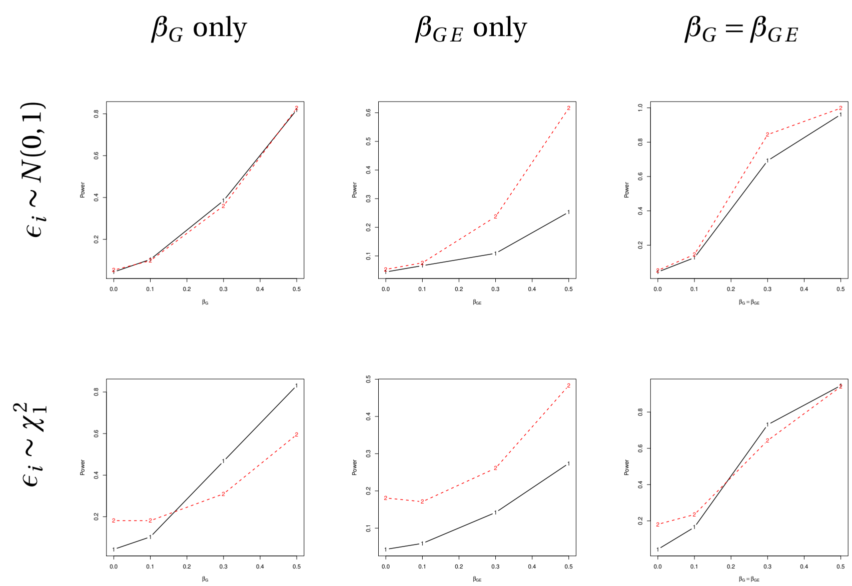


Figure 5.1 Power simulations of RoCKAT (1: black solid line) and CKLRT (2: red dashed line). The simulation uses IBS genetic marker kernel, sample size $n = 100$, and $1,000$ iterations at each setting of β. RoCKAT used 500 permutations per test, CKLRT used $N = 5000$ generated variables for the empirical null distribution of the LRT. The number of causal SNPs (k) is 20, and $\boldsymbol{X}_i \sim N(0, \mathbb{I}_2)$. Data generated using $E_i \in \{0,1\}$ and the MHC gene ($p = 787$ SNPs) with $h_G(\boldsymbol{G}_i) = \boldsymbol{G}_i^T \mathbf{1}_k$, and $h_{GE}(\boldsymbol{G}_i, \boldsymbol{E}_i) = \sum_i^k G_{ik} E_i$.

The power curves for the $E_i \in \{0,1\}$ setting are shown in Figure 5.1. When the errors are normally distributed in the $E_i \in \{0,1\}$ setting, we see that the CKLRT and RoCKAT have similar power when a genetic main effect is present. In the case of only a genetic main effect, the power curves are nearly identical, but the CKLRT has a slight power advantage when both interaction and genetic effects are present. In the case where only a gene-environment interaction effect is present, the CKLRT is substantially more powerful. As shown in Figure 5.2, these trends are similar in the multi-level case.

When the errors are Chi-squared in the $E_i \in \{0,1\}$ setting, the power estimates for the RoCKAT methodology is similar to the estimates in the normal case. The power curves for the CKLRT, however, show that the type I error is notably inflated. In the setting with only a genetic main effect and the setting with both genetic and interaction effects, we also see that RoCKAT is more powerful with observed maximum power gains of 50.6% and 13.5% over the CKLRT, respectively. As shown in Figure 5.2, these trends are repeated in the multi-level exposure case, with maximum power gains of 47.8% and 5.3%, respectively.

The bar charts displayed in Figures 5.3-5.4 compare the testing power of RoCKAT and the CKLRT when genetic and interaction effects are differentially weighted. In both the ($c_G = 0.2$, $c_{GE} = 0.3$)

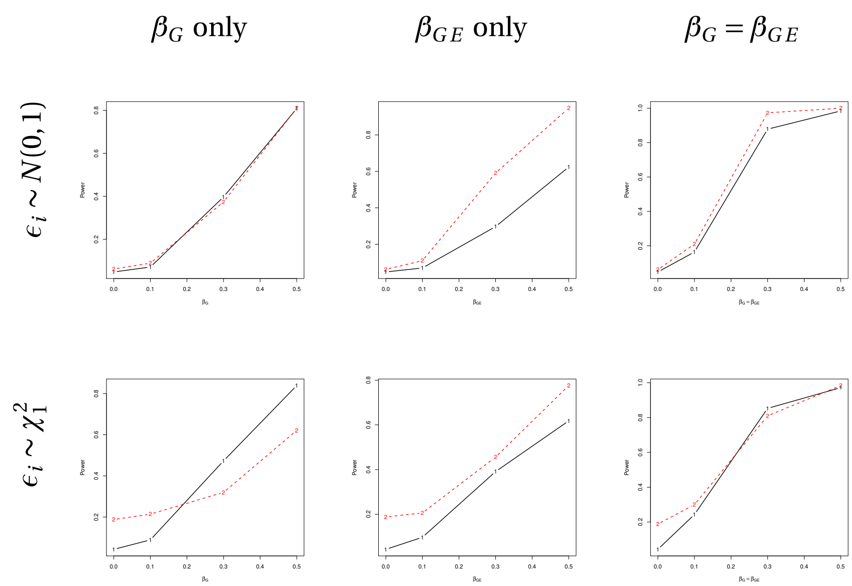


Figure 5.2 Power simulations of ROCKAT (1: black solid line) and CKLRT (2: red dashed line). The simulation uses IBS genetic marker kernel, sample size $n = 100$, and $1,000$ iterations at each setting of β. ROCKAT used 500 permutations per test, CKLRT used $N = 5000$ generated variables for the empirical null distribution of the LRT. The number of causal SNPs (k) is 20, and $\boldsymbol{X}_i \sim N(0, \mathbb{I}_2)$. Data generated using $E_i = [E_i^{(2)}\ E_i^{(3)}]^T$ as a 3-level factor variable and the MHC gene ($p = 787$ SNPs) with $h_G(\boldsymbol{G}_i) = \boldsymbol{G}_i^T \mathbf{1}_k$, and $h_{GE}(\boldsymbol{G}_i, \boldsymbol{E}_i) = \sum_i^k G_{ik}(E_i^{(2)} + 2E_i^{(3)})$.

and ($c_G = 0.3, c_{GE} = 0.2$) settings, the CKLRT is more powerful than RoCKAT when the errors are normally distributed; however, RoCKAT has a slight power advantage in both settings when the errors are skewed. These results suggest that RoCKAT's handling of skewed errors makes its testing power competitive with the CKLRT, despite RoCKAT's maximization algorithm only selecting among $\lambda = 1$ and $\lambda = 0$.

5.4 Application to CATIE Cognitive Scores Data

The Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) study was a clinical trial that compared the effectiveness of atypical antipsychotics on 1,460 unrelated patients suffering from schizophrenia (Lieberman et al., 2005). A secondary purpose, however, was to conduct one of the first double-blind, randomized, large-scale studies that investigated the connection between cognitive scores and antipsychotic medications (Keefe et al., 2003). Impaired cognitive functioning in areas such as memory, reasoning, and social skills is a notable feature of schizophrenia, and some researchers consider it a reliable predictor of poor long-term outcome (Green and Nuechterlein, 1999;

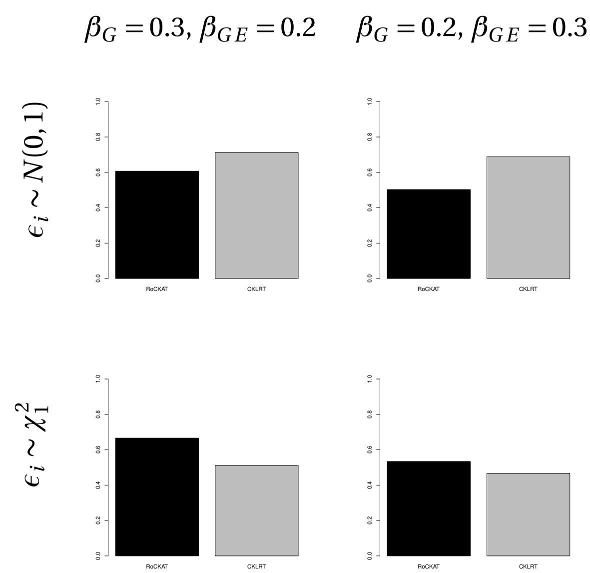


Figure 5.3 Power simulations of ROCKAT (black) and CKLRT (grey). The simulation uses IBS genetic marker kernel, sample size $n = 100$, and $1,000$ iterations at each combination of parameters. ROCKAT used 500 permutations per test, CKLRT used $N = 5000$ generated variables for the empirical null distribution of the LRT. The number of causal SNPs (k) is 20. Data generated using the MHC gene ($p = 787$ SNPs) and $\boldsymbol{E}_i$ is a scalar in $\{0,1\}$.

Keefe et al., 2006,0). Literature has also linked cognitive impairment to many negative symptoms of schizophrenia (Addington et al., 1991).

In the CATIE study, cognitive functioning was measured by five domains (processing speed, verbal memory, vigilance, reasoning, and working memory), which were borrowed from the MATRICS 7-domain scoring method (Green et al., 2004; Keefe et al., 2006). To collect these data, two testers with adequate training were assigned to each site, and 1,493 neurocognitive tests were given on patients with confirmed schizophrenia diagnoses (Keefe et al., 2003). Our data excludes 33 patients who belonged to sites with fewer than 15 patients. The scores for each domain represent the average of standardized Z-scores from tests in the given domain (Keefe et al., 2006; Stroup et al., 2005). This matches the domain scoring method for MATRICS 7-domain methodology (Green et al., 2004; Keefe et al., 2006).

As an exploratory objective, the CATIE study also involved the genotyping of 492,000 SNPs from the genetic samples of 51% of enrolled patients (Sullivan et al., 2008). In this paper, we focus on 11 genes on chromosome 6p22.1 near the Major Histocompatibility Complex (MHC) region, since genome-wide association studies have previously linked this gene region with schizophrenia (Aberg

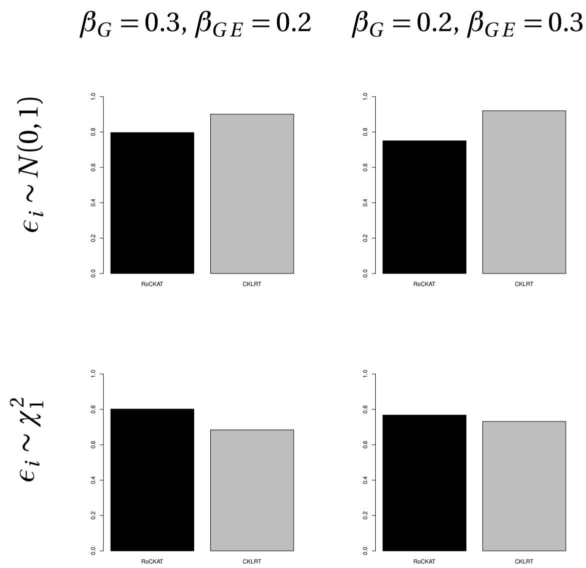
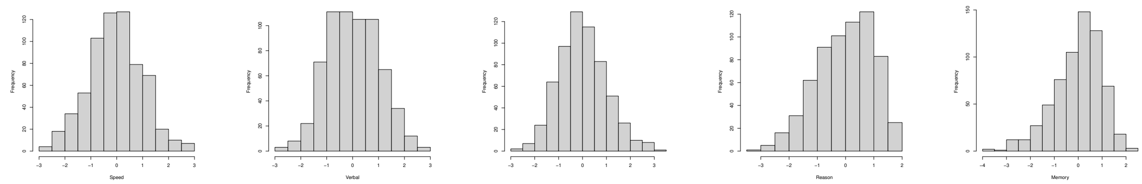


Figure 5.4 Power simulations of ROCKAT (black) and CKLRT (grey). The simulation uses IBS genetic marker kernel, sample size $n = 100$, and $1,000$ iterations at each combination of parameters. ROCKAT used 500 permutations per test, CKLRT used $N = 5000$ generated variables for the empirical null distribution of the LRT. The number of causal SNPs (k) is 20. Data generated using the MHC gene ($p = 787$ SNPs) and $\boldsymbol{E}_i$ is a 3-level factor variable.

et al., 2013; International Schizophrenia Consortium et al., 2009; Shi et al., 2009). With these data in mind, our analyses explore associations between genes in the MHC region and neurocognitive scores.

Because immune responses to infectious diseases also is linked to genetic variation in the MHC region, whether a patient has been exposed to certain viruses may be important to consider. In particular, neurotopic herpesviruses have shown significant association with the cognitive impairments associated with schizophrenia disorders (Prasad et al., 2011; Yolken et al., 2011). Researchers in the CATIE trial obtained blood samples and measured Ig class antibodies to three neurotropic herpesviruses: Herpes Simplex Virus type one (HSV-1), Herpes Simplex Virus type two (HSV-2), and Cytomegalovirus (CMV) (Yolken et al., 2011). Perhaps these neurotropic herpesvirus exposures also modulate the genetic marker effect of genes in the MHC on cognitive functioning.

Figure 5.5 shows the distribution of cognitive functioning scores for each domain. Although processing speed, verbal memory, and vigilance appear Gaussian, the distributions of the reasoning and working memory scores are notably skewed left. This suggests that an analysis of these domain scores may require methods, such as RoCKAT, that are robust to non-Gaussian error distributions.



(a) Processing Speed **(b)** Verbal Memory **(c)** Vigilance **(d)** Reasoning **(e)** Working Memory

Figure 5.5 Distribution of cognitive functioning standardized scores, separated by the five domains explored in the CATIE study.

Previous studies using robust methods have found significant associations between HSV-2 and four genes in the MHC region on Chromosome 6 (Martinez et al., 2020a). Thus, our analysis focuses on the environmental exposure, HSV-2.

In our analysis, we tested for an overall genetic effect between the MHC region and neurocognitive scores using data collected in the CATIE study. Each combination of the 12 gene-regions and five cognitive domains were analyzed with both the CKLRT and RoCKAT methods. Both statistical tests were implemented with IBS genetic kernel, linear environmental kernel, and the covariates age and sex. HSV-2 exposure was included as a binary variable, with a positive exposure defined as an Ig antibody level greater than the 1.2 optical density ratio cutoff used by Yolken et al. (2011). RoCKAT was implemented using 10,000 permutations and a Huber loss function. The CKLRT implementation used the default of 10,000 randomly generated normal variables when approximating the null distribution.

The results of our analysis are shown in Table 5.2. No significant genetic associations were detected for the processing speed, verbal memory, working memory, and vigilance domains after a Bonferroni multiple-testing correction for 12 genes and 2 methods ($\alpha^* = 0.05/24 = 0.00208$). For the reasoning cognitive functioning domain, RoCKAT found a significant association with the MHC region ($p = 0.0005$), while the CKLRT did not ($p = 0.0021$). Due to previous analyses and literature connecting the MHC to schizophrenia, perhaps there is an underlying association that the CKLRT fails to detect (Aberg et al., 2013; Martinez et al., 2020a). This conclusion would also support the power trends observed in Figures 5.1-5.2, where RoCKAT shows greater power than the CKLRT when the conditional response is skewed and there is a genetic main effect present. The CKLRT also found a significant association ($p = 0.0020$) with the *SLC17A3* gene and reasoning score, while RoCKAT

Table 5.2 Data analysis of CATIE cognitive score data using RobKAT (Huber loss) and the CKLRT, while incorporating possible gene-environment interaction with HSV-2 exposure. The significance threshold was determined via a Bonferroni correction based on 12 gene analyses and 2 methods (significance level $= 0.05/24 = 0.00208$). Significant results are shown in bold.

Gene	SNPs	Test	Verbal	Speed	Reason	Memory	Vigilance
BTN2A1	4	RoCKAT	0.1100	0.0668	0.0058	0.0314	0.2983
		CKLRT	0.1010	0.1745	0.0172	0.0510	0.3868
BTN2A2	5	RoCKAT	0.7976	0.596	0.0193	0.0066	0.3980
		CKLRT	1.0000	1.0000	0.0021	0.0803	1.0000
BTN3A2	5	RoCKAT	0.803	0.7758	0.4918	0.5337	0.6332
		CKLRT	1.0000	1.0000	0.0962	1.0000	1.0000
HIST1H2AG	2	RoCKAT	0.125	0.9181	0.3969	0.1484	0.9460
		CKLRT	0.3958	1.0000	0.5148	1.0000	1.0000
HIST1H2BJ	2	RoCKAT	0.3217	0.8536	0.3901	0.7585	0.8256
		CKLRT	1.0000	1.0000	1.0000	1.0000	0.4797
MHC	787	RoCKAT	0.0228	0.1486	**0.0005**	0.0117	0.0900
		CKLRT	0.1344	0.1014	0.0021	0.0718	0.1072
NOTCH4	24	RoCKAT	0.8244	0.1124	0.1883	0.0313	0.1598
		CKLRT	1.0000	0.0998	0.0283	0.0098	0.1589
POM121L2	4	RoCKAT	0.2276	0.2855	0.4215	0.1441	0.2970
		CKLRT	0.4710	0.2940	1.0000	0.2608	0.4769
PRSS16	1	RoCKAT	0.4837	0.5056	0.5351	0.0987	0.8178
		CKLRT	1.0000	1.0000	1.0000	0.2530	1.0000
SLC17A1	9	RoCKAT	0.4138	0.7558	0.1441	0.2745	0.6547
		CKLRT	0.5145	1.0000	0.3065	1.0000	1.0000
SLC17A3	13	RoCKAT	0.6684	0.5411	0.0362	0.1114	0.5336
		CKLRT	0.4092	0.1094	**0.0020**	0.0028	0.1836
ZNF184	8	RoCKAT	0.2003	0.6807	0.0219	0.8506	0.5069
		CKLRT	0.5367	1.0000	0.1100	1.0000	0.5554

did not ($p = 0.0362$). This may be a spurious finding from the CKLRT, since our simulation study suggests that the CKLRT's type I error rate is inflated when the conditional response is skewed.

Expanded results for HSV-1 and CMV exposures can be found in Appendix Tables C.1-C.2. When including HSV-1, RoCKAT only detects a significant association with the MHC region ($p = 0.0018$), while the CKLRT does not ($p = 0.0041$). No significant associations for any of the cognitive domains were detected when analyses included CMV. These results are partially corroborated by previous tests of CMV detecting no associations with cognitive impairment (Torniainen-Holm et al., 2018).

5.5 Discussion

In this article, we have developed a robust composite kernel association test (RoCKAT) for the overall genetic effect with no functional forms assumed for the error distribution. We accomplish this task by constructing a variance component score test and modeling the overall genetic effect's similarity structure with a weighted average of the genetic and interaction kernels. The composite kernel thus considers the genetic main effect and possible interactions with environmental exposure groups. By using a score-type test, RoCKAT only requires estimation of parameters in the linear null model, and we may estimate these parameters using robust regression to incorporate distributional robustness into our test statistic. Such robust regressions implement a Huber loss function rather than the typical least squares estimation processes. Since RoCKAT does not depend on specifying a distribution, we develop a permutation-based algorithm to approximate the null distribution of our statistic. To preserve exchangability in this permutation test, a key strategy of our algorithm is to permute within levels of the environmental factor.

As detailed in Section 5.3, we have compared the performance of the RoCKAT and CKLRT methods when the underlying errors are generated from a normal and χ_1^2 distribution. We then explored these comparisons through two simulation studies with binary and multi-level exposure variables. For both exposure types, RoCKAT and the CKLRT adequately control the type I error rate when the error is normally distributed. With this error distribution, we also see that RoCKAT has similar power to the CKLRT when a genetic main effect is present. For the case of χ_1^2 distributed errors, only RoCKAT controls the type I error rate, while the CKLRT produces substantially inflated type I error rates.

In Section 5.3.2, we also explore testing power when the interaction and genetic main effects have different signal weights. Because the CKLRT inference algorithm may choose any value $0 \le \lambda \le 1$, we see that the CKLRT is more powerful in this instance when the errors are normally distributed. Although RoCKAT's maximization algorithm only selects among $\lambda \in \{0, 1\}$, its proper handling of skewed error distributions still gives RoCKAT a power advantage when the errors are Chi-squared.

These simulation results suggest that if the analyst is confident in making a normality assumption, the CKLRT would likely be the more appropriate option. This recommendation is further supported by literature suggesting that likelihood ratio tests tend to be more powerful than score tests in kernel machine regression models, since likelihood ratio tests contain parameter estimates under both null and the alternative hypotheses (Zeng et al., 2014). If the underlying error distribution is not normally distributed, however, performing the analysis with RoCKAT is preferred. Otherwise, the analyst risks a substantial number of false positive results due to the uncontrolled type I error. We also observe in these simulations that RoCKAT is more powerful in some cases with non-normal error distributions.

As detailed in Section 5.4, we applied both RoCKAT and the CKLRT to data collected during the CATIE study. These data included SNPs from 12 genes within the MHC region on Chromosome 6, five measures of cognitive functioning, and the neurotropic herpesvirus antibody levels from patients suffering from schizophrenia. In our case study, we tested for associations between each type of cognitive score and genetic marker set, while accounting for possible interactions resulting from neurotropic herpesvirus exposure. After applying a multiple-testing correction, the only significant associations involved the reasoning domain, which has a skewed distribution. RoCKAT detected an overall genetic association between the MHC region, while the CKLRT did not. Due to previous literature supporting the MHC's connection to schizophrenia, it is possible that there exists an underlying association that the CKLRT is not powerful enough to detect. Another result from Section 5.3.2 is that the CKLRT detected a significant association between the overall effect of the *SLC17A3* gene and reasoning score; however, this may be a spurious finding, as simulations involving the CKLRT have shown inflated type I error when the error distribution is skewed.

We must also note that our RoCKAT methodology is designed for environmental exposure groups, which are represented by nominal categorical variables. For quantitative exposure variables, the exchangability condition required for a permutation test may be violated and the type I error would become inflated (Bůžková et al., 2011). While it is possible to partition a continuous exposure variable into discrete levels then apply RoCKAT, the composite kernel often does not capture the similarity structure well and results in decreased power (See Appendix C.3). Future work is required to produce a composite kernel association test that is designed to incorporate continuous environmental exposures without assuming a functional form for the error distribution and without violating the exchangability condition in its permutation algorithm.

Although RoCKAT is implemented with a permutation-based algorithm, each test with our R package only takes an average of 0.89 seconds for a sample size of 100, 2 covariates, 787 SNPs, and 500 permutations (2.9 GHz Processor). This computational speed is competitive with that of the CKLRT and allows RoCKAT to be used in gene-environment wide interaction studies. Considering its robustness to distributional errors and ability to account for possible interactions with exposure groups, RoCKAT is a useful tool for discovering novel genetic associations and enhance analyses in genetic epidemiology research.

CHAPTER

6

FUTURE RESEARCH

This dissertation has presented a variety of novel methods in robust kernel association testing. Chapter 3 described a kernel-based test for univariate responses that is robust to non-normal error distributions. In Chapter 4, this univariate method is generalized to a multivariate response comprised of both continuous and categorical phenotypes. Finally, Chapter 5 shows that these robust methods can also be tailored to other problems, such as gene-environment interaction tests. However, there are more avenues to explore in robust kernel association testing, and each of the methods proposed in this thesis have potential extensions and applications to other problems.

In the case of RobKAT, there are opportunities to tailor the method to other data types and make the software more accessible. As discussed in Chapter 2, SKAT has had a variety of extensions to microbiome data, copy number variants, and tests for epistasis. Due to the methods' similarity, there is potential for RobKAT to be easily applied to these types of data analyses. The SKAT R package also includes a pipeline to input binary format files (.bed, .bim, and .fam) that are commonly used by PLINK and other genomics software. As data cleaning for genetic data in PLINK is convenient, having this option would make RobKAT, MrKAT, and RoCKAT far more attractive for use in genome-wide association studies.

In the MrKAT methodology of Chapter 4, it may be possible to determine the specific responses that drive a significant joint association, which is a relatively unexplored topic in the literature (Larson et al., 2019). As previously discussed, the individual summands of the MrKAT statistic may be stored and analyzed. Tentative empirical evidence suggests that the proportion, T_j/T, may provide information about the contributions of each response to the global test statistic (see Appendix Tables B.6-B.7). Unfortunately, this empirical evidence currently lacks any rigorous probabilistic results to refine this intuition. Some major hurdles in this research are the possible correlations between the T_j and the fact that variation present in each T_j may vary. In their development of composite kernels as weighted averages, Wu et al. (2013) encountered a similar problem with scaling quadratic-form score statistics. Their solution was to scale the d^{th} statistic by $\gamma_d = tr\{P_0^{1/2} K_d P_0^{1/2}\}$ where $P_0^{1/2} = I - X(X^T X)^{-1} X^T$, where X contains the covariate values and K_d is the d^{th} statistic's kernel matrix. With this in mind, perhaps there is a suitable scaling of the T_j that would allow individual contributions to be compared and interpreted.

With additional time or funding, further exploration of RoCKAT and the research of Chapter 5 would likely produce additional applications. Firstly, an extension of RoCKAT for binary responses would be feasible if the robust regression estimates of the null model were modified. As in Chapter 4, the Huber ψ-function could be replaced with the weighted residuals from robust logistic regression, which are designed to downweight high leverage observations; however, further research would be required to verify that these modifications produce the same benefits as with MrKAT. Secondly, it may also be possible to extend RoCKAT to a multivariate response in the same way that RobKAT was extended to the multivariate, MrKAT. This type of analysis would allow us to jointly analyze the five cognitive functioning domain scores, rather than testing each neurocognitive score with our five separate univariate analyses as in Chapter 5. With this approach and future research, a method may be developed to better incorporate response correlations and perhaps even determine which of the five dimensions drive the overall association.

As mentioned by Larson et al (2019) in his review of kernel-based hypothesis tests, the computational burden for large sample sizes is another issue for many kernel association testing methods. In particular, the permutation-based RoCKAT suffers from this issue, as the kernel matrices must be rearranged with each permutation. Unfortunately, there are few robust KAT with known asymptotic distributions, making permutation approaches necessary. Due to the exchangability condition in permutation tests, there is also the difficulty of including continuous environmental exposures in permutation-based interaction tests (Anderson and Robinson, 2001; Bůžková et al., 2011). Dichotomizing continuous variables into binary exposures then applying RoCKAT is possible, but we see in Appendix Table C.3 that this approach is often underpowered. Additionally, a naive permutation-based KAT in a gene-environment interaction model that does not account for the continuous environmental exposures often leads to inflated type I errors.

To create a robust gene-environment interaction KAT without approximating the null distribution via permutations, an M-estimation approach may be feasible. The robust methods developed in Chapters 3 - 5 start with M-estimates, and there are well-established theoretical properties for generalized score, Wald, and likelihood ratio tests that depend only on M-estimates (Boos and Stefanski, 2013). Perhaps there is some way to mirror the CKLRT method by Zhao et al (2019) with a generalized likelihood ratio test.

Alternatively, there has been literature exploring the Tukey 1-df model in gene-environment interaction hypothesis tests, and a generalized score test may be used to create a robust test for gene-environment interaction (Chatterjee et al., 2006; Pan et al., 2011; Sengupta and Jammalamadaka, 2003). For the model $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}_X + \boldsymbol{E}_i^T \boldsymbol{\beta}_E + h_G(\boldsymbol{G}_i) + h_{GE}(\boldsymbol{G}_i, \boldsymbol{E}_i) + \epsilon_i$, one could follow Marceau et al. (2015) by constructing a genotype kernel and performing kernel PCA to represent h_G with the dimension-reduced linear form, $\boldsymbol{Z}_i^T \boldsymbol{\beta}_Z$. The Tukey 1-df test would then use the hypothesis, $H_0 : \gamma = 0$, on the model, $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}_X + \boldsymbol{E}_i^T \boldsymbol{\beta}_E + \boldsymbol{Z}_i^T \boldsymbol{\beta}_Z + \gamma \boldsymbol{Z}_i^T \boldsymbol{\beta}_Z \boldsymbol{E}_i^T \boldsymbol{\beta}_E + \epsilon_i$ to test for gene-environment interaction. Perhaps a generalized score test then could be derived using the Huber loss in an M-estimate, which would place few assumptions on the distribution of ϵ_i.

From the collection of methods discussed in this dissertation, we see that assumptions of normality are not always reasonable. While both the literature of robust statistics and of kernel methods are each important, the combination of these two research areas also provides a useful tool. There are too many data problems and research questions involving non-normality for robustness to remain overlooked. With this in mind, robust kernel association testing research is relatively new, and I expect that the area will lead to many more exciting methods in the future.

REFERENCES

Aberg, K. A., Liu, Y., Bukszár, J., McClay, J. L., Khachane, A. N., Andreassen, O. A., Blackwood, D., Corvin, A., Djurovic, S., Gurling, H., Ophoff, R., Pato, C. N., Pato, M. T., Riley, B., Webb, T., Kendler, K., O'Donovan, M., Craddock, N., Kirov, G., Owen, M., Rujescu, D., St Clair, D., Werge, T., Hultman, C. M., Delisi, L. E., Sullivan, P., and van den Oord, E. J. (2013). A Comprehensive Family-Based Replication Study of Schizophrenia Genes. *JAMA Psychiatry*, 70(6):573.

Addington, J., Addington, D., and Maticka-Tyndale, E. (1991). Cognitive functioning and positive and negative symptoms in schizophrenia. *Schizophrenia Research*, 5:123–134.

Agresti, A. (2013). *Categorical Data Analysis*. Wiley, Hoboken, NJ, 3rd edition.

Anderson, M. J. and Robinson, J. (2001). Permutation tests for linear models. *Australian and New Zealand Journal of Statistics*, 43(1):75–88.

Basu, S. and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology*, 35(7):606–619.

Boos, D. and Stefanski, L. A. (2013). *Essential Statistical Inference: Theory and Methods*. Springer, New York, New York.

Broadaway, K. A., Cutler, D. J., Duncan, R., Moore, J. L., Ware, E. B., Jhun, M. A., Bielak, L. F., Zhao, W., Smith, J. A., Peyser, P. A., Kardia, S. L., Ghosh, D., and Epstein, M. P. (2016). A Statistical Approach for Testing Cross-Phenotype Effects of Rare Variants. *The American Journal of Human Genetics*, 98(3):525–540.

Buhmann, M. D. M. D. (2003). *Radial basis functions : theory and implementations*. Cambridge University Press.

Bůžková, P., Lumley, T., and Rice, K. (2011). Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. *Annals of Human Genetics*, 75(1):36–45.

Carroll, R. J. and Pederson, S. (1993). On Robustness in the Logistic Regression Model.

Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U., and Wacholder, S. (2006). Powerful Multi-locus Tests of Genetic Association in the Presence of Gene-Gene and Gene-Environment Interactions. Technical report.

Davenport, C. A., Maity, A., Sullivan, P. F., and Tzeng, J.-Y. (2017). A Powerful Test for SNP Effects on Multivariate Binary Outcomes using Kernel Machine Regression. *Statistics in Biosciences*.

Davies, R. (1980). The distribution of a linear combination of chi-square random variables. *J.R.Stat. Soc. Ser. C Appl. Stat.*, (29):323–333.

Gauderman, W. J., Murcray, C., Gilliland, F., and Conti, D. V. (2007). Testing association between disease and multiple SNPs in a candidate gene. *Genetic Epidemiology*, 31(5):383–395.

Girirajan, S., Johnson, R. L., Tassone, F., Balciuniene, J., Katiyar, N., Fox, K., Baker, C., Srikanth, A., Yeoh, K. H., Khoo, S. J., Nauth, T. B., Hansen, R., Ritchie, M., Hertz-Picciotto, I., Eichler, E. E., Pessah, I. N., and Selleck, S. B. (2013). Global increases in both common and rare copy number load associated with autism. *Human Molecular Genetics*, 22(14):2870–2880.

Green, M. F. and Nuechterlein, K. H. (1999). Should schizophrenia be treated as a neurocognitive disorder? *Schizophrenia Bulletin*, 25(2):309–318.

Green, M. F., Nuechterlein, K. H., Gold, J. M., Barch, D. M., Cohen, J., Essock, S., Fenton, W. S., Frese, F., Goldberg, T. E., Heaton, R. K., Keefe, R. S., Kern, R. S., Kraemer, H., Stover, E., Weinberger, D. R., Zalcman, S., and Marder, S. R. (2004). Approaching a consensus cognitive battery for clinical trials in schizophrenia: The NIMH-MATRICS conference to select cognitive domains and test criteria. *Biological Psychiatry*, 56:301–307.

Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., Spencer, K. L., Kwan, S. Y., Noureddine, M., Gilbert, J. R., Schnetz-Boutaud, N., Agarwal, A., Postel, E. A., and Pericak-Vance, M. A. (2005). Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration. *Science*, 308:419–421.

Hampel, F. R. (1986). *Robust statistics : the approach based on influence functions*. Wiley, New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, 2nd edition.

Hua, W.-Y. and Ghosh, D. (2015). Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies. *Biometrics*, 71(3):812–820.

Huber, P. J. (1977). *Robust statistical procedures*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1 edition.

Huber, P. J. (1981). *Robust statistics*. Wiley, Sommerset, NJ, 1 edition.

Huber, P. J. (1996). *Robust statistical procedures*. SIAM, Philadelphia, 2 edition.

Hutter, C. M., Mechanic, L. E., Chatterjee, N., Kraft, P., Gillanders, E. M., Abnet, C. C., Amos, C., Balshaw, D., Bickeböller, H., Bierut, L. J., Boffetta, P., Bondy, M., Chanock, S., Chen, H. S., Cox, N., Vivo, I. D., Divi, R., Dupuis, J., Ellison, G., Fallin, M. D., Gauderman, W. J., Gillanders, E., Haiman, C., Hutter, C., Simonds, N. I., Iversen, E., Khoury, M. J., Marchand, L. L., McAllister, K., Mechanic, L., Peters, U., Prentice, R., Rebbeck, T., Reedy, J., Rothman, N., Schully, S., Seminara, D., Shaughnessy, D., Shete, S., Spiegelman, D., Stram, D. O., Thomas, D., Wang, M., Wang, W., Weinberg, C., Winn, D. M., and Witte, J. S. (2013). Gene-environment interactions in cancer epidemiology: A national cancer institute think tank report. *Genetic Epidemiology*, 37(7):643–57.

International Schizophrenia Consortium, I. S., Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–52.

Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A., and Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature Genetics*, 29(3):306–309.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., and Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics*.

Janeway, C., Travers, P., Walport, M., and Shlomchik, M. (2001). *Immunobiology: the immune system in health and disease*. Garland Science, New York, 5th edition.

Josse, J., Pagès, J., and Husson, F. (2008). Testing the significance of the RV coefficient. *Computational Statistics and Data Analysis*.

Kazi-Aoual, F., Hitier, S., Sabatier, R., and Lebreton, J.-D. (1995). Refined approximations to permutation tests for multivariate inference. *Computational Statistics & Data Analysis*, 20(6):643–656.

Keefe, R. S., Bilder, R. M., Harvey, P. D., Davis, S. M., Palmer, B. W., Gold, J. M., Meltzer, H. Y., Green, M. F., Miller, D. D., Canive, J. M., Adler, L. W., Manschreck, T. C., Swartz, M., Rosenheck, R., Perkins, D. O., Walker, T. M., Stroup, T. S., McEvoy, J. P., and Lieberman, J. A. (2006). Baseline neurocognitive deficits in the CATIE schizophrenia trial. *Neuropsychopharmacology*, 31:2033–2046.

Keefe, R. S., Mohs, R. C., Bilder, R. M., Harvey, P. D., Green, M. F., Meltzer, H. Y., Gold, J. M., and Sano, M. (2003). Neurocognitive assessment in the clinical antipsychotic trials of intervention effectiveness (CATIE) project schizophrenia trial: Development, methodology, and rationale. *Schizophrenia Bulletin*, 29(1):45–55.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95.

Kong, D., Maity, A., Hsu, F.-C., and Tzeng, J.-Y. (2016). Testing and estimation in marker-set association study using semiparametric quantile regression kernel machine. *Biometrics*, 72(2):364–371.

Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J., and Gauderman, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Human Heredity*, 63(2):111–119.

Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (1989). Conditionally Unbiased Bounded-Influence Estimation in General Regression Models, with Applications to Generalized Linear Models. *Journal of the American Statistical Association*, 84(406):460–466.

Kwee, L. C., Liu, D., Lin, X., Ghosh, D., and Epstein, M. P. (2008). A Powerful and Flexible Multilocus Association Test for Quantitative Traits. *The American Journal of Human Genetics*, 82(2):386–397.

Larson, N. B., Chen, J., and Schaid, D. J. (2019). A review of kernel methods for genetic association studies. *Genetic Epidemiology*, 43:122–136.

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., NHLBI GO Exome Sequencing Project—ESP Lung Project Team, N. G. E. S. P. L. P., Christiani, D. C., Wurfel, M. M., and Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American journal of human genetics*, 91(2):224–37.

Li, B. and Leal, S. M. (2008). Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *The American Journal of Human Genetics*, 83(3):311–321.

Lieberman, J. A., Stroup, T. S., McEvoy, J. P., Swartz, M. S., Rosenheck, R. A., Perkins, D. O., Keefe, R. S., Davis, S. M., Davis, C. E., Lebowitz, B. D., Severe, J., and Hsiao, J. K. (2005). Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. *New England Journal of Medicine*, 353(12):1209–1223.

Lin, X., Lee, S., Christiani, D. C., and Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics*, 14(4):667–681.

Lindström, S., Yen, Y. C., Spiegelman, D., and Kraft, P. (2009). The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. *Human Heredity*, 68:171–181.

Liu, D., Ghosh, D., and Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC bioinformatics*, 9:292.

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics*, 63(4):1079–1088.

Madsen, B. E. and Browning, S. R. (2009). A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genetics*, 5(2):e1000384.

Maity, A., Sullivan, P. F., and Tzeng, J.-Y. (2012). Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic epidemiology*, 36(7):686–95.

Maity, A., Zhao, J., Sullivan, P. F., and Tzeng, J.-Y. (2018). Inference on phenotype-specific effects of genes using multivariate kernel machine regression. *Genetic Epidemiology*, 42(1):64–79.

Marceau, R., Lu, W., Holloway, S., Sale, M. M., Worrall, B. B., Williams, S. R., Hsu, F.-C., and Tzeng, J.-Y. (2015). A Fast Multiple-Kernel Method With Applications to Detect Gene-Environment Interaction. *Genetic Epidemiology*, 39(6):456–468.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, Chichester, UK.

Martinez, K., Maity, A., Yolken, R., Sullivan, P., and Tzeng, J.-y. (2020a). Robust Kernel Association Testing (RobKAT). *Genetic Epidemiology*, 44(3):272–282.

Martinez, K., Maity, A., Yolken, R. H., Wu, M. C., and Tzeng, J.-Y. (2020b). Multivariate Robust Kernel Association Testing (MrKAT).

NCI-NHGRI Working Group on Replication in Association Studies (2007). Replicating genotype–phenotype associations. *Nature*, 447(7145):655–660.

Pan, W., Basu, S., and Shen, X. (2011). Adaptive tests for detecting gene-gene and gene-environment interactions. *Human heredity*, 72(2):98–109.

Prasad, K. M., Eack, S. M., Goradia, D., Pancholi, K. M., Keshavan, M. S., Yolken, R. H., and Nimgaonkar, V. L. (2011). Progressive gray matter loss and changes in cognitive functioning associated with exposure to herpes simplex virus 1 in Schizophrenia: A longitudinal study.

American Journal of Psychiatry, 168(8):822–830.

Schloss, P. D. and Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and environmental microbiology*, 71(3):1501–6.

Schrader, R. M. and Hettmansperger, T. P. (1980). Robust Analysis of Variance Based Upon a Likelihood Ratio Criterion. *Biometrika*, 67(1):93.

Sengupta, D. and Jammalamadaka, S. R. (2003). *Linear Models : An Integrated Approach*. World Scientific.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press, 1 edition.

Shi, J., Levinson, D. F., Duan, J., Sanders, A. R., Zheng, Y., Pe'er, I., Dudbridge, F., Holmans, P. A., Whittemore, A. S., Mowry, B. J., Olincy, A., Amin, F., Cloninger, C. R., Silverman, J. M., Buccola, N. G., Byerley, W. F., Black, D. W., Crowe, R. R., Oksenberg, J. R., Mirel, D. B., Kendler, K. S., Freedman, R., and Gejman, P. V. (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature*, 460(7256):753–7.

St Clair, D. (2009). Copy Number Variation and Schizophrenia. *Schizophrenia Bulletin*, 35(1):9–12.

Stefansson, H., Ophoff, R. A., Steinberg, S., Andreassen, O. A., Cichon, S., Rujescu, D., Werge, T., Pietiläinen, O. P. H., Mors, O., Mortensen, P. B., Sigurdsson, E., Gustafsson, O., Nyegaard, M., Tuulio-Henriksson, A., Ingason, A., Hansen, T., Suvisaari, J., Lonnqvist, J., Paunio, T., Børglum, A. D., Hartmann, A., Fink-Jensen, A., Nordentoft, M., Hougaard, D., Norgaard-Pedersen, B., Böttcher, Y., Olesen, J., Breuer, R., Möller, H.-J., Giegling, I., Rasmussen, H. B., Timm, S., Mattheisen, M., Bitter, I., Réthelyi, J. M., Magnusdottir, B. B., Sigmundsson, T., Olason, P., Masson, G., Gulcher, J. R., Haraldsson, M., Fossdal, R., Thorgeirsson, T. E., Thorsteinsdottir, U., Ruggeri, M., Tosato, S., Franke, B., Strengman, E., Kiemeney, L. A., GROUP†, Melle, I., Djurovic, S., Abramova, L., Kaleda, V., Sanjuan, J., de Frutos, R., Bramon, E., Vassos, E., Fraser, G., Ettinger, U., Picchioni, M., Walker, N., Toulopoulou, T., Need, A. C., Ge, D., Lim Yoon, J., Shianna, K. V., Freimer, N. B., Cantor, R. M., Murray, R., Kong, A., Golimbet, V., Carracedo, A., Arango, C., Costas, J., Jönsson, E. G., Terenius, L., Agartz, I., Petursson, H., Nöthen, M. M., Rietschel, M., Matthews, P. M., Muglia, P., Peltonen, L., St Clair, D., Goldstein, D. B., Stefansson, K., Collier, D. A., Kahn, R. S., Linszen, D. H., van Os, J., Wiersma, D., Bruggeman, R., Cahn, W., de Haan, L., Krabbendam, L., and Myin-Germeys, I. (2009). Common variants conferring risk of schizophrenia. *Nature*, 460(7256):744–7.

Stroup, S., Appelbaum, P., Swartz, M., Patel, M., Davis, S., Jeste, D., Kim, S., Keefe, R., Manschreck, T., McEvoy, J., and Lieberman, J. (2005). Decision-making capacity for research participation among individuals in the CATIE schizophrenia trial. *Schizophrenia Research*, 80.

Sullivan, P. F., Lin, D., Tzeng, J.-Y., van den Oord, E., Perkins, D., Stroup, T. S., Wagner, M., Lee, S., Wright, F. A., Zou, F., Liu, W., Downing, A. M., Lieberman, J., and Close, S. L. (2008). Genomewide association for schizophrenia in the CATIE study: results of stage 1. *Molecular Psychiatry*, 13(6):570–584.

Sun, J., Oualkacha, K., Forgetta, V., Zheng, H.-F., Brent Richards, J., Ciampi, A., Greenwood, C. M., and UK10K Consortium, U. (2016). A method for analyzing multiple continuous phenotypes in rare variant association studies allowing for flexible correlations in variant effects. *European journal of human genetics : EJHG*, 24(9):1344–51.

Theodoridis, S. and Koutroumbas, K. (2009). *Pattern recognition*. Academic Press, Burlington MA, 4th edition.

Torniainen-Holm, M., Suvisaari, J., Lindgren, M., Härkänen, T., Dickerson, F., and Yolken, R. H. (2018). Association of cytomegalovirus and Epstein-Barr virus with cognitive functioning and risk of dementia in the general population: 11-year follow-up study. *Brain, Behavior, and Immunity*.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–21.

Tzeng, J.-Y., Magnusson, P. K., Sullivan, P. F., and Szatkiewicz, J. P. (2015). A New Method for Detecting Associations with Rare Copy-Number Variants. *PLoS Genetics*, 11(10):e1005403.

Tzeng, J.-Y. and Zhang, D. (2007). Haplotype-based association analysis via variance-components score test. *American journal of human genetics*, 81(5):927–938.

Wahba, G. (1990). *Spline models for observational data*. Society for Industrial and Applied Mathematics.

Wang, T. and Elston, R. C. (2007). Improved Power by Use of a Weighted Score Test for Linkage Disequilibrium Mapping. *The American Journal of Human Genetics*, 80(2):353–360.

Wang, Y., Li, D., and Wei, P. (2015a). Powerful Tukey's One Degree-of-Freedom Test for Detecting

Gene-Gene and Gene-Environment Interactions. *Cancer informatics*, 14(Suppl 2):209–18.

Wang, Z., Maity, A., Luo, Y., Neely, M. L., and Tzeng, J.-Y. (2015b). Complete effect-profile assessment in association studies with multiple genetic and multiple environmental factors. *Genetic Epidemiology*, 39(2):122–133.

Williams, C. K. I. (2003). Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. *Journal of the American Statistical Association*, 98(462):489–489.

Wu, B. and Pankow, J. S. (2016). Sequence Kernel Association Test of Multiple Continuous Phenotypes. *Genetic Epidemiology*, 40(2):91–100.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010). Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *The American Journal of Human Genetics*, 86(6):929–942.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics*, 89(1):82–93.

Wu, M. C., Maity, A., Lee, S., Simmons, E. M., Harmon, Q. E., Lin, X., Engel, S. M., Molldrem, J. J., and Armistead, P. M. (2013). Kernel machine SNP-set testing under multiple candidate kernels. *Genetic Epidemiology*, 37(3):267–275.

Yolken, R. H., Torrey, E. F., Lieberman, J. A., Yang, S., and Dickerson, F. B. (2011). Serological evidence of exposure to Herpes Simplex Virus type 1 is associated with cognitive deficits in the CATIE schizophrenia sample. *Schizophrenia Research*, 128(1-3):61–65.

Zeng, P., Zhao, Y., Zhang, L., Huang, S., and Chen, F. (2014). Rare variants detection with kernel machine learning based on likelihood ratio test. *PLoS ONE*, 9(3).

Zhan, X., Girirajan, S., Zhao, N., Wu, M. C., and Ghosh, D. (2016). A novel copy number variants kernel association test with application to autism spectrum disorders studies. *Bioinformatics*, 32(23):btw500.

Zhan, X., Tong, X., Zhao, N., Maity, A., Wu, M. C., and Chen, J. (2017a). A small-sample multivariate kernel machine test for microbiome association studies. *Genetic Epidemiology*, 41(3):210–220.

Zhan, X. and Wu, M. C. (2017). Reader Reaction: A note on testing and estimation in marker-set association study using semiparametric quantile regression kernel machine. *Biometrics*.

Zhan, X., Zhao, N., Plantinga, A., Thornton, T. A., Conneely, K. N., Epstein, M. P., and Wu, M. C. (2017b).

Powerful Genetic Association Analysis for Common or Rare Variants with High-Dimensional Structured Traits. *Genetics*, 206(4):1779–1790.

Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. Technical Report 1.

Zhang, Z. (2016). Residuals and regression diagnostics: focusing on logistic regression. *Annals of translational medicine*, 4(10):195.

Zhao, G., Marceau, R., Zhang, D., and Tzeng, J.-Y. (2015a). Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression. *Genetics*, 199(3):695–710.

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H., and Wu, M. C. (2015b). Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. *The American Journal of Human Genetics*, 96(5):797–807.

Zhao, N., Zhang, H., Clark, J. J., Maity, A., and Wu, M. C. (2019). Composite kernel machine regression based on likelihood ratio test for joint testing of genetic and gene–environment interaction effect. *Biometrics*, 75(2):625–637.

Zhu, W. and Zhang, H. (2009). Why do we test multiple traits in genetic association studies? *Journal of the Korean Statistical Society*, 38(1):1–10.

APPENDICES

APPENDIX

A

SUPPLEMENTAL INFORMATION FOR CHAPTER 3

A.1 Type I Error Results for Additional Kernels and Loss Functions with n=100

Table A.1 Simulation study for type I error that includes sample size $n = 100$, 9 SNPs, and IBS kernel (mirroring Table 1) as well as linear and quadratic kernels. The study compares RobKAT (Huber loss), QRKM (LAD loss), and SKAT as well as tests with redescending loss functions, RobKAT (Bisquare loss) and RobKAT (Hampel loss). Type I errors are identical for any $h(\boldsymbol{Z}_i)$ function. Estimates show proportion of $10,000$ iterations with p-value below α cutoff.

			Loss Function (ρ)				
Kernel	Error	α	Huber	LAD	Bisquare	Hampel	SKAT
Linear	t_3	0.01	0.0095	0.0093	0.0113	0.0109	0.0081
		0.05	0.0477	0.0469	0.0461	0.0463	0.0466
Linear	χ_1^2	0.01	0.0107	0.0109	0.0097	0.0107	0.0080
		0.05	0.0518	0.0515	0.0496	0.0499	0.0475
Linear	$N(0,1)$	0.01	0.0112	0.0096	0.0111	0.0110	0.0105
		0.05	0.0468	0.0498	0.0473	0.0466	0.0455
Linear	Cauchy	0.01	0.0104	0.0104	0.0104	0.0108	0.0032
		0.05	0.0482	0.0503	0.0514	0.0515	0.0319
Linear	10% Mix	0.01	0.0104	0.0108	0.0113	0.0112	0.0088
		0.05	0.0492	0.0486	0.0505	0.0500	0.0494
Linear	30% Mix	0.01	0.0102	0.0115	0.0105	0.0101	0.0096
		0.05	0.0509	0.0489	0.0504	0.0504	0.0489
Quadratic	t_3	0.01	0.0090	0.0095	0.0099	0.0106	0.0081
		0.05	0.0478	0.0468	0.0449	0.0463	0.0466
Quadratic	χ_1^2	0.01	0.0101	0.0103	0.0099	0.0106	0.0080
		0.05	0.0492	0.0494	0.0463	0.0462	0.0475
Quadratic	$N(0,1)$	0.01	0.0109	0.0101	0.0109	0.0108	0.0105
		0.05	0.0471	0.0485	0.0474	0.0464	0.0455
Quadratic	Cauchy	0.01	0.0115	0.0115	0.0092	0.0113	0.0032
		0.05	0.0488	0.0496	0.0501	0.0518	0.0319
Quadratic	10% Mix	0.01	0.0111	0.0104	0.0112	0.0106	0.0088
		0.05	0.0512	0.0478	0.0510	0.0512	0.0494
Quadratic	30% Mix	0.01	0.0097	0.0111	0.0088	0.0092	0.0096
		0.05	0.0500	0.0499	0.0504	0.0491	0.0489
IBS	t_3	0.01	0.0100	0.0083	0.0096	0.0101	0.0081
		0.05	0.0456	0.0458	0.0468	0.0478	0.0466
IBS	χ_1^2	0.01	0.0105	0.0113	0.0101	0.0103	0.0080
		0.05	0.0532	0.0507	0.0493	0.0508	0.0475
IBS	$N(0,1)$	0.01	0.0104	0.0089	0.0106	0.0107	0.0105
		0.05	0.0473	0.0492	0.0474	0.0478	0.0455
IBS	Cauchy	0.01	0.0120	0.0109	0.0101	0.0113	0.0032
		0.05	0.0488	0.0509	0.0508	0.0504	0.0319
IBS	10% Mix	0.01	0.0115	0.0094	0.0108	0.0109	0.0088
		0.05	0.0484	0.0476	0.0498	0.0489	0.0494
IBS	30% Mix	0.01	0.0108	0.0103	0.0113	0.0110	0.0096
		0.05	0.0484	0.0487	0.0467	0.0489	0.0489

A.2 Type I Error Results for Additional Kernels and Loss Functions with n=300

Table A.2 Simulation study for type I error with the larger sample size of $n = 300$. The simulation includes IBS kernel and 9 SNPs (mirroring Table 1) as well as linear and quadratic kernels. Methods tested include RobKAT (Huber loss), QRKM (LAD loss), and SKAT as well as tests with redescending loss functions, RobKAT (Bisquare loss) and RobKAT (Hampel loss). Estimates show proportion of 10,000 iterations with p-value below α cutoff.

			Loss Function (ρ)				
Kernel	Error	α	Huber	LAD	Bisquare	Hampel	SKAT
Linear	t_3	0.01	0.0108	0.0097	0.0099	0.0107	0.0090
		0.05	0.0491	0.0464	0.0482	0.0481	0.0479
Linear	χ_1^2	0.01	0.0107	0.0094	0.0095	0.0099	0.0101
		0.05	0.0478	0.0478	0.0459	0.0489	0.0480
Linear	$N(0,1)$	0.01	0.0100	0.0100	0.0101	0.0101	0.0091
		0.05	0.0488	0.0489	0.0489	0.0485	0.0479
Linear	Cauchy	0.01	0.0104	0.0081	0.0096	0.0101	0.0037
		0.05	0.0507	0.0468	0.0508	0.0529	0.0335
Linear	10% Mix	0.01	0.0103	0.0103	0.0111	0.0106	0.0101
		0.05	0.0511	0.0507	0.0505	0.0518	0.0520
Linear	30% Mix	0.01	0.0101	0.0094	0.0103	0.0099	0.0102
		0.05	0.0469	0.0479	0.0480	0.0463	0.0468
Quadratic	t_3	0.01	0.0092	0.0099	0.0094	0.0103	0.0090
		0.05	0.0486	0.0465	0.0462	0.0469	0.0479
Quadratic	χ_1^2	0.01	0.0107	0.0101	0.0112	0.0099	0.0101
		0.05	0.0489	0.0478	0.0467	0.0475	0.0480
Quadratic	$N(0,1)$	0.01	0.0108	0.0117	0.0107	0.0108	0.0091
		0.05	0.0486	0.0491	0.0472	0.0488	0.0479
Quadratic	Cauchy	0.01	0.0097	0.0088	0.0092	0.0101	0.0037
		0.05	0.0494	0.0464	0.0471	0.0502	0.0335
Quadratic	10% Mix	0.01	0.0103	0.0111	0.0114	0.0111	0.0101
		0.05	0.0475	0.0499	0.0503	0.0495	0.0520
Quadratic	30% Mix	0.01	0.0096	0.0100	0.0096	0.0098	0.0102
		0.05	0.0486	0.0454	0.0473	0.0479	0.0468
IBS	t_3	0.01	0.0106	0.0100	0.0103	0.0100	0.0090
		0.05	0.0493	0.0476	0.0480	0.0493	0.0479
IBS	χ_1^2	0.01	0.0094	0.0095	0.0092	0.0092	0.0101
		0.05	0.0491	0.0457	0.0442	0.0473	0.0480
IBS	$N(0,1)$	0.01	0.0107	0.0094	0.0107	0.0109	0.0091
		0.05	0.0463	0.0479	0.0454	0.0465	0.0479
IBS	Cauchy	0.01	0.0108	0.0077	0.0111	0.0105	0.0037
		0.05	0.0498	0.0469	0.0508	0.0519	0.0335
IBS	10% Mix	0.01	0.0109	0.0098	0.0104	0.0103	0.0101
		0.05	0.0502	0.0493	0.0488	0.0507	0.0520
IBS	30% Mix	0.01	0.0105	0.0105	0.0100	0.0101	0.0102
		0.05	0.0434	0.0438	0.0444	0.0440	0.0468

A.3 Power Results for Additional Kernels and Loss Functions with n=100

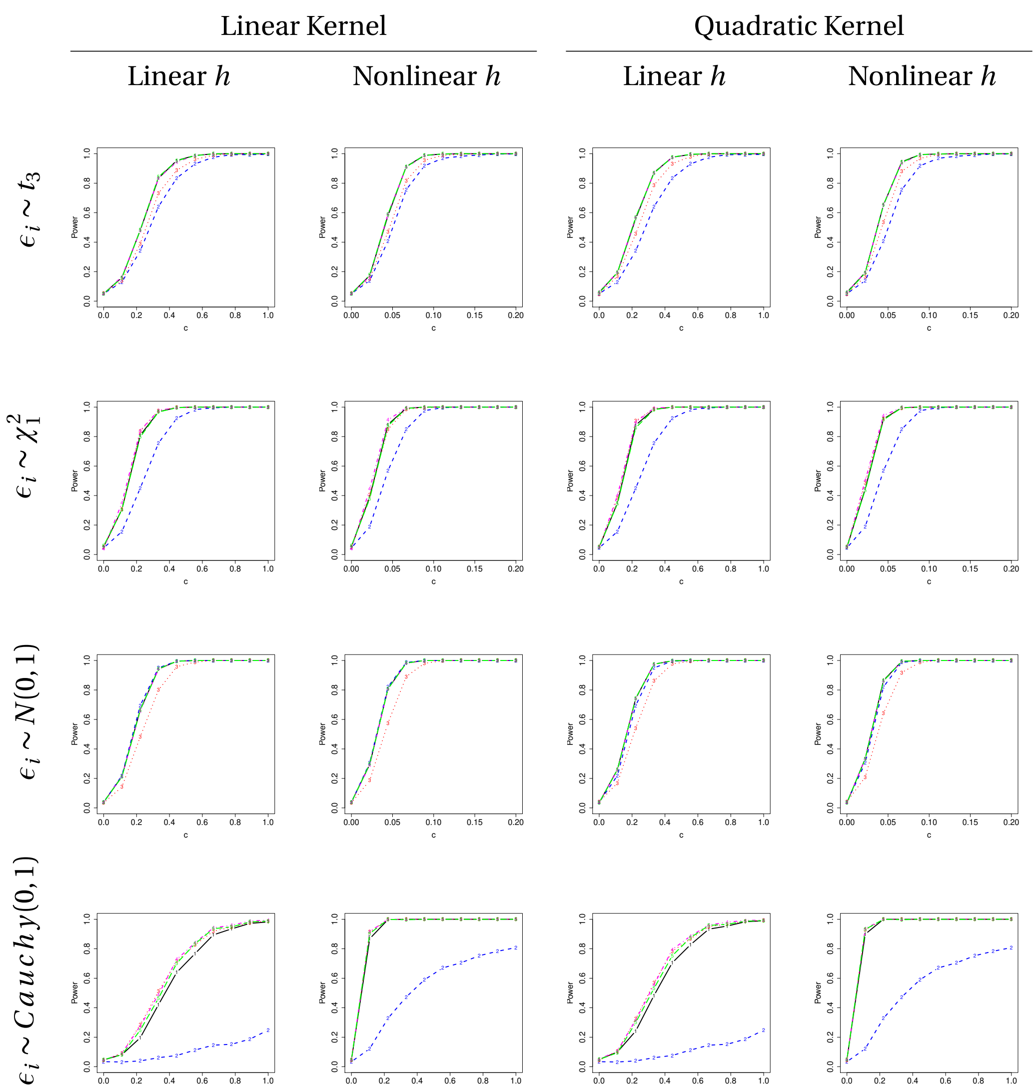


Figure A.1 Power simulations with linear and quadratic kernels of RobKAT Huber (1:black solid line), SKAT (2:blue dashed line), and LAD (3:red dotted line) as well as tests with redescending loss functions, RobKAT Bisquare (4:magenta dot-dashed line) and RobKAT Hampel (5: green long-dash line). The simulation uses sample size $n = 100$, $p = 9$ SNPs, and 1,000 iterations at each setting of c. The linear case is $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \mathbf{1}_9$, and the nonlinear case is $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{p} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{p} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$.

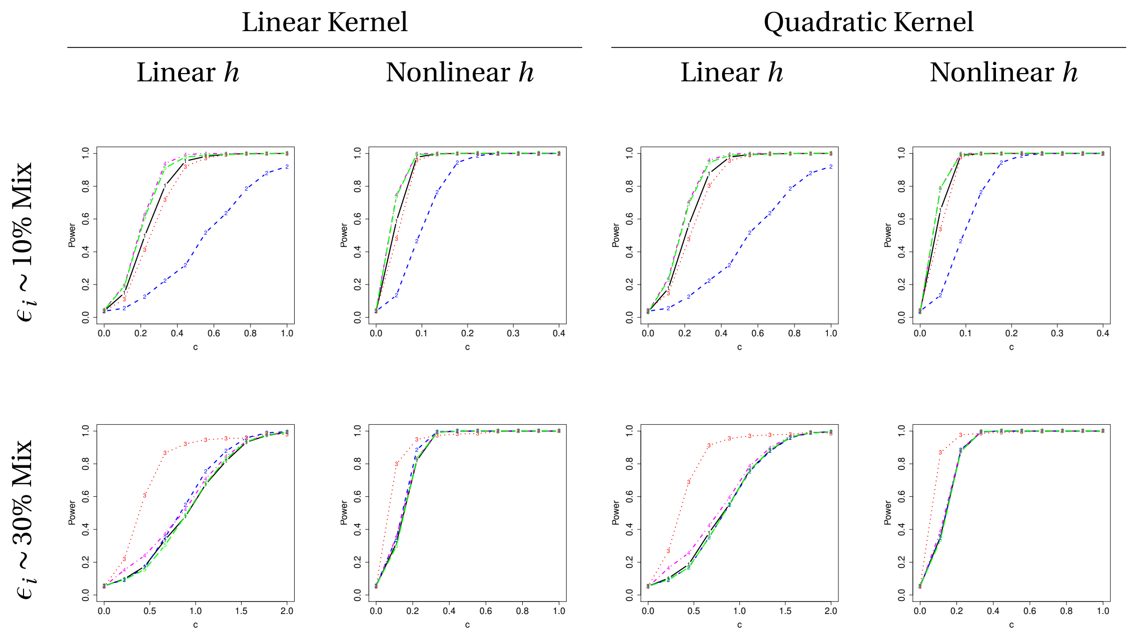


Figure A.2 Power simulations with bimodal errors that include linear and quadratic kernels of RobKAT Huber (1:black solid line), SKAT (2:blue dashed line), and LAD (3:red dotted line), as well as tests with redescending loss functions, RobKAT Bisquare (4: magenta dot-dashed line) and RobKAT Hampel (5: green long-dash line). The simulation uses sample size $n = 100$, $p = 9$ SNPs, and $1,000$ iterations at each setting of c. The linear case is $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \mathbf{1}_9$, and the nonlinear case is $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{p} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{p} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$.

A.4 Power Results for Additional Kernels and Loss Functions with n=300

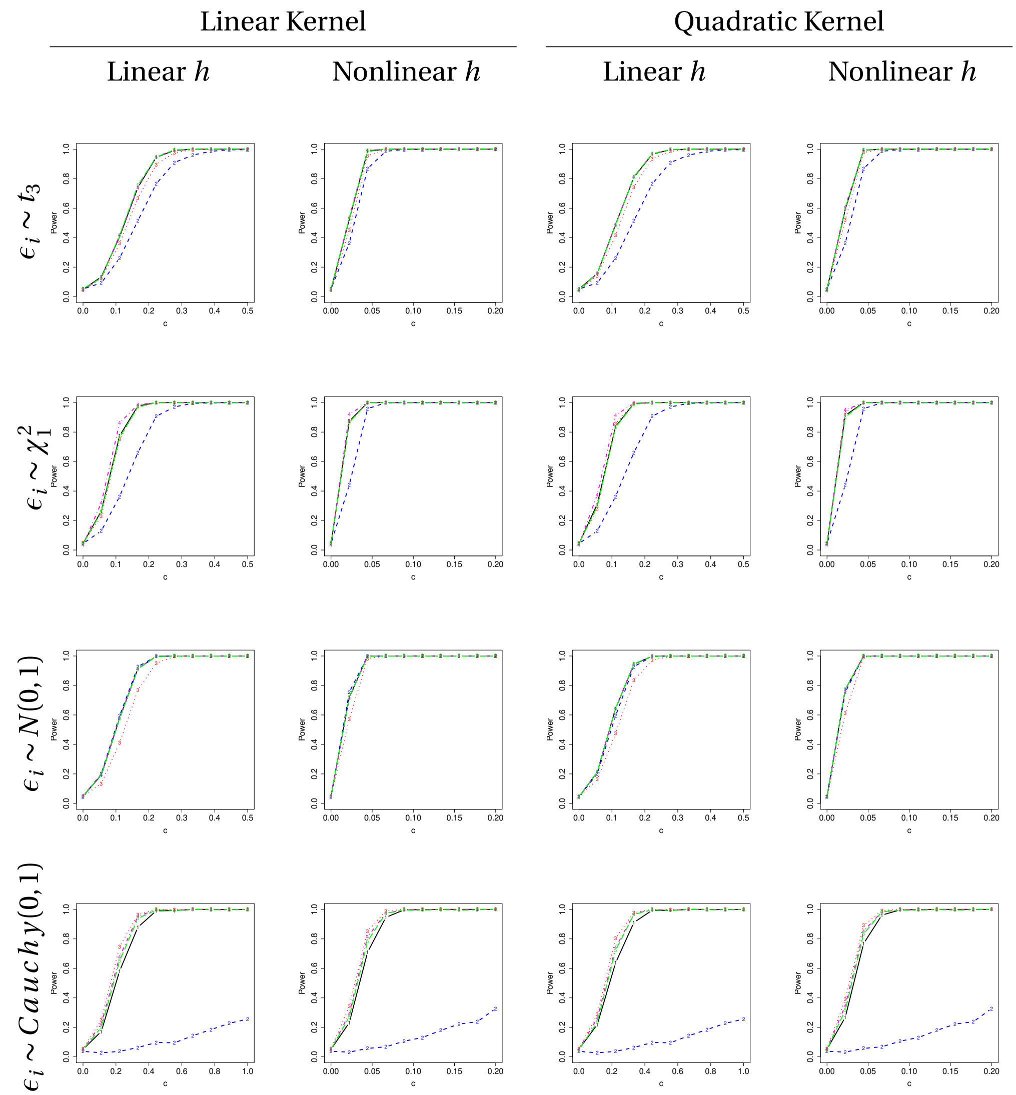


Figure A.3 Power simulations with linear and quadratic kernels of RobKAT Huber (1:black solid line), SKAT (2:blue dashed line), and LAD (3:red dotted line) as well as tests with redescending loss functions, RobKAT Bisquare (4:magenta dot-dashed line) and RobKAT Hampel (5: green long-dash line). The simulation uses sample size $n = 300$, $p = 9$ SNPs, and 1,000 iterations at each setting of c. The linear case is $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \mathbf{1}_9$, and the nonlinear case is $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{p} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{p} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$.

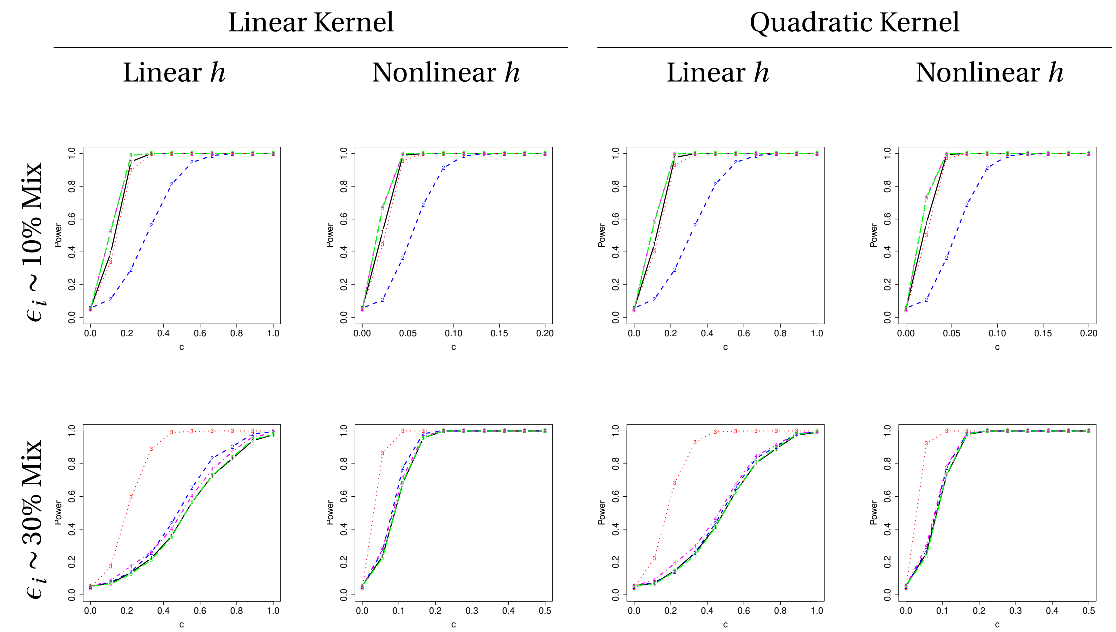


Figure A.4 Power simulations with bimodal errors that include linear and quadratic kernels of RobKAT Huber (1:black solid line), SKAT (2:blue dashed line), and LAD (3:red dotted line) as well as tests with redescending loss functions, RobKAT Bisquare (4: magenta dot-dashed line) and RobKAT Hampel (5: green long-dash line). The simulation uses sample size $n = 300$, $p = 9$ SNPs, and 1,000 iterations at each setting of c. The linear case is $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \mathbf{1}_9$, and the nonlinear case is $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{p} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{p} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$.

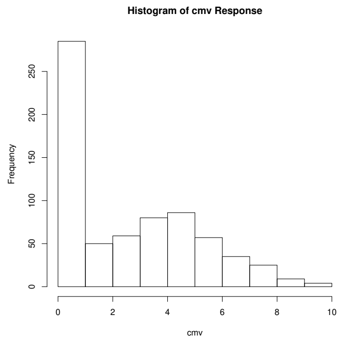
A.5 Full CATIE Data Analysis

Table A.3 RobKAT analysis on CATIE antibodies, mirroring Section 4 and Table 3 in the main text. Here, we include redescending loss functions (Hampel, Bisquare) in addition to Huber loss, LAD loss, and SKAT. The significance threshold was determined via a Bonferroni correction based on 12 gene analyses, 3 kernels, and 5 methods ($\alpha = 0.05/(12*3*5) = 0.00028$). Significant results are shown in bold.

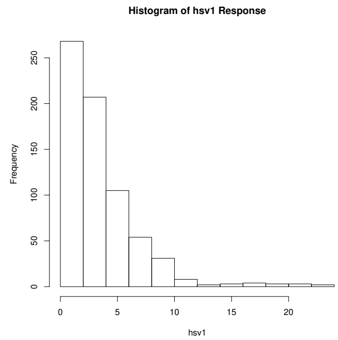
		HSV1			HSV2			CMV		
gene	ψ	linear	quadratic	IBS	linear	quadratic	IBS	linear	quadratic	IBS
BTN2A1	SKAT	0.07679	0.15524	0.10567	0.00176	0.00335	0.00494	0.00150	0.00759	0.00236
	huber	0.00803	0.02631	0.01436	0.00541	0.00639	0.02336	0.00155	0.00863	0.00251
	LAD	0.03475	0.10657	0.04764	0.00106	0.00150	0.00514	0.01024	0.03866	0.01116
	hampel	0.00627	0.02151	0.01173	0.01653	0.01675	0.05807	0.00157	0.00874	0.00253
	bisquare	0.00747	0.02340	0.01460	0.02529	0.02669	0.06940	0.00163	0.00894	0.00256
BTN2A2	SKAT	0.05244	0.08935	0.05702	0.05251	0.06792	0.07835	0.00324	0.00763	0.00286
	huber	0.01515	0.03424	0.01131	0.01443	0.02229	0.01920	0.00517	0.01266	0.00425
	LAD	0.11892	0.21728	0.08933	0.03188	0.07616	0.03467	0.01244	0.01766	0.01149
	hampel	0.01907	0.04121	0.01417	0.04623	0.07018	0.04208	0.00508	0.01254	0.00415
	bisquare	0.02446	0.05532	0.01591	0.15323	0.31035	0.15821	0.00531	0.01271	0.00443
BTN3A2	SKAT	0.03494	0.02282	0.01762	0.07271	0.06011	0.08863	0.04910	0.03126	0.07955
	huber	0.01431	0.00673	0.01121	0.15886	0.19427	0.19263	0.07666	0.04850	0.11918
	LAD	0.00384	0.00212	0.00339	0.84444	0.73582	0.90418	0.31518	0.23253	0.44299
	hampel	0.01803	0.00918	0.01687	0.24953	0.29447	0.29717	0.07092	0.04428	0.11209
	bisquare	0.02563	0.01315	0.02508	0.33129	0.41909	0.38807	0.06861	0.04290	0.10830
HIST1H2AG	SKAT	0.92213	0.93267	0.70116	0.19853	0.23608	0.25445	0.43686	0.66758	0.21429
	huber	0.72157	0.81800	0.85490	0.06201	0.12390	0.03250	0.38483	0.55682	0.25441
	LAD	0.48233	0.38472	0.80218	0.00241	0.00306	0.00407	0.56707	0.73559	0.45521
	hampel	0.67595	0.74697	0.83445	0.06554	0.14932	0.02053	0.38927	0.56423	0.25241
	bisquare	0.73319	0.75167	1.00000	0.12553	0.20302	0.09458	0.41578	0.60284	0.25119
HIST1H2BJ	SKAT	0.79122	0.58380	0.70037	0.01310	0.01899	0.01649	0.18609	0.30461	0.11541
	huber	1.00000	1.00000	1.00000	0.00121	0.00236	0.00129	0.16770	0.25322	0.13759
	LAD	1.00000	1.00000	1.00000	**0.00028**	**0.00018**	0.00098	0.27022	0.21404	0.26863
	hampel	1.00000	1.00000	1.00000	0.00164	0.00388	0.00125	0.16514	0.25126	0.13363
	bisquare	1.00000	1.00000	1.00000	0.10073	0.12027	0.13510	0.17812	0.27359	0.13751
MHC	SKAT	0.04048	0.08628	0.04483	**0.00010**	**0.00016**	**0.00013**	0.10080	0.14007	0.10371
	huber	0.00877	0.01932	0.00660	**2.2314e-06**	**1.7107e-06**	**2.4198e-06**	0.05149	0.06349	0.05627
	LAD	0.03487	0.02926	0.02175	0.00419	0.00159	0.00180	0.07406	0.07532	0.11273
	hampel	0.00847	0.01841	0.00561	**0.00003**	**0.00002**	**0.00003**	0.05116	0.06379	0.05574
	bisquare	0.02272	0.04267	0.01571	0.30234	0.40056	0.34272	0.04891	0.06097	0.05394
NOTCH4	SKAT	0.04606	0.04072	0.06212	0.26256	0.40066	0.28563	0.60714	0.93015	0.64687
	huber	0.03791	0.03982	0.05217	0.06431	0.21994	0.08140	0.43858	0.79330	0.48190
	LAD	0.04826	0.05320	0.07832	0.16485	0.44452	0.25259	0.28569	0.53319	0.33174
	hampel	0.04201	0.04878	0.05736	0.05894	0.21162	0.06706	0.45052	0.80802	0.49656
	bisquare	0.05590	0.05960	0.07839	0.38195	0.49109	0.50436	0.43229	0.79199	0.47528
POM121L2	SKAT	0.26827	0.36745	0.20218	0.00073	0.00146	0.00076	0.10933	0.16350	0.07037
	huber	0.21433	0.29563	0.15850	**0.00023**	0.00042	**0.00021**	0.09777	0.13930	0.07341
	LAD	0.03044	0.05003	0.02184	0.00725	0.01584	0.00337	0.01474	0.01903	0.00989
	hampel	0.30392	0.40973	0.21487	0.00051	0.00079	0.00051	0.09872	0.14196	0.07171
	bisquare	0.24868	0.34978	0.17056	0.99552	0.90453	0.97208	0.09860	0.13897	0.07300
PRSS16	SKAT	0.41067	0.41067	0.41067	0.65578	0.65578	0.65578	0.14173	0.14173	0.14173
	huber	0.24503	0.24503	0.24226	0.32980	0.32980	0.33131	0.07214	0.07214	0.07305
	LAD	0.09568	0.10038	0.10510	0.26890	0.26890	0.26107	0.03620	0.03620	0.03802
	hampel	0.26569	0.26569	0.26187	0.43627	0.43627	0.45089	0.07809	0.07809	0.07887
	bisquare	0.31691	0.31691	0.31277	1.00000	1.00000	0.81803	0.07719	0.07719	0.07807
SLC17A1	SKAT	0.19398	0.17851	0.32453	0.00122	0.00396	0.00250	0.00529	0.01382	0.00961
	huber	0.08473	0.09419	0.14602	**0.00020**	0.00070	0.00062	0.00492	0.01555	0.00913
	LAD	0.24536	0.33055	0.39316	0.00400	0.00265	0.00420	0.03292	0.07575	0.04848
	hampel	0.06916	0.08305	0.11766	0.00037	0.00119	0.00106	0.00491	0.01532	0.00915
	bisquare	0.06557	0.08100	0.11257	0.34847	0.25176	0.24522	0.00565	0.01695	0.01029
SLC17A3	SKAT	0.57656	0.40083	0.66907	0.02843	0.07313	0.05747	0.09016	0.05664	0.06619
	huber	0.48986	0.39620	0.43752	0.01814	0.05846	0.03350	0.10804	0.06930	0.10833
	LAD	0.97778	0.78567	0.66117	0.05555	0.17217	0.07699	0.16769	0.14177	0.13404
	hampel	0.50210	0.43765	0.43122	0.07575	0.15748	0.10140	0.10698	0.06823	0.10676
	bisquare	0.42802	0.44934	0.35395	0.32164	0.37201	0.28366	0.11015	0.06636	0.10430
ZNF184	SKAT	0.58984	0.64917	0.61829	0.29280	0.19342	0.33406	0.32298	0.32206	0.44491
	huber	0.46517	0.51445	0.63247	0.42303	0.39284	0.61290	0.30613	0.30321	0.39351
	LAD	0.41699	0.46274	0.54990	0.58686	0.69201	0.61074	0.10163	0.10491	0.15728
	hampel	0.47430	0.50862	0.65502	0.74338	0.62134	1.00000	0.31411	0.31137	0.40549
	bisquare	0.39693	0.39867	0.55873	0.78451	0.65241	1.00000	0.30434	0.30272	0.39215

Table A.4 RobKAT analysis on CATIE antibodies as described in Section 4, but instead using log-transformed antibody responses and including redescending loss functions (Hampel, Bisquare). The significance threshold was determined via a Bonferroni correction based on 12 gene analyses, 3 kernels, and 5 methods ($\alpha = 0.05/(12*3*5) = 0.00028$). Significant results are shown in bold.

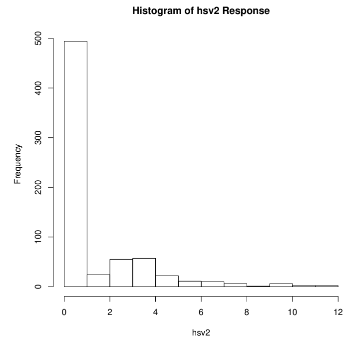
		HSV1			HSV2			CMV		
		linear	quadratic	IBS	linear	quadratic	IBS	linear	quadratic	IBS
BTN2A1	SKAT	0.00043	0.00215	0.00096	0.00046	0.00066	0.00236	0.00054	0.00259	0.00131
	huber	0.00050	0.00270	0.00123	0.00056	0.00073	0.00344	0.00029	0.00167	0.00079
	LAD	0.06215	0.13806	0.08185	0.00114	0.00131	0.00536	0.00608	0.02538	0.00819
	hampel	0.00050	0.00263	0.00123	0.00060	0.00082	0.00351	0.00030	0.00172	0.00085
	bisquare	0.00046	0.00252	0.00111	0.00041	0.00056	0.00264	0.00036	0.00211	0.00092
BTN2A2	SKAT	0.00107	0.00437	0.00051	0.00829	0.01471	0.01082	0.00110	0.00317	0.00089
	huber	0.00120	0.00518	0.00048	0.00980	0.01787	0.01229	0.00093	0.00289	0.00066
	LAD	0.07623	0.11484	0.05847	0.02774	0.03912	0.02327	0.01145	0.01781	0.01383
	hampel	0.00113	0.00486	0.00045	0.00946	0.01717	0.01201	0.00087	0.00267	0.00062
	bisquare	0.00124	0.00534	0.00049	0.00885	0.01690	0.01102	0.00112	0.00329	0.00081
BTN3A2	SKAT	0.02343	0.01009	0.02416	0.20429	0.24686	0.22426	0.13090	0.09418	0.20471
	huber	0.01923	0.00748	0.02219	0.29643	0.34122	0.31866	0.12944	0.08839	0.20861
	LAD	0.00364	0.00217	0.00366	0.83200	0.72011	0.85070	0.05205	0.01981	0.09857
	hampel	0.02242	0.00897	0.02496	0.29702	0.33894	0.32503	0.13711	0.09392	0.21766
	bisquare	0.01799	0.00695	0.01980	0.27512	0.31934	0.30176	0.12756	0.08838	0.20416
HIST1H2AG	SKAT	0.16071	0.19460	0.20646	0.07240	0.13132	0.04615	0.14159	0.25024	0.07972
	huber	0.21564	0.25611	0.27324	0.04116	0.09070	0.01913	0.14309	0.24872	0.08733
	LAD	0.54097	0.45194	0.75342	0.00206	0.00299	0.00308	0.56707	0.67698	0.60459
	hampel	0.20897	0.24852	0.26543	0.04664	0.09938	0.02270	0.14189	0.24766	0.08583
	bisquare	0.22299	0.26006	0.28861	0.04175	0.09001	0.02075	0.16726	0.28389	0.10243
HIST1H2BJ	SKAT	0.49672	0.54134	0.65553	0.00198	0.00363	0.00228	0.06033	0.09746	0.05222
	huber	0.53584	0.59749	0.65911	0.00100	0.00204	0.00102	0.06882	0.10653	0.06503
	LAD	1.00000	1.00000	1.00000	**0.00016**	**0.00011**	0.00057	0.27492	0.35330	0.33107
	hampel	0.53660	0.59732	0.66185	0.00103	0.00212	0.00104	0.06658	0.10383	0.06241
	bisquare	0.56778	0.62094	0.71714	0.00103	0.00201	0.00113	0.07764	0.12006	0.07194
MHC	SKAT	0.00272	0.00718	0.00266	**8.7e-06**	**0.00001**	**0.00001**	0.02226	0.03543	0.02841
	huber	0.00125	0.00312	0.00104	**8.38e-07**	**8.67e-07**	**8.43e-07**	0.01605	0.02368	0.01938
	LAD	0.01508	0.02202	0.01394	0.00082	0.00191	0.00222	0.03513	0.02815	0.04447
	hampel	0.00124	0.00313	0.00104	**6.5e-07**	**6.68e-07**	**6.55e-07**	0.01475	0.02194	0.01790
	bisquare	0.00144	0.00347	0.00124	**7.53e-07**	**8.19e-07**	**7.49e-07**	0.01577	0.02201	0.01917
NOTCH4	SKAT	0.00712	0.01106	0.01136	0.10963	0.33275	0.14997	0.23368	0.55807	0.26421
	huber	0.00769	0.01098	0.01233	0.08378	0.27088	0.11515	0.26686	0.58140	0.29398
	LAD	0.03400	0.05464	0.06487	0.18388	0.47369	0.28816	0.17461	0.33604	0.17000
	hampel	0.00731	0.01074	0.01171	0.08567	0.27312	0.11927	0.25741	0.56783	0.28377
	bisquare	0.00831	0.01177	0.01359	0.09295	0.28921	0.12960	0.24446	0.53772	0.27090
POM121L2	SKAT	0.19270	0.23583	0.17564	0.00076	0.00161	0.00060	0.07210	0.11318	0.05031
	huber	0.18253	0.22679	0.16704	0.00061	0.00135	0.00042	0.06710	0.10504	0.04447
	LAD	0.03095	0.05625	0.02184	0.00994	0.01889	0.00429	0.03481	0.02924	0.04001
	hampel	0.18764	0.23256	0.17154	0.00062	0.00137	0.00043	0.06673	0.10518	0.04472
	bisquare	0.16296	0.20227	0.15109	0.00065	0.00143	0.00043	0.06104	0.09244	0.04390
PRSS16	SKAT	0.13207	0.13207	0.13207	0.33001	0.33001	0.33001	0.04146	0.04146	0.04146
	huber	0.13900	0.13900	0.13909	0.29159	0.29159	0.28911	0.04128	0.04128	0.04256
	LAD	0.10038	0.09568	0.10043	0.24981	0.24981	0.23406	0.03620	0.03620	0.03802
	hampel	0.13728	0.13728	0.13732	0.30152	0.30152	0.29892	0.03980	0.03980	0.04108
	bisquare	0.13566	0.13566	0.13582	0.30291	0.30291	0.30077	0.03732	0.03732	0.03869
SLC17A1	SKAT	0.01666	0.02739	0.03670	0.00037	0.00109	0.00085	0.00077	0.00326	0.00146
	huber	0.01159	0.02047	0.02778	0.00030	0.00082	0.00071	0.00071	0.00289	0.00144
	LAD	0.16536	0.20482	0.29532	0.00241	0.00215	0.00304	0.03599	0.08460	0.05064
	hampel	0.01220	0.02136	0.02889	**0.00026**	0.00077	0.00063	0.00068	0.00281	0.00138
	bisquare	0.01300	0.02212	0.03020	0.00031	0.00076	0.00069	0.00092	0.00374	0.00182
SLC17A3	SKAT	0.28193	0.22871	0.19521	0.02268	0.05884	0.04014	0.04898	0.04257	0.04197
	huber	0.28986	0.23930	0.23899	0.03010	0.06998	0.04569	0.05496	0.04597	0.05045
	LAD	0.77521	0.74479	0.62417	0.02699	0.10134	0.04518	0.13866	0.14578	0.11608
	hampel	0.28474	0.23400	0.23218	0.02157	0.05556	0.03370	0.04840	0.04120	0.04561
	bisquare	0.29599	0.24032	0.23673	0.02563	0.06366	0.04053	0.06014	0.04991	0.05326
ZNF184	SKAT	0.52845	0.52178	0.71230	0.41907	0.34434	0.58803	0.32582	0.33276	0.38268
	huber	0.49687	0.48631	0.67116	0.48083	0.42036	0.67621	0.33242	0.33776	0.39019
	LAD	0.41699	0.37620	0.51461	0.65516	0.69846	0.73871	0.13120	0.11966	0.18275
	hampel	0.49567	0.48590	0.67039	0.47578	0.40326	0.65908	0.32776	0.33305	0.38367
	bisquare	0.49216	0.48012	0.66667	0.47937	0.41481	0.67595	0.31835	0.31798	0.37039



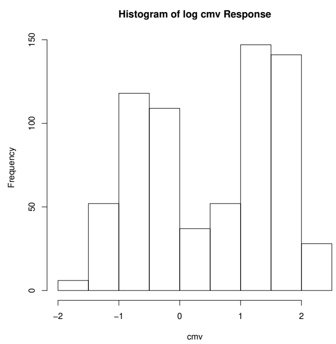
(a) CMV Response



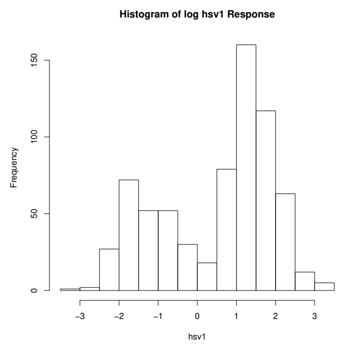
(b) $HSV1$ Response



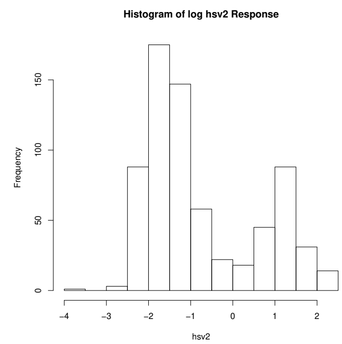
(c) $HSV2$ Response



(d) $log(CMV)$ Response



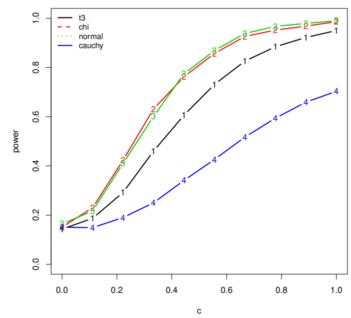
(e) $log(HSV1)$ Response



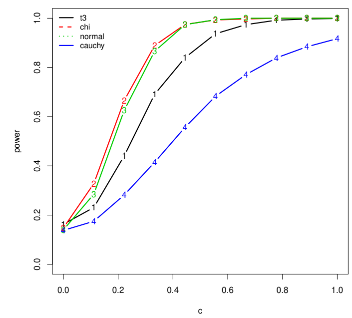
(f) $log(HSV2)$ Response

Figure A.5 Distribution of the continuous responses from the CATIE study. Before log-transformation (a-c), all distributions are notably skewed. After log-transformation (d-f), all distributions appear bimodal. This suggests a need for a robust testing method.

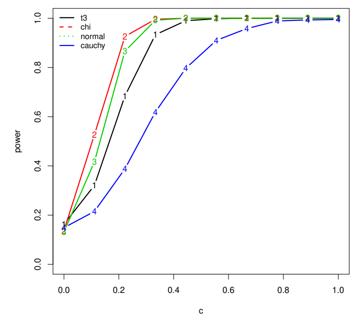
A.6 RobKAT Naive χ^2 Mixture Null Distribution



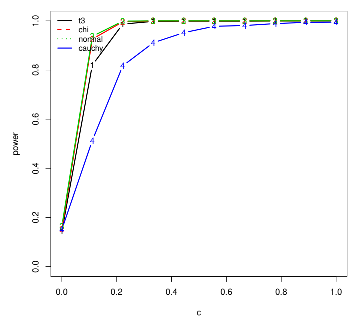
(a) $n = 25$



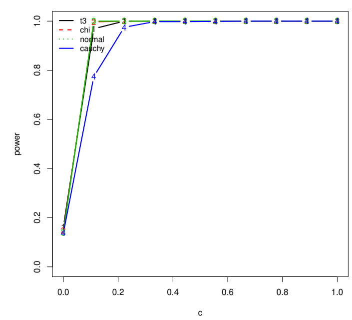
(b) $n = 50$



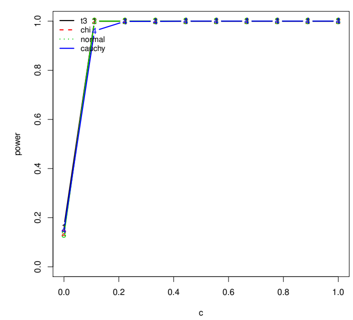
(c) $n = 100$



(d) $n = 25$



(e) $n = 50$



(f) $n = 100$

Figure A.6 Power simulations naively using the LSKM null distribution with the RobKAT statistic. The LSKM null distribution of the test statistic for the overall genetic marker effect follows a Chi-squared mixture, which is approximated using a Satterthwaite approximation (Liu et al. 2007). The simulation includes IBS kernel, Huber loss, sample sizes $n \in \{25, 50, 100\}$, $p = 9$ SNPs, and 1,000 iterations at each setting of c. Data was generated from $Y_i = \boldsymbol{X}_i^T \beta + c * h(\boldsymbol{Z}_i) + \epsilon_i$ with ϵ_i iid from t_3 (1: black), χ_1^2 (2: red), $N(0,1)$ (3: green), and $Cauchy(0,1)$ (4: blue). The top row corresponds to the linear case $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \mathbf{1}_9$, and the bottom row corresponds to the nonlinear case $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{p} \boldsymbol{Z}_{ik} \eta_k + 2 \sum_{k=2}^{p} \boldsymbol{Z}_{i1} \boldsymbol{Z}_{ik}$. When using this naive approach with incorrect null distribution, testing power is suboptimal compared to the permutation approach.

Table A.5 Type I error simulations naively using the LSKM null distribution with the RobKAT statistic. The LSKM null distribution of the test statistic for the overall genetic marker effect follows a Chi-squared mixture, which is approximated using a Satterthwaite approximation (Liu et al. 2007). The simulation includes IBS kernel, Huber loss, sample sizes $n \in \{25, 50, 100\}$, $p = 9$ SNPs, and $1,000$ iterations at each setting of c. Data was generated from $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + c * h(\boldsymbol{Z}_i) + \epsilon_i$ with ϵ_i iid from t_3, χ^2_1, $N(0,1)$, and $Cauchy(0,1)$. The linear case is $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^T \mathbf{1}_9$, and the nonlinear case is $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^p \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^p \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$. When using this naive approach with incorrect null distribution, Type I error is inflated.

Data Type	$h(Z_i)$	$n = 25$		$n = 50$		$n = 100$	
		< 0.01	< 0.05	< 0.01	< 0.05	< 0.01	< 0.05
$t3$	linear	0.0542	0.1469	0.0562	0.1513	0.0549	0.1482
χ^2_1	linear	0.0570	0.1556	0.0558	0.1442	0.0543	0.1401
$N(0,1)$	linear	0.0565	0.1518	0.0555	0.1462	0.0525	0.1487
$Cauchy(0,1)$	linear	0.0523	0.1465	0.0525	0.1377	0.0508	0.1386
$t3$	nonlinear	0.0542	0.1469	0.0562	0.1513	0.0549	0.1482
χ^2_1	nonlinear	0.0570	0.1556	0.0558	0.1442	0.0543	0.1401
$N(0,1)$	nonlinear	0.0565	0.1518	0.0555	0.1462	0.0525	0.1487
$Cauchy(0,1)$	nonlinear	0.0523	0.1465	0.0525	0.1377	0.0508	0.1386

A.7 High-Dimensional SNP-set Simulation with with n=300

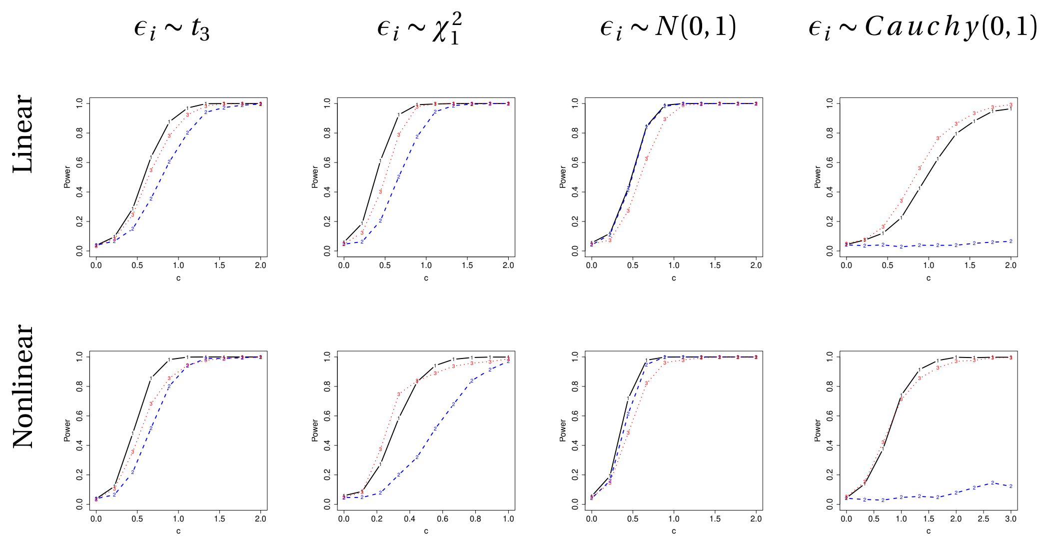
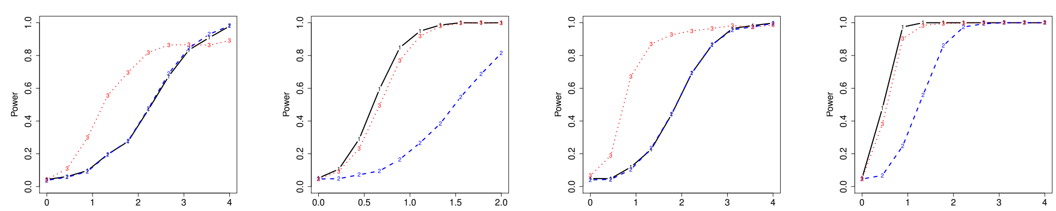


Figure A.7 Power simulations of RobKAT (black solid line), SKAT (blue dashed line), and LAD (red dotted line) for the setting of 787 simulated SNPs. This mirrors the simulation in Section 3.1 for $p = 787$ in the main text (Figure 3) but with a larger sample size ($n = 300$). The simulation uses IBS kernel, 1,000 iterations at each setting of c, and 2 causal SNPs. The top row corresponds to the linear case $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_{i1} + \boldsymbol{Z}_{i2}$, and the bottom row corresponds to the nonlinear case $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{2} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{2} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$.



(a) 30% mixture; n=300 $h(\boldsymbol{Z}_i)$ linear

(b) 10% mixture; n=300 $h(\boldsymbol{Z}_i)$ linear

(c) 30% mixture; n=300 $h(\boldsymbol{Z}_i)$ nonlinear

(d) 10% mixture; n=300 $h(\boldsymbol{Z}_i)$ nonlinear

Figure A.8 Power simulations with bimodal errors using RobKAT (black solid line), SKAT (blue dashed line), and LAD (red dotted line) for the setting of 787 simulated SNPs. This mirrors the simulation in Section 3.1 for $p = 787$ in the main text (Figure 4) but with a larger sample size ($n = 300$). The simulation uses IBS kernel, 1,000 iterations at each setting of c, and 2 causal SNPs. Figures (a) and (b) show the linear case $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_{i1} + \boldsymbol{Z}_{i2}$, and figures (c) and (d) display nonlinear case $h(\boldsymbol{Z}_i) = 1 + \sum_{k=1}^{2} \boldsymbol{Z}_{ik}\eta_k + 2\sum_{k=2}^{2} \boldsymbol{Z}_{i1}\boldsymbol{Z}_{ik}$.

Table A.6 Simulation study for type I error of RobKAT (Huber loss), QRKM (LAD loss), and SKAT in the setting of 787 simulated SNPs. This mirrors the simulation in Section 3.1 for $p = 787$ in the main text (Table 2) but with larger sample size ($n = 300$). The simulation includes IBS kernel, and estimates show proportion of 10,000 iterations with p-value below α cutoff.

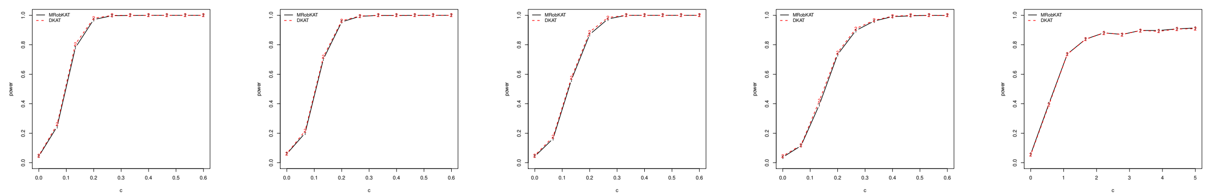
Error	α	Loss Function (ρ)		
		Huber	LAD	SKAT
	0.01	0.0092	0.0088	0.0064
t_3	0.05	0.0478	0.0476	0.0416
	0.01	0.0101	0.0085	0.0084
χ_1^2	0.05	0.0486	0.0469	0.0422
	0.01	0.0100	0.0095	0.0067
$N(0,1)$	0.05	0.0440	0.0467	0.0370
	0.01	0.0112	0.0102	0.0065
$Cauchy$	0.05	0.0488	0.0512	0.0332
	0.01	0.0104	0.0116	0.0062
10% Mix	0.05	0.0480	0.0485	0.0403
	0.01	0.0122	0.0098	0.0098
30% Mix	0.05	0.0540	0.0447	0.0436

APPENDIX

B

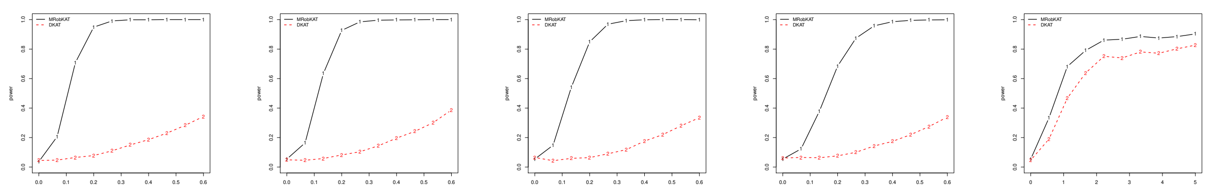
SUPPLEMENTAL INFORMATION FOR CHAPTER 4

B.1 Power Results for the Standard and Robust Settings with Nonlinear h



(a) 0 Binary Resp. **(b)** 1 Binary Resp. **(c)** 2 Binary Resp. **(d)** 3 Binary Resp. **(e)** 4 (all) Binary Resp.

Figure B.1 Power simulation results from the standard setting (no high leverage points). MrKAT is shown in black, and DKAT is shown in red. 4 Responses; IBS kernel, n=100; $h_1 = h_2 = h_3 = h_4 = h(z) = 2(z_1 - z_2)^2 + z_2 z_3 + 3sin(2z_3)z_4 + z_5^2 + 2cos(z_4)z_5$; $\beta = -5\mathbf{1}_4$



(a) 0 Binary Resp. **(b)** 1 Binary Resp. **(c)** 2 Binary Resp. **(d)** 3 Binary Resp. **(e)** 4 (all) Binary Resp.

Figure B.2 Power simulation results from the robust setting (1 high leverage point). MrKAT is shown in black, and DKAT is shown in red. 4 Responses; IBS kernel, n=100; $h_1 = h_2 = h_3 = h_4 = 2(z_1 - z_2)^2 + z_2 z_3 + 3sin(2z_3)z_4 + z_5^2 + 2cos(z_4)z_5$; $\beta = -5\mathbf{1}_4$; $Y_{100} \sim N(10,1)$ or $Y_{100} = 1$, $X_{100}\ N(3,1)$

B.2 CATIE Data Analysis on Binary Responses

Table B.1 Applying MrKAT and DKAT to the binary form of HSV1, HSV2, and CMV antibody levels. The bold values highlight significant p-values after a Bonferroni correction $(0.05/(2*3*12) = 0.00069)$ for the 2 methods, 3 kernels, and 12 genes.

Gene	Test	linear	quadratic	IBS
BTN2A1	DKAT	**1.11732e-07**	**2.51799e-06**	**1.82289e-06**
	MrKAT	**6.88427e-08**	**1.65947e-06**	**1.17446e-06**
BTN2A2	DKAT	**6.11241e-06**	**9.87614e-05**	**9.13882e-07**
	MrKAT	**4.92310e-06**	**8.53123e-05**	**6.81298e-07**
BTN3A2	DKAT	0.03206	0.02171	0.06166
	MrKAT	0.02976	0.02062	0.05785
HIST1H2AG	DKAT	0.20793	0.33845	0.18148
	MrKAT	0.18630	0.31744	0.15690
HIST1H2BJ	DKAT	0.01376	0.02749	0.01605
	MrKAT	0.01398	0.02907	0.01531
MHC	DKAT	**6.89945e-07**	**7.22392e-07**	**1.37308e-06**
	MrKAT	**6.24832e-07**	**6.89728e-07**	**1.17706e-06**
NOTCH4	DKAT	0.03659	0.08357	0.05563
	MrKAT	0.03508	0.07833	0.05538
POM121L2	DKAT	0.00173	0.00267	0.00266
	MrKAT	0.00192	0.00305	0.00281
PRSS16	DKAT	0.06557	0.06557	0.06557
	MrKAT	0.05208	0.05208	0.05377
SLC17A1	DKAT	**2.53081e-06**	**3.24997e-05**	**2.13407e-05**
	MrKAT	**2.59972e-06**	**3.29762e-05**	**2.16830e-05**
SLC17A3	DKAT	0.05219	0.11881	0.05580
	MrKAT	0.06034	0.12949	0.06743
ZNF184	DKAT	0.66038	0.61805	0.78087
	MrKAT	0.67307	0.62420	0.79012

B.3 Influence and Leverage Measurements

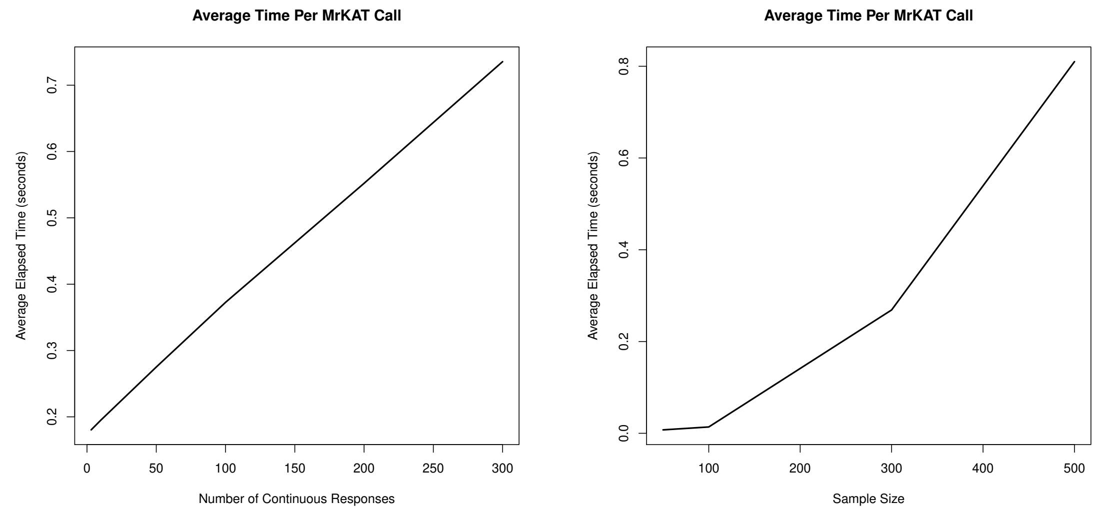
Table B.2 Approximate influence of added outlier to simulation study described in Section 4.3.2. Simulated Cook's Distance and leverage were based on 1000 iterations mirroring simulation study settings of nonlinear $h_j(\boldsymbol{Z}_i)$ for $j \in \{1,2,3,4\}$, four responses, and $c = 0.5$. Leverage is only dependent on covariate matrix, $\boldsymbol{X}_i$.

Response	Outlier Added	Cooks Distance	Leverage
Continuous	No	0.0098	0.0398
Continuous	Yes	9.1494	0.2959
Binary	No	0.0151	0.0398
Binary	Yes	2.2143	0.2959

Table B.3 Maximum influence (measured by Cook's Distance to linear or generalized linear model) and maximum leverage for the CATIE antibody data analyzed in Section 4.4. Since leverage values sum to $q = \#$ Covariates $= 2$ and the expected leverage value would be $q/n = 0.0028$, we do not see evidence of extreme leverage observations in this data set. These metrics also do not show strong evidence of highly influential observations.

	hsv1	hsv2	cmv	max leverage
continuous	0.0014	0.0017	0.0015	0.0114
binary	0.0014	0.0016	0.0014	0.0114

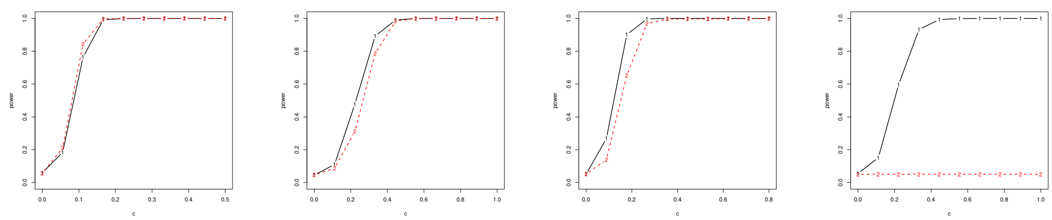
B.4 MrKAT Time of Computation



(a) Time Across Response Size **(b)** Time Across Sample Size

Figure B.3 Estimated average times for one MrKAT test from 1000 simulated iterations. For $i = 1, 2, ..., n$ and $j = 1, 2, ..., r$, each simulation setting included data generated from $Y_{ij} = \boldsymbol{X}_i^T \boldsymbol{\beta} + h_j(\boldsymbol{Z}_i) + \epsilon_{ij}$, where $X_i \sim N(0, 1)$, $\boldsymbol{\beta} = -5\mathbf{1}_5$, $\boldsymbol{\epsilon}_i \sim N_r(0, I_r)$, $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i \mathbf{1}_p$, and $\boldsymbol{Z}_i$ contained $p = 200$ SNPs with overall minor allele frequencies matching the *SLC17A1* gene. For Plot (a), times were estimated for $n = 300$ and $r = 3, 10, 50, 100, 200, 300$ responses. For Plot (b), times were estimated for $n = 50, 100, 300, 500$. Times were calculated on a 2.9 GHz Intel Core i5 machine with 8 GB 1867 MHz memory.

B.5 Simulation Results for the Non-normal Setting with 200 Responses



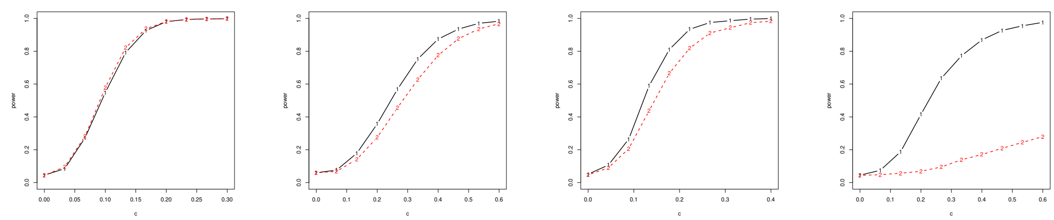
(a) $N(0,1)$ Errors **(b)** χ_1^2 Errors **(c)** t_3 Errors **(d)** Cauchy Errors

Figure B.4 Power simulation results for 200 continuous, multivariate responses that have errors that are symmetric ($N(0,1)$), skewed (χ_1^2), heavy-tailed (t_3), and substantially heavy-tailed (Cauchy). Of the 200 responses, five were associated ($Y_{ij} = \boldsymbol{X}_i\beta + c * h_j(\boldsymbol{Z}_i) + \epsilon_{ij}$) and 195 were not associated ($Y_{ij} = \epsilon_{ij}$) with the covariates. MrKAT is shown in black, and DKAT is shown in red. Implementation included an IBS genotype kernel, sample size of $n = 100$, 1000 iterations at each value of c, $\beta = -5\mathbf{1}_4$, and $h_1 = h_2 = h_3 = h_4 = 2z_1 + 3z_2 + z_3 + 2z_4 + z_5$.

Table B.4 Simulation study for type I error exploring the testing properties from 200 continuous, multivariate responses that have errors that are symmetric ($N(0,1)$), skewed (χ_1^2), heavy-tailed (t_3), and substantially heavy-tailed (Cauchy). Of the 200 responses, five were associated ($Y_{ij} = \boldsymbol{X}_i\beta + c * h_j(\boldsymbol{Z}_i) + \epsilon_{ij}$) and 195 were not associated ($Y_{ij} = \epsilon_{ij}$) with the covariates. Data was generated with IBS kernel, sample size $n = 100$, $\beta = -5\mathbf{1}_4$, and estimates averaged over 1000 iterations.

Error	α	MrKAT	DKAT
	0.01	0.007	0.004
$N(0,1)$	0.05	0.060	0.057
	0.10	0.114	0.119
	0.01	0.008	0.007
χ_1^2	0.05	0.046	0.046
	0.10	0.096	0.098
	0.01	0.009	0.004
t_3	0.05	0.051	0.052
	0.10	0.102	0.106
	0.01	0.009	0.016
Cauchy	0.05	0.055	0.051
	0.10	0.104	0.090

B.6 Simulation Results for the Non-normal Setting with 25 Responses



(a) $N(0,1)$ Errors **(b)** χ_1^2 Errors **(c)** t_3 Errors **(d)** Cauchy Errors

Figure B.5 Power simulation results for 4 continuous, multivariate responses that have errors that are symmetric ($N(0,1)$), skewed (χ_1^2), heavy-tailed (t_3), and substantially heavy-tailed (Cauchy). Of the 4 responses, all were generated with $Y_{ij} = \boldsymbol{X}_i\beta + c * h_j(\boldsymbol{Z}_i) + \epsilon_{ij}$. MrKAT is shown in black, and DKAT is shown in red. Implementation included an IBS genotype kernel, sample size of $n = 25$, 1000 iterations at each value of c, $\beta = -5\mathbf{1}_4$, and $h_1 = h_2 = h_3 = h_4 = 2z_1 + 3z_2 + z_3 + 2z_4 + z_5$.

Table B.5 Simulation study for type I error exploring the testing properties from 4 continuous, multivariate responses that have symmetric ($N(0,1)$), skewed (χ^2_1), heavy-tailed (t_3), and substantially heavy-tailed (Cauchy) error distributions. Of the 4 responses, all were generated with $Y_{ij} = \boldsymbol{X}_i\beta + c * h_j(\boldsymbol{Z}_i) + \epsilon_{ij}$. Data was generated with IBS kernel, sample size $n = 25$, $\beta = -5\mathbf{1}_4$, and estimates averaged over 10,000 iterations.

Error	α	MrKAT	DKAT
	0.01	0.0083	0.0088
$N(0,1)$	0.05	0.0467	0.0467
	0.10	0.0958	0.0947
	0.01	0.0104	0.0089
χ^2_1	0.05	0.0520	0.0486
	0.10	0.1019	0.0989
	0.01	0.0118	0.0106
t_3	0.05	0.0515	0.0514
	0.10	0.1012	0.0997
	0.01	0.0089	0.0088
Cauchy	0.05	0.0484	0.0476
	0.10	0.0960	0.0991

B.7 Percent Contribution Exploration

Table B.6 Simulated T_j/T for each of $j = 1, 2, 3$ continuous responses, averaged over 1000 iterations. Data was generated from $Y_j = \boldsymbol{X}_i\beta + c_j h_j(\boldsymbol{Z}_i) + \epsilon_{ij}$ for $i = 1, 2, ..., n$ samples and $\boldsymbol{Z}_i$ a randomly sampled 9-SNP from the *SLC17A1* gene. Independent errors represent $\epsilon_{ij} \sim N(0,1)$, and correlated ϵ_{ij} is the $(i,j)^{th}$ entry of E^*, where $E_{ij} \sim N(0,1)$ and the errors have the correlation structure

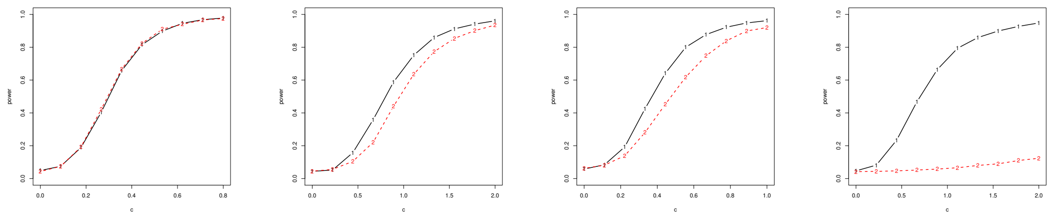
$$E^* = E \cdot chol\left(\begin{bmatrix} 0.95 & 0.07 & 0.23 \\ 0.07 & 0.86 & 0.24 \\ 0.23 & 0.24 & 0.89 \end{bmatrix}\right)$$

c_1	c_2	c_3	Sample Size	Error	Y_1	Y_2	Y_3
0.4	0.4	0.4	$n = 50$	Independent	0.3343	0.3326	0.3331
				Correlated	0.3291	0.3369	0.3340
			$n = 100$	Independent	0.3330	0.3322	0.3348
				Correlated	0.3280	0.3365	0.3355
			$n = 300$	Independent	0.3326	0.3340	0.3334
				Correlated	0.3276	0.3381	0.3344
			$n = 500$	Independent	0.3335	0.3333	0.3332
				Correlated	0.3283	0.3375	0.3342
0	0.4	0.4	$n = 50$	Independent	0.0401	0.4795	0.4804
				Correlated	0.0386	0.4826	0.4788
			$n = 100$	Independent	0.0205	0.4879	0.4916
				Correlated	0.0198	0.4908	0.4894
			$n = 300$	Independent	0.0067	0.4971	0.4962
				Correlated	0.0065	0.4995	0.4940
			$n = 500$	Independent	0.0040	0.4981	0.4979
				Correlated	0.0038	0.5006	0.4956
0	0.2	0.4	$n = 50$	Independent	0.0515	0.3300	0.6185
				Correlated	0.0490	0.3426	0.6084
			$n = 100$	Independent	0.0268	0.3312	0.6420
				Correlated	0.0255	0.3449	0.6297
			$n = 300$	Independent	0.0088	0.3390	0.6522
				Correlated	0.0084	0.3529	0.6388
			$n = 500$	Independent	0.0052	0.3381	0.6567
				Correlated	0.0049	0.3522	0.6429

Table B.7 Simulated T_j/T for each of $j = 1,2,3$ continuous responses, averaged over 1000 iterations. Data was generated from $Y_j = \boldsymbol{X}_i\beta + c_j h(\boldsymbol{Z}_i) + \epsilon_{ij}$ for $i = 1,2,...,n$ samples, $X_{ij} \sim N(0,1)$, $\beta = -5\mathbf{1}_2$, $\boldsymbol{Z}_i$ a randomly sampled 9-SNP from the *SLC17A1* gene, $h_1(\boldsymbol{Z}_i) = h_2(\boldsymbol{Z}_i) = h_3(\boldsymbol{Z}_i) = 2z_1 + 3z_2 + z_3 + 2z_4 + z_5$, independent errors from $\epsilon_{ij} \sim N(0, \sigma_j^2)$, and an IBS kernel. To vary spreads of the different responses, $\sigma_1 = 1$, $\sigma_2 = 5$, and $\sigma_3 = 10$.

c_1	c_1	c_1	Sample Size	Y_1	Y_2	Y_3
0.4	0.4	0.4	$n = 50$	0.565	0.264	0.171
			$n = 100$	0.576	0.259	0.165
			$n = 300$	0.583	0.262	0.155
			$n = 500$	0.588	0.260	0.153
0	0.4	0.4	$n = 50$	0.099	0.546	0.354
			$n = 100$	0.054	0.579	0.367
			$n = 300$	0.019	0.617	0.364
			$n = 500$	0.011	0.623	0.366
0	0	0.4	$n = 50$	0.183	0.181	0.636
			$n = 100$	0.118	0.109	0.773
			$n = 300$	0.047	0.045	0.908
			$n = 500$	0.029	0.029	0.942
0	0	0	$n = 50$	0.334	0.331	0.334
			$n = 100$	0.345	0.322	0.333
			$n = 300$	0.340	0.328	0.332
			$n = 500$	0.336	0.331	0.334
0	0.2	0.4	$n = 50$	0.146	0.340	0.514
			$n = 100$	0.087	0.334	0.579
			$n = 300$	0.032	0.351	0.617
			$n = 500$	0.019	0.350	0.631

B.8 Simulation Results in the Non-normal Setting with p=787 SNPs



(a) $N(0,1)$ Errors **(b)** χ_1^2 Errors **(c)** t_3 Errors **(d)** Cauchy Errors

Figure B.6 Power simulation results for four continuous, independent, multivariate responses that are symmetric ($N(0,1)$), skewed (χ_1^2), heavy-tailed (t_3), and substantially heavy-tailed (Cauchy). MrKAT is shown in black, and DKAT is shown in red. Genetic marker sets ($p = 787$ SNPs) were randomly sampled from the MHC region. This simulation utilized an IBS kernel, sample size $n = 100$, 1000 iterations at each value of c, $\beta = -5\mathbf{1}_4$, and the linear form $h_1(\boldsymbol{Z}_i) = ... = h_4(\boldsymbol{Z}_i) = 2z_1 + 3z_2 + z_3 + 2z_4 + z_5$;

APPENDIX

C

SUPPLEMENTAL INFORMATION FOR CHAPTER 5

C.1 CATIE Data Analysis Results (HSV-1)

Table C.1 Data analysis of CATIE cognitive score data using RobKAT (Huber loss) and the CKLRT, while incorporating possible gene-environment interaction with HSV-1 exposure. The significance threshold was determined via a Bonferroni correction based on 12 gene analyses and 2 methods ($\alpha^* = 0.05/24 = 0.00208$). Significant results are shown in bold.

Gene	SNPs	Test	Verbal	Speed	Reason	Memory	Vigilance
BTN2A1	4	RoCKAT	0.2072	0.0645	0.0039	0.0233	0.3092
		CKLRT	0.1835	0.1508	0.0113	0.0299	0.1975
BTN2A2	5	RoCKAT	0.6486	0.722	0.0327	0.014	0.3720
		CKLRT	1.0000	1.0000	0.0775	0.0997	1.0000
BTN3A2	5	RoCKAT	0.8608	0.7828	0.4957	0.6051	0.5947
		CKLRT	1.0000	1.0000	0.3143	1.0000	0.1474
HIST1H2AG	2	RoCKAT	0.0807	0.8957	0.5744	0.1829	0.9163
		CKLRT	0.1491	1.0000	1.0000	0.4092	1.0000
HIST1H2BJ	2	RoCKAT	0.2431	0.8938	0.4766	0.7244	0.8085
		CKLRT	0.4551	1.0000	1.0000	1.0000	1.0000
MHC	787	RoCKAT	0.0538	0.1726	**0.0018**	0.0071	0.0836
		CKLRT	0.1353	0.1786	0.0041	0.0351	0.0441
NOTCH4	24	RoCKAT	0.8444	0.1218	0.2025	0.0281	0.1688
		CKLRT	1.0000	0.0976	0.0830	0.0190	0.1604
POM121L2	4	RoCKAT	0.2570	0.2879	0.4703	0.1519	0.3244
		CKLRT	0.1571	0.1173	1.0000	0.2356	0.4087
PRSS16	1	RoCKAT	0.5041	0.5907	0.6054	0.1133	0.6417
		CKLRT	1.0000	1.0000	1.0000	0.2139	1.0000
SLC17A1	9	RoCKAT	0.4366	0.7999	0.1179	0.2259	0.7405
		CKLRT	0.1712	1.0000	0.1920	0.3654	1.0000
SLC17A3	13	RoCKAT	0.9132	0.5080	0.0586	0.1580	0.5840
		CKLRT	1.0000	0.3773	0.0472	0.1229	0.3203
ZNF184	8	RoCKAT	0.2145	0.7543	0.0228	0.8867	0.5681
		CKLRT	0.4916	1.0000	0.1073	1.0000	0.4086

C.2 CATIE Data Analysis Results (CMV)

Table C.2 Data analysis of CATIE cognitive score data using RobKAT (Huber loss) and the CKLRT, while incorporating possible gene-environment interaction with CMV exposure. The significance threshold was determined via a Bonferroni correction based on 12 gene analyses and 2 methods ($\alpha^* = 0.05/24 = 0.00208$). Significant results are shown in bold.

Gene	SNPs	Test	Verbal	Speed	Reason	Memory	Vigilance
BTN2A1	4	RoCKAT	0.1967	0.1121	0.0086	0.0512	0.3335
		CKLRT	0.2133	0.2161	0.0325	0.0753	0.3455
BTN2A2	5	RoCKAT	0.8689	0.7489	0.0401	0.0155	0.3856
		CKLRT	1.0000	1.0000	0.0801	0.1247	0.4801
BTN3A2	5	RoCKAT	0.8730	0.6520	0.5358	0.5736	0.6371
		CKLRT	1.0000	0.4160	1.0000	1.0000	1.0000
HIST1H2AG	2	RoCKAT	0.0756	0.7953	0.5797	0.1594	0.9634
		CKLRT	0.0576	1.0000	1.0000	1.0000	1.0000
HIST1H2BJ	2	RoCKAT	0.2232	0.9325	0.5427	0.7643	0.8647
		CKLRT	0.2770	1.0000	1.0000	1.0000	1.0000
MHC	787	RoCKAT	0.0559	0.2037	0.0068	0.0212	0.0817
		CKLRT	0.2051	0.2736	0.0126	0.0487	0.0910
NOTCH4	24	RoCKAT	0.8665	0.1220	0.2014	0.0324	0.1550
		CKLRT	1.0000	0.1031	0.1043	0.0182	0.1563
POM121L2	4	RoCKAT	0.2790	0.2769	0.6063	0.1819	0.2161
		CKLRT	0.2543	0.0578	1.0000	0.2519	0.0870
PRSS16	1	RoCKAT	0.5206	0.6266	0.6176	0.1327	0.8449
		CKLRT	1.0000	1.0000	1.0000	0.2423	1.0000
SLC17A1	9	RoCKAT	0.5439	0.7102	0.2390	0.3598	0.6848
		CKLRT	1.0000	1.0000	0.3795	1.0000	1.0000
SLC17A3	13	RoCKAT	0.8375	0.2707	0.0616	0.1777	0.5732
		CKLRT	1.0000	0.2010	0.0567	0.1572	0.4068
ZNF184	8	RoCKAT	0.2186	0.7218	0.0259	0.9038	0.5519
		CKLRT	0.4957	1.0000	0.1203	1.0000	0.3539

C.3 Power Results with Continuous E

Table C.3 Power simulations of ROCKAT and the CKLRT. The number of causal SNPs (k) is 20, and no covariates. Data generated with $E_i \sim N(0,1)$ using the MHC gene ($p = 787$ SNPs) with $h_G(\boldsymbol{G}_i) = \sum_{i=1}^{k} G_{ij}$, and $h_{GE}(\boldsymbol{G}_i, E_i) = \sum_{i=1}^{k} G_{ij} E_i$. The simulation tests use $E_i^* = I\{E_i \leq med(\boldsymbol{E})\}$, IBS genetic marker kernel, sample size $n = 100$, and $1,000$ iterations at each setting of β. ROCKAT used 500 permutations per test, CKLRT used $N = 5000$ generated variables for the empirical null distribution of the LRT. There is low power within this β window, as the K_{GE} built on the binary representation does not model the data with continuous E_i well.

	Error	Test	$\beta = 0$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$
β_G Only	$N(0,1)$	RoCKAT	0.058	0.064	0.322	0.706
		CKLRT	0.056	0.071	0.267	0.690
	χ_1^2	RoCKAT	0.055	0.081	0.322	0.690
		CKLRT	0.166	0.148	0.278	0.510
β_{GE} Only	$N(0,1)$	RoCKAT	0.058	0.050	0.049	0.047
		CKLRT	0.056	0.066	0.082	0.0820
	χ_1^2	RoCKAT	0.055	0.040	0.049	0.051
		CKLRT	0.166	0.106	0.095	0.081
$\beta_{GE} = \beta_G$	$N(0,1)$	RoCKAT	0.058	0.068	0.087	0.093
		CKLRT	0.056	0.064	0.081	0.090
	χ_1^2	RoCKAT	0.055	0.057	0.076	0.094
		CKLRT	0.166	0.119	0.093	0.102